

第22回 日本SASユーザー会総会
および研究発表会
論文集

2003年7月31日(木)~8月1日(金)

SAS、SAS を構成するプロダクト群は、SAS Institute Inc. の登録商標です。

その他、本論文集に記載されている会社名、製品名は、一般にそれぞれ各社の商標または登録商標です。

本論文集の一部または全部を無断転載することは、著作権法上の例外を除き、禁止されています。本論文集の内容を実際に運用した結果の影響については、責任を負いかねます。

目 次

口頭論文発表

◆ 医薬品開発

- 臨床試験の早期中止の検討における…………… 3
ベイズ流予測確率と条件付き検出力の利用について
堺 伸也（イーピーエス株式会社／東京理科大学）
菅波 秀規（興和株式会社／東京理科大学）
- Group Sequential計画のためのパワーシミュレータの開発…………… 11
本田 圭一（塩野義製薬株式会社）
田崎 武信
太田 裕二（住商情報システム株式会社）
佐賀野 修一
- 投与期間別の副作用発現率を解釈するために…………… 19
古川 雅史（塩野義製薬株式会社）
片山 和夫
田崎 武信
- SAS Integration Technologies + ASPによる…………… 31
解析帳票作成 Web システム構築の試み
岩本 光司（武田薬品工業株式会社）
矢野 尚（株式会社富士通ビー・エス・シー）
- <医薬特別セッション:シグモイド型用量反応曲線の解析>
- 2値および計量値のシグモイド曲線…………… 41
— 曲線の推定および逆推定と95%信頼区間 —
杉山 公仁（昭和薬品化工株式会社）
馬場 淳（明治製菓株式会社）
天竺桂 裕一郎（興和株式会社）
高橋 行雄（中外製薬株式会社）
- 陰性および陽性対照があるシグモイド曲線…………… 51
— ダミー変数を持つ非線型回帰モデルの応用 —
山田 雅之（キッセイ薬品工業株式会社）
吉田 光宏（グラクソ・スミスクライン株式会社）
高橋 行雄（中外製薬株式会社）
- 計量値のシグモイド用量反応曲線の同時推定…………… 61
— 効力比とその95%信頼区間 —
高橋 行雄（中外製薬株式会社）

◆ チュートリアル

生存時間解析における症例数設計	73
濱田 知久馬 (東京理科大学)	
藤井 陽介	

◆ 統計解析

区間打ち切り生存時間データのセミパラメトリックな解析法の紹介	101
SASプログラムの紹介	

ーギブス・サンプラーを利用した周辺尤度アプローチ	
西山 智 (アベンティス ファーマ株式会社/東京理科大学)	
吉村 健一 (東京大学)	

イベント発生確率推定時における連続変数のカテゴリー化、およびカテゴリー変数の実数化	113
上條 史夫 (株式会社数理技研)	
川崎 章弘	

SASによる生存時間の多重イベントの解析	121
ー糖尿病合併症を例にー	
広本 篤 (東京大学)	
金子 徹治	
大橋 靖雄	

再発事象に対するモデルを用いた解析方法の検討	131
中 牧子 (東京大学)	
大橋 靖雄	

MIXEDプロシジャを用いた線形混合効果モデルの交互作用の指定方法	141
寒水 孝司 (東京理科大学)	
菅波 秀規 (興和株式会社/東京理科大学)	

要因配置実験の効果成分の表示から生じる不定性	151
柴山 忠雄 (前：名古屋市工業研究所)	

◆ 統計教育

CROにおけるSASプログラムの育成教育	161
竹田 眞 (株式会社CRCソリューションズ)	
佐藤 智美	

◆ システム

- CALL EXECUTEを用いたマクロの再帰呼び出しと統計計算への応用…………… 169
伊藤 要二 (アストラゼネカ株式会社)
- SAS未経験者をSAS内部構造を理解したDATAステップSASプログラマに…………… 179
短期間で育成するカリキュラムの紹介
山田 大志 (アストラゼネカ株式会社)
小澤 康彦
宮浦 千香子
- Microsoft AccessとSASによるデータマネジメントシステム…………… 189
中村 竜児 (メディカル統計株式会社)
松沢 享
- SAS® Metadata, Authorization and Management Services…………… 199
—Working Together for You
SASによるメタデータマネジメント
Michelle Ryals (SAS Institute Inc.)
翻訳 鹿渡 圭二郎 (SAS Institute Japan 株式会社)
李 錦実
江口 英男
- Enterprise Guide 2.0によるadd-in機能について…………… 211
木下 貴文 (SAS Institute Japan 株式会社)
- SAS/SHAREサーバーアクセスログの分析…………… 219
中村 崇文 (SAS Institute Japan 株式会社)
- 簡易 運用入門…………… 229
弘田 貴 (SAS Institute Japan 株式会社)
- MEANS、TABULATE、DATASETSプロシージャの機能紹介…………… 239
檜皮 孝史 (SAS Institute Japan 株式会社)
渋谷 佳枝
迫田 奈緒子

◆ 経営・経済

- SASソフトウェアを利用したCIR++モデルの…………… 251
パラメータ推定と金利パス生成
岸田 則生 (株式会社CRCソリューションズ)
- コンシューマ・クレジット業の利益指向の新与信モデル…………… 261
小野 潔 (株式会社UFJ銀行)

非補償型ロジットモデルを用いた企業倒産確率の予測モデル…………… 269
—NLP Procedureによる非補償型ロジットモデルに対するパラメータ推定—
坂巻 英一（株式会社金融工学研究所／東京工業大学）

SAS Risk Dimensionsによる統合リスク分析のご紹介…………… 281
嘉陽 亜希子（SAS Institute Japan 株式会社）
鬼頭 拓郎
尾高 雅代
田中 愛

◆ 調査・マーケティング

建築生産における建築物の耐久性確保に関する実務者の意識と実態…………… 295
小島 隆矢（独立行政法人建築研究所）
小野 久美子（国土交通省国土技術政策総合研究所）
植木 暁司

JMPによるワインの顧客価値分析…………… 305
林 俊克（株式会社資生堂）
平野 広隆（株式会社アーキテクト）

看護師のセクシャルハラスメントに対する意識について…………… 315
田久 浩志（中部学院大学）
岩本 晋（NPO 福祉法人 OIEMASE）

Life Time Valueを基準とした施策の最適化方法…………… 319
—遺伝的アルゴリズムによる解析事例—
小谷田 知行（株式会社浜銀総合研究所）
堀 彰男

Bioinformaticsの手法を活用したクレジットカード取引履歴データの…………… 329
途上審査モデルへの適用事例
堀 彰男（株式会社浜銀総合研究所）
小谷田 知行

◆ SAS ソリューション

ゲノム創薬向け統合ソリューション…………… 341
SAS Scientific Discovery Solutionsの紹介
段谷 高章（SAS Institute Japan 株式会社）

ポスターセッション

◆ 統計解析

一般化推定方程式およびSASの解析ツール..... 351

王 露萍 (アベンティス ファーマ株式会社)

野口 知雄

高田 康行 (持田製薬株式会社)

NLMIXEDプロシジャーを用いたItemResponseModelのシミュレーション..... 361

板東 説也 (有限会社電助システムズ)

宮岡 悦良 (東京理科大学)

緑川 修一

高原 佳奈

変量効果モデルによるメタ・アナリシス..... 369

DerSimonian-Laird法のSASマクロの作成

中西 豊支 (東京理科大学)

浜田 知久馬

メタ・アナリシスにおける公表バイアスの評価..... 379

trim-and-fill法のSASマクロの作成

松岡 伸篤 (東京理科大学)

浜田 知久馬

◆ 統計教育

看護系大学における疫学・生物統計学教育の実態調査..... 391

田中 司朗 (東京大学)

◆ システム

SASを用いたXMLデータの作成 -ODM ver. 1.1対応-..... 403

岡下 邦博 (株式会社日本アルトマーク)

進藤 三富子

SASデータセットのエクスポート..... 409

羽田野 実 (SAS Institute Japan 株式会社)

◆ 経営・経済

労働市場の時系列分析 -JMPを利用して-..... 417

浦澤 浩一 (株式会社八千代銀行/青山学院大学)

アジルなSupply Chainを実現する予測プロセスの自動化.....	431
—SAS® High-Performance Forecastingのご紹介—	
松舘 学 (SAS Institute Japan 株式会社)	

◆ 調査・マーケティング

地方における実演芸術鑑賞の実態.....	443
—県民芸術劇場(兵庫県)の来場者調査より—	
有馬 昌宏 (神戸商科大学)	

青年期女性の自意識と完全主義傾向の関連.....	453
中村 晃士 (東京慈恵会医科大学)	
牛島 定信	
縣 俊彦	
清水 英佑	

個人レベルの選好を基にしたクラスタリング.....	459
河崎 一益 (株式会社日本アルトマーク)	
松沢 利繁 (株式会社インターナショナル・クリエイティブ・マーケティング)	

患者参加型医療情報交換システムのニーズ調査.....	465
義澤 宣明 (株式会社三菱総合研究所)	
船曳 淳	
小山 博史 (東京大学)	

◆ グラフィック・統計教育

SAS/GRAPH入門 —社内における教育研修事例—.....	477
林 行和 (株式会社CRCソリューションズ)	
畑中 雄介	
小出 起美雅	
山口 孝一	

◆ グラフィック・レポートニング

SASグラフによる動く万華鏡の作成.....	489
岸本 容司 (神戸商科大学)	

口頭論文発表 医薬品開発

日本SASユーザー会 (SUG J)

臨床試験の早期中止の検討における ベイズ流予測確率と条件付き検出力の利用について

○堺伸也*^{1,2} 菅波秀規*^{1,3}

*¹ 東京理科大学大学院工学研究科

*² イーピーエス株式会社統計解析部 *³ 興和株式会社臨床解析部

Comparison of Bayesian predictive probability and conditional power as a criterion
for early termination of clinical trials

Shinya Sakai*^{1,2} Hideki Suganami*^{1,3}

*¹ Graduate School of Engineering, Tokyo University of Science

*² Statistics Analysis Dept., EPS Co., Ltd.

*³ Biostatistics and Data Management Dept., Kowa Co., Ltd.

要 旨

近年、中間解析を実施する試験が増えてきている。中間解析は「臨床試験のための統計的原則」にも明記されており、臨床試験での標準的な実験デザインの一つとして認知されている。臨床試験での薬剤の有効性に関して中間解析を行う目的は大きく2つに分類される。①有効性が示されたと判断して、試験を早期中止する(帰無仮説 H_0 を棄却する)。②将来有効性を示すことが難しいと判断して、試験を早期中止する。本稿では、②に関してしばしば利用される「ベイズ流予測確率」と「条件付き検出力」についての検討を行った。これらの手法は、将来有意差の得られる確率を示すため、数値の意味を臨床家へ説明し易く利用性は高いと思われる。しかし、 β エラーを制御することは第一の目標とされていないため利用に際しては注意が必要である。本稿では単純な臨床試験の状況を想定し「条件付き検出力」と「ベイズ流予測確率」を算出するためのSASプログラムを作成し、 β エラー、試験を停止させる確率等の性能をSAS8.2を利用して評価した。

キーワード： ベイズ流予測確率、条件付き検出力、早期中止

ベイズ流予測確率と条件付き検出力の算出式

手法の詳細は、宇野,松井,小山(2000)で述べられている。論文中的でのプロトタイプケースの設定を本稿では利用した。

<記号>

簡単のため 1 標本で試験を考え、各症例 i のデータ Y_i が平均値 δ (薬剤効果)、分散 σ^2 (既知)の正規分布に従うとする。帰無仮説 $H_0: \delta = 0$ を片側有意水準 α で検定するものとする。中間解析時に m 症例のデータを回収済みで $\{y_1, y_2, \dots, y_m\}$ 、最終解析までに残り n 症例のデータを回収する予定とする $\{Y_{m+1}, Y_{m+2}, \dots, Y_{m+n}\}$ 。中間解析時の m 症例のデータの平均値を x_m 、最終解析時の $m+n$ 症例のデータの平均値を X_{m+n} とする。このとき帰無仮説 $H_0: \delta = 0$ を検定するための統計量は $Z_m = x_m \sqrt{m} / \sigma$ 、 $Z_{m+n} = x_{m+n} \sqrt{m+n} / \sigma$ となる。

<中止基準の設定>

将来有意であることを示すことが難しいとき、試験を早期中止する場合を考える。

—ベイズ流予測確率(無情報事前分布)の算出式—

中間解析時点のパラメータ δ に関する情報によって無情報事前分布を更新し、最終解析時に有意となる確率 P_m を求める。中間解析時点での試験の進捗の割合を f とすると ($f = m/(m+n)$)、

$$P_m = \Phi \left\{ \frac{z_m - \Phi^{-1}(1-\alpha)\sqrt{f}}{\sqrt{1-f}} \right\}$$

H_0 を棄却する確率を算出し、これがある基準値 $1-\gamma$ より低いとき試験を早期中止するという基準を構成できる。Spiegelhalter, Freedman, Parmar(1994)は楽観的な事前分布を設定したベイズ流予測確率で判断を行うことを薦めている。なお、事前分布を設定したベイズ流予測確率の算出式は同論文を参照頂きたい。関心のあるベイズ流予測確率として、無情報事前分布、楽観的な事前分布、悲観的な事前分布を設定したときの数値を評価している。

Spiegelhalter, Freedman, Parmar (1994)は、ベイズ流予測確率で中間解析を行ったとき α エラーを制御するために、悲観的な事前分布を設定する方法も紹介している。

—条件付き検出力の算出式—

z_m の条件付きで、最終解析時に有意となる確率 $C_m(\delta)$ を求める。

$$C_m(\delta) = 1 - \Phi \left\{ \frac{\Phi^{-1}(1-\alpha) - z_m\sqrt{f} - \delta \frac{\sqrt{m+n}}{\sigma}(1-f)}{\sqrt{1-f}} \right\} \quad (\delta \text{ は薬剤効果})$$

H_0 を棄却する確率を算出し、条件付き検出力 ($\delta = \delta_1$: 当初想定した薬剤効果)がある基準値 $1-\gamma$ より低いとき試験を早期中止する(Lan, Simon, Halperin(1982))。ベイズ流予測確率の算出方法と異なり、 z_m は既に得られたデータとしては利用されるが、残り n 例の挙動には影響を与えない。また、逆に H_0 を棄却することを目的に試験を早期中止するときは $\delta = \delta_0$ (帰無仮説)を設定する。

関心のある条件付き検出力として、 $\delta = \delta_1$: 想定した薬剤効果、 $\delta = \delta_0$: 帰無仮説、 $\delta = \hat{\delta}$: 中間解析時の平均値、 $\delta = \hat{\delta}_U$: 中間解析時の平均値の信頼上限、 $\delta = \hat{\delta}_L$: 中間解析時の平均値の信頼下限の値などが用いられる。

条件付き確率 ($\delta = \delta_1$)での判定を複数回行ったとき β エラーの上限は β/γ となる(Lan, Simon, Halperin(1982))。

早期中止の検討に際しては、基準値を定めず関心のある複数の数値を算出し、独立データモニタリング委員会が総合的に判断する場合もあるが、企業の方針も重要な判断材料であるため、早期中止の基準はプロトコールに明記する方が望ましい。

なお早期中止の基準値としては5%、10%、20%として設定し、考察していることが多い。

生存時間解析の例

生存時間をエンドポイントとした例を示す。A 群、B 群 の 2 群間比較(優越性試験)で対数ハザード比 0.3 を想定し、 α エラー2.5%(片側)、 β エラー20%のもとで総イベント数 350 例を見積もったとする。イベント数 175 例集めたときに中間解析を行い、そのときの対数ハザード比は $x_m=0.10$ であった(図 1)。このときの「ベイズ流予測確率」と「条件付き検出力」は次のようになる。

- ・ ベイズ流予測確率(無情報事前分布) 0.153
- ・ 条件付き検出力($\delta = \delta_0=0.30$) 0.450
- ・ 条件付き検出力($\delta = \hat{\delta}=0.10$) 0.074

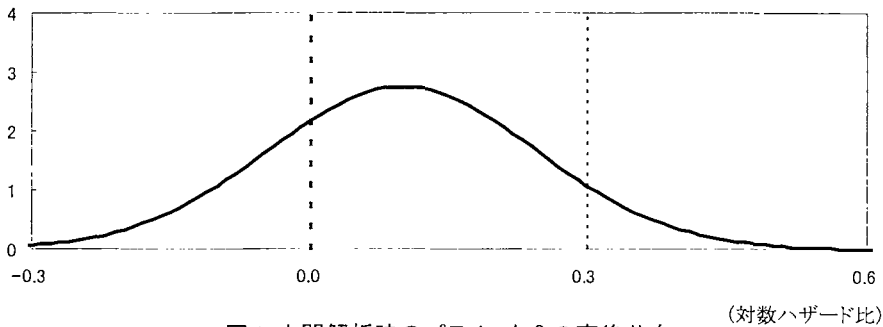


図 1 中間解析時のパラメータ δ の事後分布

Spiegelhalter, Freedman, Parmar (1994) はベイズ流による解析の紹介で、生存時間解析の群間比較(両群とも例数は同じ、対数ハザード比 $\neq 0$)として δ を対数ハザード比、 $x_m = 4L/m$ (L は Log-rank スコア)、 $\sigma^2=4$ とする近似式を用いている。このとき m, n は症例数ではなくイベント数となる。条件付き検出力の算出においてもこの方法を用いた。

また、条件付き検出力($\delta = \hat{\delta}$)を中止基準として利用することは妥当ではないかもしれないが、比較のため示した。

中間解析時に得られたデータの対数ハザード比 x_m は 0.1 だったが、この値を変化させたとしてベイズ流予測確率と条件付き検出力の値を計算した(図 2)。

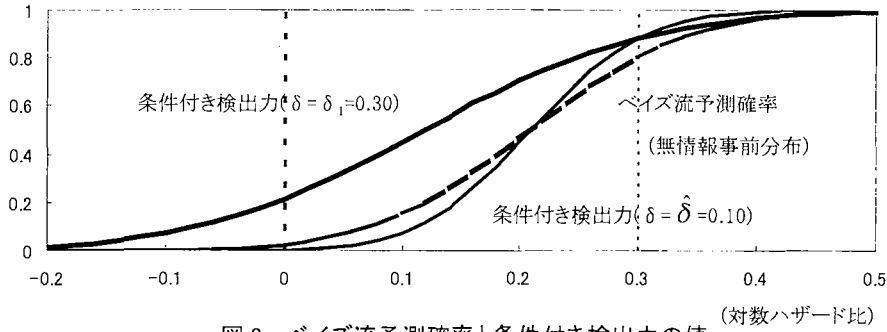


図2 ベイズ流予測確率と条件付き検出力の値

ベイズ流予測確率(無情報事前分布)と条件付き検出力($\delta = \hat{\delta} = x_m$)は非常に近い値をとることが図から読み取れる。先の設定で、条件付き検出力($\delta = \hat{\delta} = x_m$)を求めて分布関数の中身を \sqrt{f} 倍するとベイズ流予測確率(無情報事前分布)と同じ式になり(Jennison, Turnbull(2000))、また、同様に先の設定で、ベイズ流予測確率を算出する際の予測分布を求めるところでパラメータ δ の分布を $\delta = x_m$ の1点に集中させると、ベイズ流予測確率(無情報事前分布)と条件付き検出力($\delta = \hat{\delta} = x_m$)は等しくなる。

性能評価

ベイズ流予測確率と条件付き検出力の性能評価を行った。中間解析1回、最終解析1回で、中間解析時に将来有効性を示すことが難しいか否か検討を行うケースを想定して、SASのPROBBNRM関数(二変量正規分布の分布関数)を用いてプログラムを作成した。

早期中止を判断する基準値を $1 - \gamma$ とすると、ベイズ流予測確率の式より $z_m < \Phi^{-1}(1 - \gamma)\sqrt{1 - f} + \Phi^{-1}(1 - \alpha)\sqrt{f}$ を満たすとき、試験を中止することになる。

期待する仮説($\delta_1 = (z_\alpha + z_\beta)\sqrt{\frac{\sigma^2}{m+n}}$)のもとで、 $z_m \sim N((z_\alpha + z_\beta)\sqrt{f}, 1)$ 、 $z_{m+n} \sim N((z_\alpha + z_\beta)h, 1)$ 、

$Cov(z_m, z_{m+n}) = \sqrt{f}$ となることからPROBBNRM関数で次の確率が算出できる。

- ・中間解析時に早期中止せず、最終解析で有意となる確率
- ・中間解析時に早期中止せず、最終解析で有意とならない確率
- ・中間解析時に早期中止し、(もし継続していたら)最終解析で有意となる確率
- ・中間解析時に早期中止し、(もし継続していたら)最終解析で有意とならない確率

$z_\alpha = \Phi^{-1}(1 - 0.025)$ 、 $z_\beta = \Phi^{-1}(1 - 0.200)$ として、これらの数値から表1、表2を作成した。

表1 性能：想定通りの薬剤効果がある($\delta = \delta_1$)とき

有意差の出る確率0.8000(検出力)の内訳を示した。ベイズ流予測確率で基準値0.05で判断を行うと、0.0071で試験を中止し、0.7929で最終解析まで行い有意差ありとなる。

中間解析実施時点 f	基準値	ベイズ流 予測確率	条件付き 検出力 ($\delta = \delta$)	条件付き 検出力 ($\delta = \delta_1$)
0.50	0.05	0.0071	0.0207	0.0001
	0.10	0.0162	0.0342	0.0004
	0.20	0.0385	0.0588	0.0026
0.30	0.05	0.0109	0.0517	0.0000
	0.10	0.0246	0.0729	0.0000
	0.20	0.0574	0.1063	0.0004

表2 性能：薬剤効果がない($\delta = \delta_0 = 0$)とき

有意差の出ない確率0.9750(1-有意水準)の内訳を示した。ベイズ流予測確率で基準値0.05で判断を行うと、0.5875で試験を中止し、0.3875で最終解析まで行い有意差なしとなる。

中間解析実施時点 f	基準値	ベイズ流 予測確率	条件付き 検出力 ($\delta = \delta$)	条件付き 検出力 ($\delta = \delta_1$)
0.50	0.05	0.5875	0.7117	0.1965
	0.10	0.6829	0.7691	0.3118
	0.20	0.7824	0.8283	0.4795
0.30	0.05	0.3803	0.6221	0.0060
	0.10	0.4988	0.6822	0.0250
	0.20	0.6405	0.7481	0.0989

表には「条件付き検出力」でもっとも関心の高いと思われる $\delta = \hat{\delta}$ を中止基準としたときの結果も含め、正規の中止基準である $\delta = \delta_1$ としたときの結果も示した。ベイズ流予測確率は楽観的な事前分布を設定する方法もあるが、簡便さから無情報事前分布に基づく結果を示した。

中間解析の実施時点 f は 0.5 のケースと 0.3 のケースの 2 つとした。f=0.3 は、ベイズ流予測確率と条件付き検出力 ($\delta = \hat{\delta}$) において、試験開始直後の少ないデータから判断を行うことについて、その危険性の有無を確認するために設定した。

あと、「薬剤効果が想定の中点」のときについても表を作成した(表 3、表 4)。

表3 性能：薬剤効果が想定の中点($\delta = \delta_1/2$)のとき 1

有意差の出る確率0.2880の内訳を示した。ベイズ流予測確率で基準値0.05で判断を行うと、0.0054で試験を中止し、0.2826で最終解析まで行い有意差ありとなる。

中間解析実施時点 f	基準値	ベイズ流 予測確率	条件付き 検出力 ($\delta = \delta$)	条件付き 検出力 ($\delta = \delta_1$)
0.50	0.05	0.0054	0.0150	0.0001
	0.10	0.0119	0.0239	0.0003
	0.20	0.0267	0.0393	0.0021
0.30	0.05	0.0071	0.0306	0.0000
	0.10	0.0153	0.0419	0.0000
	0.20	0.0337	0.0590	0.0003

表4 性能：薬剤効果が想定の半分($\delta = \delta_1/2$)のとき 2
 有意差の出ない確率0.7120の内訳を示した。ベイズ流予測確率で基準値0.05で判断を行うと、0.2159で試験を中止し、0.4961で最終解析まで行い有意差なしとなる。

中間解析実施時点 f	基準値	ベイズ流予測確率	条件付き検出力 ($\delta = \hat{\delta}$)	条件付き検出力 ($\delta = \delta_1$)
0.50	0.05	0.2159	0.3197	0.0325
	0.10	0.2929	0.3792	0.0689
	0.20	0.3942	0.4505	0.1468
0.30	0.05	0.1352	0.2967	0.0005
	0.10	0.2065	0.3475	0.0032
	0.20	0.3117	0.4094	0.0197

考察

試験が薬剤の有効性を示すことが目的であることから考えて、表 1 で将来有効性の示せるデータを 5%以上の割合で早期中止させる基準は問題があると考えた。条件付き検出力($\delta = \hat{\delta}$)の全ての基準値とベイズ流予測確率の基準値 0.20 はこれに該当する。条件付き検出力($\delta = \hat{\delta}$)は関心の高い数値だが、中止基準としての利用は難しいと思われた。ベイズ流予測確率の基準値 0.20 は試験の開始直後での利用は注意が必要である。これらは検出力を大きく低下させる場合がある。

表 2 から考察し、ベイズ流予測確率、条件付き検出力($\delta = \delta_1$)の基準値 0.20 については、「薬剤効果がない」ときの早期中止の性能は高いと思われた。試験の半分近くを中止することができる。

すこし読み取りづらいが表 3 と表4からも同様のことが見てとれる。

以上のように中間解析が 1 回の場合は SAS の PROBBNRM 関数を用いて容易に考察を行うことが出来る。なお、 α エラーと β エラーを保持することに主眼をおいた手法は Jennison,Turnbull(2000)に紹介がある。

謝辞

本稿の作成について浜田知久馬助教授(東京理科大学)に助言を頂きました。御礼申し上げます。

参考文献

- 1) Jennison,C. and Turnbull,B.W.(2000), Group Sequential Methods with Applications to Clinical Trials, Chapman & Hall/CRC
- 2) Lan,K.K.G., Simon,R. amd Halperin,M.(1982), Stochastically curtailed tests in long-term clinical trials., Commun.Statist.C,1,207-219
- 3) Spiegelhalter,D.J., Freedman,L.S. and Parmar,M.K.B.(1994), Bayesian approaches to randomized trials., J.R.Statist.Soc.A,157,357-416

- 4) 医薬審 第 1047 号, 「臨床試験のための統計的原則」について, (1998 年 11 月 30 日)
- 5) 宇野一・松井茂之・小山鴨之(2000), 中間解析におけるベイズ流アプローチ: 最近の理論的展開, 計量生物学, 21 巻特集号, 125-149

使用したプログラム

```

/* 変数の説明
P1 : 中間解析時に早期中止せず、最終解析時に有意差あり
P2 : 中間解析時に早期中止、最終解析時に有意差あり
P3 : 中間解析時に早期中止せず、最終解析時に有意差なし
P4 : 中間解析時に早期中止、最終解析時に有意差なし
P12 : P1+P2
P34 : P3+P4
P2_P12 : P2/P12
P4_P34 : P4/P34
METHOD : 0:ベイズ流予測確率 1:条件付き確率 ( $\delta = \delta^*$ ) 2:条件付き確率 ( $\delta = \delta_1$ )
*/

data wk1 ;
do delta=0 , (probit(0.975)+probit(0.80))/2 , (probit(0.975)+probit(0.80)) ;
do f=0.3 , 0.5 ;
do method=0,1,2 ;
do g=0.05,0.10,0.20 ;
if method=0 then z= probit(0.975)*sqrt(f)+probit(g)*sqrt(1-f) ;
if method=1 then z= probit(0.975)*sqrt(f)+probit(g)*sqrt(1-f)*sqrt(f) ;
if method=2 then z=( probit(g)*sqrt(1-f)+probit(0.975)
- (probit(0.800)+probit(0.975))*(1-f) )/sqrt(f) ;

p1=PROBNORM(-probit(0.975)+delta, -z+sqrt(f)*delta, sqrt(f)) ;
p2=PROBNORM(-probit(0.975)+delta)-p1 ;
p3=PROBNORM(-z+sqrt(f)*delta)-p1 ;
p4=1-p1-p2-p3 ;

p12=p1+p2 ;
p34=p3+p4 ;

p2_p12=p2/p12 ;
p4_p34=p4/p34 ;
output ;

end ;
end ;
end ;
end ;
run ;

proc print data=wk1 ; format _numeric_ 7.4 ; run ;

```

日本SASユーザー会 (SUGI-J)

Group Sequential 計画のためのパワーシミュレータの開発

○本田圭一* 太田裕二** 佐賀野修一** 田崎武信*

* 塩野義製薬株式会社 解析センター

** 住商情報システム株式会社 産業システム第二事業部

Power simulator for group sequential clinical trial design

Keiichi Honda*, Yuji Oota**, Shuichi Sagano** and Takenobu Tasaki*

* Biostatistics Dept., Shionogi Co., Ltd.

** Industrial Systems Div.2., Sumisho Computer Systems Co., Ltd.

要 旨

群逐次 (Group Sequential) 型の比較臨床試験のパワーを見通しよく評価するためのプログラムを開発した。中間解析ごとの標本サイズと有意水準を具体的に与えたもとの、パワーをシミュレーションの反復結果として求めた。このためのプログラムを、評価変数の分布のタイプ別に準備し、マクロ言語を利用して統合し、システム化を試みた。

キーワード： パワー、群逐次臨床試験、中間解析、マクロ言語

1. はじめに

比較臨床試験では、その試験が十分なパワーをもつかどうか、すなわち主要な評価変数のもとの薬剤効果の差を検出できるかどうか、が試験をデザインする段階での主たる関心事のひとつとなる。固定標本サイズの比較臨床試験のもとの、主要な評価変数の分布として単純な確率分布を想定できる場合、パワーを理論的に求めようと試みるであろう。しかし、有効な治療あるいは無効な治療と、許容できない有害事象を早期に発見するという倫理的な配慮から中間解析を計画し、群逐次 (Group Sequential) 型の試験デザインを採用する場合、パワーを理論的に求めることは簡単でない。そこで、リスクをマネジメントする観点から、群逐次型の試験デザインのもとも計算機の力を借りて、試験がもつパワーを手軽に計算するためのプログラムを開発し、システム化を目指す。

群逐次型の試験デザインとして設定した状況のもと、予想される結果の周辺でデータをランダムに発生させる。そして、このシミュレーションを十分な回数だけくり返し、理論パワーに代わる経験パワーを獲得する。プログラムは、評価変数の分布のタイプごとにマクロ化する。ここに評価変数の分布のタイプとして、2値、順序カテゴリカル値、正規連続値、2成分混合正規連続値、および生存時

間をとりあげる。中間解析における情報量の尺度としては、試験治療(対照も含む)が終了した症例の集積率を利用する。しかし、生存時間を評価変数とするもとは、症例の集積率を利用する場合のほかに、試験開始後の経過時間を利用する場合も考える。この後者の場合は、群逐次型のデザインである必要はなく、症例登録がすべて終わったのちに中間解析がおこなわれることをむしろ考えたい。中間解析における情報量として、試験治療が終了した症例の集積率を利用する試験での症例登録の進捗の典型例を図1(a)に、試験開始後の経過時間を利用する試験での症例登録の進捗の典型例を図1(b)に示す。図1(a)のタイプの試験は、図1(b)のタイプの試験に比べて一般的にエントリー期間が長いと考えられる。

開発したプログラムにおいて、群逐次型の試験デザインのもとで中間解析として想定している状況を2節で説明し、3節でシステムの入力パラメータと出力情報を示す。4節では、生存時間を評価変数としてシステムを実行した例を紹介する。5節では、バリデーションに対する考えかたについて述べる。

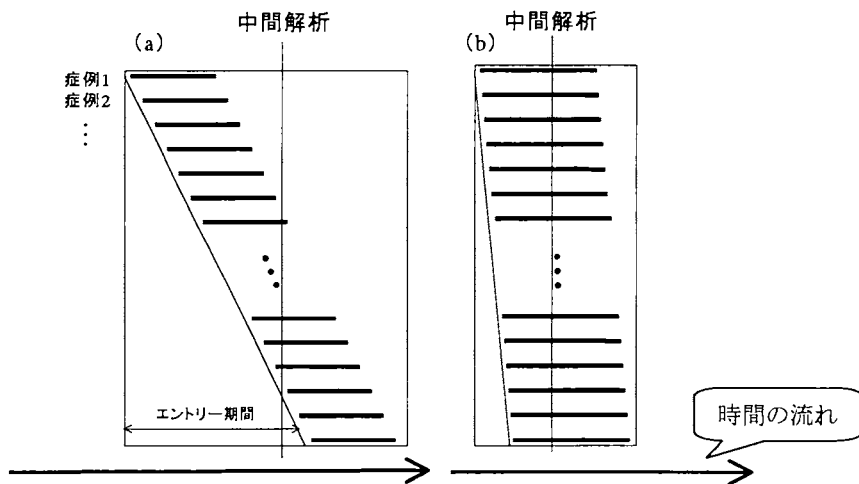


図1. 症例登録の進捗の様子: (a)試験治療が終了した症例の集積率を情報量として定義する場合の試験、(b)試験開始後の経過時間を情報量として定義する場合の試験

2. 群逐次型の試験デザインのもとで想定する状況

2.1. 想定する試験デザインのもとでの中間解析

中間解析が計画された状況で、試験群と対照群の2群比較試験を考える。比較のための主要な評価変数はここでは1つとする。比較試験の帰無仮説は2群の薬剤効果に差がないこと、対立仮説は2群の薬剤効果に差があることとする。すなわち両側検定を考える。

中間解析ごとに主要な評価変数に対する検定をくりかえす。そして、帰無仮説が棄却された場合に限り試験を中止し、それ以外では試験を継続する。したがって、帰無仮説を受容して試験を中止する場合、たとえば Enas and Offen(1993)の Early Stopping Procedure for Ineffectivenessなどを、ここでは扱わないことに注意する。

中間解析を実施すると、試験全体での第1種の過誤 α の膨張を抑えつつ、望ましいパワー $(1 - \beta)$ を獲得することが大事である。中間解析のデザインに関する理論的かつ技術的な議論は、たとえば試験全体の α を制御するために中間解析ごとに消費する α をどのように定めるか、あるいは同じことにつながるが、中間解析ごとの検定統計量に対する棄却限界値をどのように求めるか、すなわち中間解析ごとの有意水準をいくりに設定するか、という試験の中止規準を導出することに関して活発である。しかしながら、ここではそのような議論をおこなわない。なお、試験の中止規準の導出を含め、中間解析の最近の理論的な展開については、松井ほか(2000)と宇野ほか(2000)で詳述されており、勉強になる。

中間解析ごとに消費する α と中間解析ごとの有意水準は同じものではないが、両者には対応がある。つまり、ある試験において、それまでの各中間解析で使用した有意水準から、中間解析のおのおので消費した α を求めることができるし、その逆の流れで、すなわち消費した α から有意水準を導くことができる。ただし、計算が比較的容易なのは各中間解析での検定統計量が独立な関係にある場合である。本システムでは、各中間解析での検定統計量が独立であるかどうかの議論を回避するために、中間解析ごとに消費する α ではなく、中間解析ごとの有意水準を指定するように設計している。したがって、本システムは各中間解析での検定統計量が独立であるかどうかに関係なく、経験パワーを計算する。

各中間解析での検定統計量が独立な関係にあれば、有意水準ではなく、中間解析ごとに消費する α を利用してパワーを導くことも可能である。それは、たとえば Lan and DeMets(1983)の alpha spending function などを利用することで、適切に設定されていることが望ましいが、この設定の手続きの適切さはこのシステムのもとで前提条件ではない。設定が適切であったかどうかは、本システムによるシミュレーションの結果として、検定サイズを調べることによりおよそ把握することができる。

誤解を避けるために述べておくと、alpha spending function は中間解析で消費する α を事前に決めるために提案されたものではない。臨床試験を進めながら、中間解析を実施する回数や時期を柔軟に設定し、その都度、消費する α を決めるということが本来の用途である。うえでは、特定の条件下で臨床試験のパワーを読むために、その条件のひとつとなる、中間解析ごとに消費する α を、たとえば alpha spending function を利用して設定すると述べたにすぎない。

2.2. システムを利用するためのさらなる想定

中間解析の実施回数と実施時期については、前節のおわりに述べたとおり、本システムを利用するには定めておく必要があり、各中間解析までの1群あたりの累積標本サイズを与えることで両者を暗に指定する。試験のデザイン段階での利用を念頭においているので、標本サイズは2群で等しいことを前提とする。中間解析の実施時期は、試験治療が終了した症例の集積率に対応する。ただし、群逐次型の試験を必ずしも想定していない場合、すなわち本システムのもとでは、生存時間を評価変数とし、試験開始後の経過時間を情報量の尺度とする場合については例外となる。この場合は、各中間解析の時点として、試験開始後の経過時点を指定する。標本サイズも与えるが、それは最終解析まで試験が継続した場合の標本サイズに相当する。

標本サイズと同様に、中間解析ごとに有意水準を与える。ここに、指定する値は、中間解析ごとに消費する α でもなく、中間解析ごとの検定統計量に対する棄却限界値でもないことに注意する。

評価変数の分布のタイプとしては、2値、順序カテゴリカル値、正規連続値、2成分混合正規連続値、および生存時間のいずれかを選択できる。いずれのタイプの評価変数であっても、試験群と対照群の評価変数に関する分布について、想定する特性値を指定する。たとえば、評価変数の分布のタイプとして2値を選択すると、試験群での想定有効率と対照群での想定有効率を指定することが必要になる。2群の評価変数の分布において、差がない状況を想定すれば検定サイズを、差がある状況を想定すればパワーを試算することになる。

これらのことを想定したもとの、シミュレーションを十分な回数だけ、たとえば 10000 回くらい、くり返す。その結果として、実際に薬剤効果の差を検出できた回数の割合が経験パワーであり、我々の知りたい理論パワーをおよそ代替するものであると考えたい。

これまでに述べた想定するおもな状況を図 2 に示す。

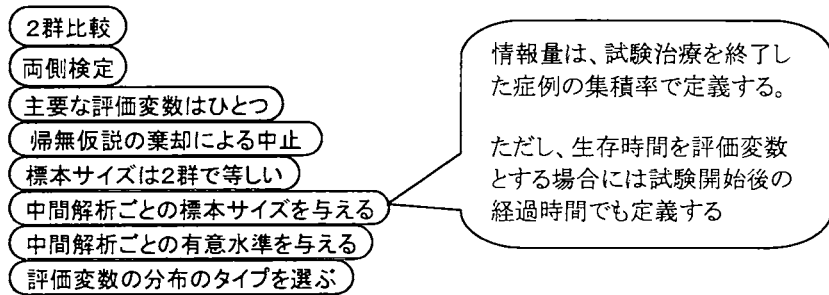


図 2. 想定する状況

2.3. 生存時間を評価変数とする場合のさらなる想定

ある事象(たとえば、治癒、再発、死亡、など)が生起するまでの時間を評価変数とする。事象が生起するまでの時間が指数分布 $\lambda_E \exp(-\lambda_E t)$ に従う場合、長さ T の試験期間において事象が生起する症例の割合 π_E は $1 - \exp(-\lambda_E T)$ で与えられる。逆に、 λ_E は π_E から $\lambda_E = -\log(1 - \pi_E)/T$ で与えられる。

脱落症例の発生を考慮する。脱落症例についての評価変数の値は、脱落時点での中途打ち切り観測値とする。脱落は試験期間においてランダムに、すなわち指数分布に従って生起すると想定する。ここでの脱落率は、試験期間において脱落し、評価変数が中途打ち切りで観測される症例の割合である。長さ T の試験期間における脱落率を π_D とすると、脱落が生起するまでの時間を表す指数分布のパラメータ λ_D は $\lambda_D = -\log(1 - \pi_D)/T$ で与えられる。

全症例に共通の最長追跡期間を想定する。それは、ここでは試験治療が継続される最長の期間を意味する。事象が生起するまでの時間が、全症例に共通の最長追跡期間を超える場合、評価変数の値はその期間での中途打ち切り観測値とする。

エントリー期間を与える。この期間の中で患者は一様分布に従ってランダムにエントリーされると想定する。

3. プログラムの仕様

開発したプログラムでは、中間解析ごとの標本サイズを与えて、試験がもつパワーを見積もることができる。プログラムの基本的な入力パラメータは、1群あたりの標本サイズ、試験群と対照群の評価変数に関する情報、検定法、有意水準である。備えていない検定法でも、User 定義検定のプログラム名とパス名を指定すれば利用することができる。これらのパラメータについての指定方法を表 1 に示す。そのうち、試験群と対照群の評価変数に関する情報の詳細については表 2 に、選択可能な検定法については表 3 に示す。

表 1. 入力パラメータ

パラメータ	指定方法
1群あたりの標本サイズ	中間解析までの累積標本サイズをスペース区切りで指定 (注意)生存時間で試験開始後の経過時間を情報量の尺度とする場合には、試験全体での1群あたりの標本サイズを指定
試験群と対照群の評価変数に関する情報	<表 2 に掲載>
検定法	<表 3 に掲載>
有意水準(両側)	中間解析ごとにスペース区切りで指定
User 定義検定のプログラム名とパス名	検定法において、User 定義検定を指定した場合に限定

表 2 において、生存時間その1の表記は、試験治療が終了した症例の集積率によって情報量を定義した場合、生存時間その2の表記は、試験開始後の経過時間によって情報量を定義した場合にそれぞれ該当する。後者の場合は、その2に加えて、その1についても指定が必要である。

出力情報を表 4 に示す。入力パラメータおよび解析時点別の経験パワーと試験全体での経験パワーは、評価変数の分布のタイプに関係なく、いずれの場合でも出力する。評価変数に関する2群の要約統計量も出力するが、これについては、評価変数の分布のタイプによって情報が異なり、2値では有効率、順序カテゴリカル値ではカテゴリーの相対頻度、正規連続値と2成分混合正規連続値では平均値と標準偏差、生存時間では事象生起率がそれぞれ出力される。さらに2成分混合正規連続値については、生成した各成分のデータの頻度プロットを出力する。

4. 実行例

実際に出力した例を表 5 に示す。評価変数の分布のタイプは生存時間、情報量の定義は試験治療が終了した症例の集積率に相当する。試験の途中でログランク検定を用いて 2 回の中間解析をおこなう。1 回目は 1 群あたりの標本サイズが 50 例のとき、2 回目は 90 例のとき、そして最終解析は 110 例のときである。各中間解析での有意水準はそれぞれ 0.01、0.025 とし、最終解析で 0.05 とする。最長追跡期間を 26 週とし、その期間における試験群での事象生起率を 35%、対照群での事象生起率を 57%と想定し、脱落症例は考慮しない。

上記の設定のもとで、シミュレーションを 10000 回くりかえした結果、1 回日の中間解析での経験

表 2. 試験群と対照群の評価変数に関する情報

評価変数の分布のタイプ	評価変数に関する情報
2値	試験群での想定有効率 対照群での想定有効率
順序カテゴリカル値	試験群での想定カテゴリー相対頻度% 対照群での想定カテゴリー相対頻度%
正規連続値	試験群での想定平均値 試験群での想定標準偏差 対照群での想定平均値 対照群での想定標準偏差
2成分混合正規連続値	試験群における第1成分の割合 試験群における第1成分の想定平均値 試験群における第2成分の想定平均値 試験群における第1成分の想定標準偏差 試験群における第2成分の想定標準偏差 対照群における第1成分の割合 対照群における第1成分の想定平均値 対照群における第2成分の想定平均値 対照群における第1成分の想定標準偏差 対照群における第2成分の想定標準偏差 (注) User 定義分布のプログラム名とパス名を指定することも可能
生存時間その1 <試験治療が終了した症例の集積率>	脱落症例を考慮するか否かの指定 2群に共通の想定脱落率 試験群での想定事象生起率 対照群での想定事象生起率 全症例に共通の最長追跡期間 (注) User 定義による事象生起までの時間分布のプログラム名とパス名を指定することも可能
生存時間その2 <試験開始後の経過時間> その1の情報を右の情報を追加指定	エントリー期間(一様分布) 最終解析を含む中間解析の実施時期(試験開始からの経過時間をスペース区切りで指定) (注) User 定義による事象生起までの時間分布のプログラム名とパス名を指定することも可能

表 3. 検定法

評価変数の分布のタイプ	検定法
2値	カイ二乗検定、Fisher の直接確率計算法、User 定義検定
順序カテゴリカル値	Wilcoxon 順位和検定、累積カイ二乗検定、ロジスティック回帰分析、User 定義検定
正規連続値	Welch 検定、Student のt検定、Wilcoxon 順位和検定、User 定義検定
2成分混合正規連続値	Welch 検定、Student のt検定、Wilcoxon 順位和検定、User 定義検定
生存時間	ログランク検定、User 定義検定

表 4. 出力情報

入力パラメータ	
試験群と対照群の *** の要約統計量(平均値、標準偏差、最大値、最小値)	
***→	有効率<2値>、カテゴリーの相対頻度<順序カテゴリカル値>、 平均値と標準偏差<正規連続値、2成分混合正規連続値>、 事象生起率<生存時間>
生成したデータの頻度プロット<2成分混合正規連続値>	
解析時点別の経験パワーと試験全体での経験パワー	

表 5. 実行例

シミュレーション実行パラメータ							1
脱落症例を考慮しない場合							
解析回数	: 3回						
標本サイズ							
中間解析1回目	: 50						
中間解析2回目	: 90						
最終解析時	: 110						
試験回数	: 10000回						
試験群の想定事象生起率	: 35%						
対照群の想定事象生起率	: 57%						
試験期間	: 26						
検定法	: Log-Rank検定						
有意水準							
中間解析1回目	: 0.01						
中間解析2回目	: 0.025						
最終解析時	: 0.05						
時点別の事象生起率の平均							2
MEANS プロシジャ							
変数	ラベル	N	平均値	標準偏差	最小値	最大値	
Key1_01	中間解析1回目の試験群事象生起率	10000	34.9184000	6.8011445	12.0000000	62.0000000	
Key2_01	中間解析1回目の対象群事象生起率	10000	56.9286000	7.0614367	32.0000000	84.0000000	
時点別の事象生起率の平均							3
MEANS プロシジャ							
変数	ラベル	N	平均値	標準偏差	最小値	最大値	
Key1_02	中間解析2回目の試験群事象生起率	6225	36.5599286	4.6903497	21.1111111	58.8888889	
Key2_02	中間解析2回目の対象群事象生起率	6225	55.3222668	4.8176503	40.0000000	72.2222222	
時点別の事象生起率の平均							4
MEANS プロシジャ							
変数	ラベル	N	平均値	標準偏差	最小値	最大値	
Key1_03	最終解析時の試験群事象生起率	2084	39.0036643	3.8755618	27.2727273	56.3636364	
Key2_03	最終解析時の対象群事象生起率	2084	52.8398185	3.8806390	40.0000000	65.4545455	
シミュレーション10000のときのPowerの内訳							5
	試験時期	有意数	累積有意数	Power			
	中間解析1回目	3775	3775	37.75			
	中間解析2回目	4141	7916	79.16			
	最終解析時	1307	9223	92.23			

パワーはおよそ 37.75%、2 回目の中間解析までの累積経験パワーはおよそ 79.16%となり、最終解析までに得られるこの臨床試験がもつ経験パワーはおよそ 92.23%になった。

5. バリデーションに対する考え

3節で述べたプログラムの仕様のもとで、本システムはシミュレーションにより経験パワーを試算する。この一連の手続きが妥当であったか否かを判断することは、かなり難しい。理論的にアプローチできるようにあれば、シミュレーションに頼るパワー計算を考えることはなかったかもしれないからである。そうすると、別のシミュレーションの結果と比較して、同様の結果が得られるかどうかという手段が

考えられる。そして、その別のシミュレーションの結果というのが文献などで公表され、妥当性について検証されたものであれば、なおのこと望ましい。

生存時間を評価変数とした場合に限ると、Gu and Lai(1999)で、シミュレーションによりパワーを計算するためのソフトウェアが提供されている。そこでは、入力を必要とする情報が詳細にまで及び、設定が複雑である。それらの情報をおおまかに分類すると、ルックに関する情報、症例の集積に関する情報、中止境界を生成する方法、検定統計量の選択、生存関数についてのベースライン分布と対立分布の規定、中途打ち切り分布の規定、不遵守率の規定などであり、それぞれについて、細目が用意されている。これらのなかには本システムで考慮していない情報がいくつも含まれており、バリデーションも一筋縄ではいかないというのが実状である。

生存時間を評価変数とした場合に限らず、本システムのバリデーションをいかに実施するかが今後の課題である。

6. おわりに

評価変数の分布のタイプとして、2値、順序カテゴリカル値、正規連続値、2成分混合正規連続値、生存時間のいずれであろうと、中間解析ごとの標本サイズと有意水準を与えて、群逐次型の臨床試験がもつ早期中止の場合の経験パワーと最後まで試験を継続した場合の経験パワーの計算を試みた。このシステムは臨床試験のデザイン段階で利用されることを意図して開発したが、用途をそこに限る必要はない。進行中の臨床試験における中間解析の断面で、2群の評価変数の分布に関する情報を入手することが許されていれば、最後まで試験を継続した場合の臨床試験がもつパワーを、より確実性のあるシミュレーション結果の経験パワーでもって占うことができると考える。

中間解析を備えた臨床試験、とくに群逐次型の試験が議論される機会は、最近の医薬品開発の場において確実に増えてきていると実感している。このことが動機となって本システムを開発したが、2節で述べたように、本システムが想定している状況は必ずしも十分とはいえず、ある程度までに限定されている。将来には想定範囲を拡大する必要に迫られるかもしれない。そのことも視野におさめながら、利便性をよくするために、本システムのSAS/PH-Clinicalへの実装を目指すつもりである。

参考文献

- Enas,G.G. and Offen,W.W.(1993).A simple stopping rule for declaring treatment ineffectiveness in clinical trials. *Journal of Biopharmaceutical Statistics*.3(1),13-22.
- 松井茂之・宇野 一・小山暢之 (2000).中間解析における頻度論的アプローチ: 最近の理論的展開. *計量生物学*,21,87-124.
- 宇野 一・松井茂之・小山暢之 (2000).中間解析におけるベイズ流アプローチ: 最近の理論的展開. *計量生物学*,21,125-149.
- Lan,K.K.G. and DeMets,D.L.(1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70,659-663.
- Gu,M. and Lai,T.L.(1999).Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analysis. *Controlled Clinical Trials*, 20,423-438.

日本S A Sユーザー会 (S U G I - J)

投与期間別の副作用発現率を解釈するために

○古川 雅史, 片山 和夫, 田崎 武信

塩野義製薬株式会社

解析センター

On an approach useful in interpreting the dosing period-specific incidence of adverse drug reactions

Masashi Furukawa, Kazuo Katayama, Takenobu Tasaki

Biostatistics Department, Shionogi & Co., Ltd.

要 旨

医薬品の市販後調査における目的の一つに、副作用の発現に影響を及ぼす要因の探索がある。例えば、男女間で副作用発現率に差があれば、性を副作用発現の影響要因のひとつとして疑うことができる。副作用は服用しはじめたところに現れるのか、それとも長期に服用することで現れるかを把握することも必要である。このため、投与期間が副作用の発現に及ぼす影響を定量的に評価しなければならない。ところが、投与期間別の副作用発現率は曲者である。ここでは、生存時間データの解析のアプローチを適用し、投与期間別の副作用発現率をハザード関数として表現することで投与期間と副作用の発現との関係を解釈する。

キーワード： ハザード関数, 市販後調査, 適合度検定

1. はじめに

医薬品の安全性を定量的に表現するものとして副作用発現率が一般的である。医薬品の使用成績調査で行うデータ解析の目的の一つに、副作用の発現に影響を及ぼす要因の探索がある。この目的を遂行する第 1 歩として、例えば性別に計算した副作用発現率が均一でなければ、性を副作用発現の影響要因として疑うことになる。このような影響要因の探索において、慣習的にその要因を構成するカテゴリーで層別し、順序関係をもたないカテゴリー、あるいは 2 値カテゴリーで観測されるものであれば、カテゴリー間の副作用発現率の均一性がカイ 2 乗検定や直接確率計算法で検定されてきた。そして、順序関係をもつカテゴリーで観測されるものであれば、均一性の検定に加えて線形トレンドの有意性検定が、ふつうは Cochran-Armitage 検定のみが行われてきた。このような解析は影響要因の探索で初期のアプローチとして妥当であろう。しかしながら、投与期間の影響を調べるとき、投与期間別の副作用発現率は曲者である。投与期間をなぜ特別な項目として取り上げざるを得ないのかを以下で説明

する。なお、総投与量別の副作用発現率も投与期間と同じ考え方が通用する。

2. 投与期間別副作用発現率

投与期間別の副作用発現率の求め方を説明するにあたって、図1に示す架空の患者10例における投薬と副作用の発現状況を使用する。図1はGoldman(1992)でイベントチャートとよばれている。ここでは、投与期間を等間隔に投薬開始から順に期間1、期間2、…、期間6の6カテゴリーに区分している。○は投薬が終了または中止されたことを表し、×は副作用が発現したことを表す。したがって、患者1では期間6で投薬が終了し、全投与期間にわたって副作用は発現しなかったことがわかる。患者3では期間1で副作用が発現して投薬が中止された。そして、患者4では期間1で副作用が発現しても投薬が継続され、期間4で投薬が終了した。

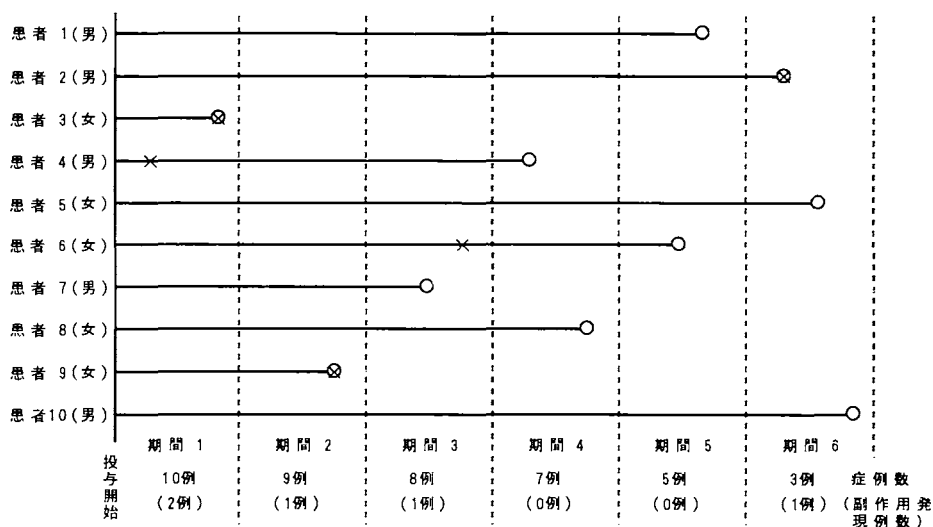


図1. 投与期間と副作用の発現状況

図1では患者の性もあわせて示していることから、最初に、性別の副作用発現率を考える。性別の副作用発現率を表1に示す。性の場合、男と女のカテゴリーで患者に重複は起こりえない。男と女のように、患者集団がカテゴリー間で異なる項目では、カテゴリー別の副作用発現率が一定であるか否かを検定することで、その項目が影響要因であるかどうかを調べることができる。この検定にカイ2乗検定や直接確率計算法が用いられる。要因が順序カテゴリーで構成される場合には、順序カテゴリーに伴う副作用発現率の線形トレンドを評価することもできる。

表 1. 性別の副作用発現状況

性	例数	副作用発現例数	副作用発現率
男	5	2	40.0%
女	5	3	60.0%

つぎに、投与期間カテゴリー別の副作用発現率を考える。表 2a の投与期間別副作用発現率は、投薬が開始から当該の期間カテゴリーまで行われたすべての患者を対象にして、その期間カテゴリーの間に副作用が発現した患者の割合で求めている。参考までに、投与期間別の副作用発現率というタイトルから想像する集計結果は表 2b のようなものである。

表 2a. 投与期間別の副作用発現率

投与期間	調査例数	副作用発現例数	副作用発現率
期間 1	10	2	20.0%
期間 2	9	1	11.1%
期間 3	8	1	12.5%
期間 4	7	0	0.0%
期間 5	5	0	0.0%
期間 6	3	1	33.3%

表 2b. そのタイトルから想像する投与期間別の副作用発現率

投与期間	調査例数	副作用発現例数	副作用発現率
期間 1	1	1	100.0%
期間 2	1	1	100.0%
期間 3	1	0	0.0%
期間 4	2	1	50.0%
期間 5	2	1	50.0%
期間 6	3	1	33.3%
全体	10	5	50.0%

表 2a において、投与期間の各カテゴリーに属する患者集団は、常に、時間が先行するカテゴリーの患者集団の一部となっている。このため、投与期間別の副作用発現率を比較するのに、カイ 2 乗検定や直接確率計算法を適用することができない。実際に、そうした検定は行われていないはずである。つまり、各カテゴリーに属する患者集団がカテゴリー間で異なるような性に代表される項目と、カテゴリー別患者集団が重複する投与期間に代表される項目とでは、副作用発現率の解釈や検定の方法を変更する必要がある。後者の項目でのカテゴリー別の副作用発現率は条件付きの副作用発現率であり、それは、生存時間データの解析の分野で、ハザードと呼ばれているものに相当する。表 2a における副作用発現率の求めかたは生存時間データをグループ(カテゴリー)分けする仕方に類似しているが、副作用が発現しても投与が継続された 2 例(患者 4 と患者 6)が表 2a では生かされていることに注意したい。すなわち、表 2a では期間 1 において、症例数は 10、副作用発現例数は 2 である。そして、期間 2 で

の症例数は9となっている。生存時間データであれば、期間1で副作用が発現した2例は、期間2に移行しないので症例数は8となる。

ところで、図1において副作用発現または投与終了(中止)までの日数は表3のようであったとする。

表3. 副作用発現までの日数

患者	日数	副作用発現の有無	患者番号	日数	副作用発現の有無
1	33	なし	6	19(32)	あり
2	37	あり	7	17	なし
3	6	あり	8	27	なし
4	2(23)	あり	9	12	あり
5	39	なし	10	41	なし

注) 表中において、患者4と6は副作用が発現したにもかかわらず、投与が継続された症例である。これらの症例について、括弧内で与えた日数は最終投与期間を表している。

副作用の発現しなかった患者についての投与日数を中途打ち切り観測値と取り扱って、累積非副作用発現率と累積副作用発現率を推定するため、生存時間データの解析で用いられるKaplan-Meier法を適用できる。前者は生存時間解析の分野において累積生存率、後者は累積死亡率と呼ばれているものに相当する。表3のデータにKaplan-Meier法を適用した結果を表4と図2(実線)に示す。

表4. 日数データ(表3)に基づくKaplan-Meier推定値

日数	リスク下にある症例数	副作用発現例数	中途打ち切り例数	累積非副作用発現率	累積副作用発現率
0	10	—	—	100.0%	0.0%
2	10	1	0	90.0%	10.0%
6	9	1	0	80.0%	20.0%
12	8	1	0	70.0%	30.0%
17	7	0	1	70.0%	30.0%
19	6	1	0	58.3%	41.7%
27	5	0	1	58.3%	41.7%
33	4	0	1	58.3%	41.7%
37	3	1	0	38.9%	61.1%
39	2	0	1	38.9%	61.1%
41	1	0	1	38.9%	61.1%

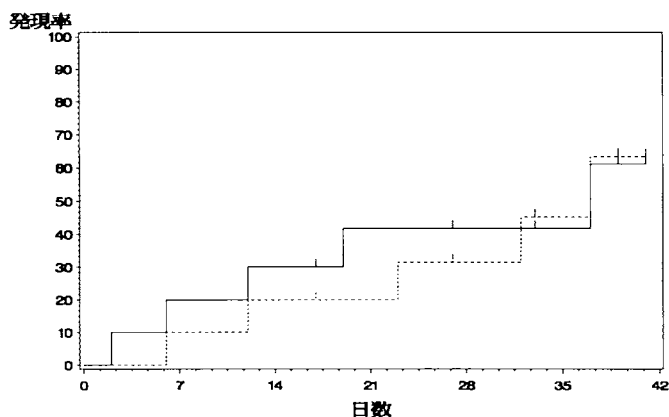


図 2. 累積副作用発現率

図 2 において点線は、患者 4 と患者 6 について表 3 の括弧内の日数を用いた場合の累積副作用発現率に対応する。いまの事例で、実線は点線に比べて、日数の早い時点で高く、遅い時点で低くなっている。

ところで、表 3 の実数を週単位でカテゴリー化したデータに Kaplan-Meier 法を適用した結果を表 5 に示す。表 4 とつきあわせると、いまの事例でカテゴリー化(日数の週区分)は大きく影響していないことがわかる。

表 5. カテゴリー化したデータに基づく Kaplan-Meier 推定値

週数	リスク下にある症例数	副作用発現例数	中途打ち切り例数	累積非副作用発現率	累積副作用発現率
0 週	10	—	—	100.0%	0.0%
1 週(7 日)	10	2	0	80.0%	20.0%
2 週(14 日)	8	1	0	70.0%	30.0%
3 週(21 日)	7	1	1	60.0%	40.0%
4 週(28 日)	5	0	1	60.0%	40.0%
5 週(35 日)	4	0	1	60.0%	40.0%
6 週(42 日)	3	1	2	40.0%	60.0%

生存時間データの解析法を用いるという観点からは、表 3 あるいは表 5 を出発点とすることが望ましいと考えられる。しかし、医療の現場では、副作用が発現しても投与が継続される患者がいるという状況から、カテゴリー別の患者集団の規定の仕方は表 2a が当を得ていると考えられる。いずれを対象として評価するのが適切かは悩ましい問題である。

3. 投与期間別の副作用発現率の評価方法

表 2a の副作用発現率はハザードと見なせることから、投与期間にかかわらず副作用発現率が一定か否かの問題は、ハザードが投与期間にかかわらず一定か否かの問題に置き換えて

扱うことができる。さらに、ハザードが投与期間にかかわらず一定か否かの問題は、副作用発現までの時間分布が指数分布に従うか否かで確かめることができる。したがって、帰無仮説は「副作用発現までの時間は指数分布に従う」で、これは「投与期間の各カテゴリー間でのハザードは一定である」と読み替えることができる。ここでの帰無仮説の評価は、生存時間データの解析法の一つである加速モデルのあてはめが利用できる。具体的に SAS の LIFEREG プロシジャを用いた指数分布の適合度検定により行える。なお、この方法は、大橋・浜田(1995)の 4.1 章における「区分指数モデルによるグループ化された生存時間データの解析」に相当している。

表 2a における調査例数、副作用発現例数、副作用発現率をそれぞれ、患者集団、事象発現例数、ハザードと呼び代えて整理し直した表 6 に対して LIFEREG プロシジャを適用した、指数分布の適合度検定を行った結果、検定における p 値は 0.9594 であり、帰無仮説は棄却されなかった。すなわち、投与期間カテゴリー間でハザード、すなわち副作用発現率は一定であったと解釈できる。

表 6. 投与期間別のハザード

投与期間	患者集団	事象発現例数	打ち切り例数	ハザード
期間 1	10	2	0	20.0
期間 2	9	1	0	11.1
期間 3	8	1	1	12.5
期間 4	7	0	2	0.0
期間 5	5	0	2	0.0
期間 6	3	1	2	33.3

しかし、仮に、この適合度検定における p 値が 0.05 未満であったとしよう。この場合、仮説は棄却され、投与期間カテゴリー間で、ハザードは一定でないとして解釈できる。ただし、この方法では、いずれのカテゴリー間でハザードが異なっていたのか、および投与期間の増加にともなってハザードが増加あるいは減少しているのかといった情報を得ることはできない。

一方、副作用発現までの時間分布に指数分布ではなく、ワイブル分布を想定すれば、LIFEREG プロシジャを適用し、尺度パラメータの推定値と 95%信頼区間、および形状パラメータの推定値と 95%信頼区間を求めることができる。尺度パラメータの推定値の 95%信頼区間が 1 を含まない場合、副作用発現までの時間は指数分布に従うとの仮定は否定される。そして、形状パラメータの推定値を用いて、投与期間に伴うハザードの増加または減少といったトレンドを評価することができる。すなわち、形状パラメータの推定値が 1 より大きくその 95%信頼区間が 1 を含まないとき、ハザードは時間とともに増加し、1 より小さくその 95%信頼区間が 1 を含まないとき、ハザードは時間とともに減少していると解釈できる。

4. 適用例

いくつかの薬剤について、加速モデルに基づき、副作用発現までの時間の特徴を吟味した。分布が指数分布に従うか否かを評価した。A 剤について、評価症例数 319 例のうち、副作用が発現した症例数は 263 例で、副作用が発現しなかった打ち切り症例数は 56 例であった。この事例の特徴は、打ち切り症例数の評価対象症例数に占める割合が 17.6%で、打ち切り割合が比較的低いことである。A 剤ほど副作用発現率が高くなかった薬剤として、B 剤, C 剤, D 剤, E 剤, F 剤をとりあげた。これら計 6 薬剤について評価症例数, 事象発現症例数, 事象発現率, 打ち切り割合を表 7 に示した。なお, A 剤のみが抗癌剤であった。

表 7. 薬剤毎の評価症例数, 事象発現率, 打ち切り割合

薬剤	評価症例数	事象発現症例数	事象発現率(%)	打ち切り割合(%)
A剤	319	263	82.45	17.6
B剤	10,568	686	6.49	93.5
C剤	9,564	169	1.77	98.2
D剤	14,002	860	6.14	93.9
E剤	10,818	94	0.87	99.1
F剤	2,925	393	13.44	86.6

4.1 投与期間カテゴリー別の副作用発現率と 95%信頼区間

最初に、薬剤毎の投与期間別の副作用発現状況を概括するため、投与期間別の事象発現率と 95%信頼区間を算出し、表 8.1 から表 8.6 と図 8.1 から図 8.6 に示した。図における横軸の投与期間はそれぞれ表での各カテゴリーに対応している。いずれの薬剤においても投与期間カテゴリー別のハザードは一定でなかったことがうかがえた。また、薬剤毎には以下のことがうかがえた。

- ・A 剤:ハザードは投与期間の増加にともなって増加していた。
- ・B 剤:ハザードは、「3 日以下」から「15 日以上 21 日以下」までは投与期間の増加にともなって増加していたが、その後減少していた。
- ・C 剤:ハザードは、「7 日以下」でいくぶん高く、「8 日以上 14 日以下」から「22 日以上 28 日以下」では低く、その後「29 日以上 84 日以下」、「85 日以上 168 日以下」、「169 日以上」では高くなっていた。
- ・D 剤:ハザードと投与期間との間で一定の増減関係は見られなかった。
- ・E 剤:ハザードは、「3 日以下」から「15 日以上 21 日以下」までは投与期間の増加にともなって増加していたが、その後一度減少し、再度増加していた。
- ・F 剤:ハザードは、「3 日以下」から「15 日以上 21 日以下」までは投与期間の増加にともなって増加し、その後一旦減少し、「36 日以上 42 日以下」で再度増加していた。

表8.1 A剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 29日以下	319	62	19.44	15.24	24.21
2. 30日以上 59日以下	236	74	31.36	25.49	37.69
3. 60日以上 89日以下	140	59	42.14	33.85	50.77
4. 90日以上 416日	78	68	87.18	77.68	93.68

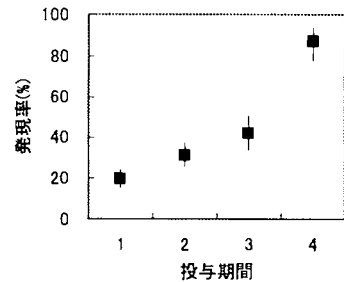


図8.1 A剤

表8.2 B剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 3日以下	10568	155	1.47	1.25	1.71
2. 4日以上 7日以下	9614	275	2.86	2.54	3.21
3. 8日以上 14日以下	4450	186	4.18	3.61	4.81
4. 15日以上 21日以下	978	54	5.52	4.17	7.14
5. 22日以上 28日以下	296	11	3.72	1.87	6.55
6. 29日以上 78日	113	5	4.42	1.45	10.02

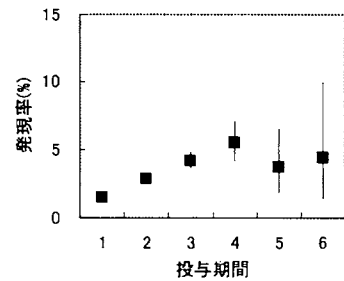


図8.2 B剤

表8.3 C剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 7日以下	9564	54	0.56	0.42	0.74
2. 8日以上 14日以下	7905	16	0.20	0.12	0.33
3. 15日以上 21日以下	6177	21	0.34	0.21	0.52
4. 22日以上 28日以下	5165	11	0.21	0.11	0.38
5. 29日以上 84日以下	4247	55	1.30	0.98	1.68
6. 85日以上 168日以下	924	10	1.08	0.52	1.98
7. 169日以上 402日	173	2	1.16	0.14	4.11

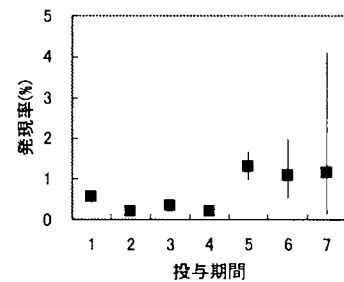


図8.3 C剤

表8.4 D剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 6日以下	14002	115	0.82	0.68	0.99
2. 7日以上 13日以下	13887	92	0.66	0.53	0.81
3. 14日以上 27日以下	13781	142	1.03	0.87	1.21
4. 28日以上 41日以下	13598	127	0.93	0.78	1.11
5. 42日以上 55日以下	13389	80	0.60	0.47	0.74
6. 56日以上 83日以下	13105	126	0.96	0.80	1.14
7. 84日以上 167日以下	11697	155	1.33	1.13	1.55
8. 168日以上 894日	2497	23	0.92	0.58	1.38

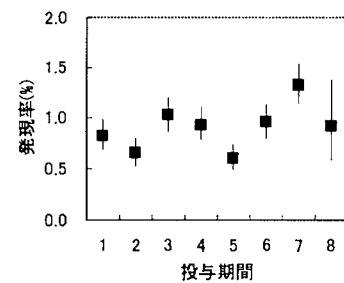


図8.4 D剤

表8.5 E剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 3日以下	10818	19	0.18	0.11	0.27
2. 4日以上 7日以下	10176	38	0.37	0.26	0.51
3. 8日以上 14日以下	3830	26	0.68	0.44	0.99
4. 15日以上 21日以下	870	7	0.80	0.32	1.65
5. 22日以下 28日以下	452	1	0.22	0.01	1.23
6. 29日以上 272日	208	3	1.44	0.30	4.16

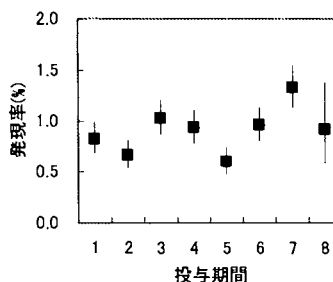


図8.5 E剤

表8.6 F剤の投与期間カテゴリー別の事象発現率と95%信頼区間

カテゴリー	症例数	事象 発現例	事象 発現率	下限	上限
1. 3日以下	2926	47	1.61	1.18	2.13
2. 4日以上 7日以下	2641	119	4.51	3.75	5.37
3. 8日以上 14日以下	1919	130	6.77	5.69	7.99
4. 15日以上 21日以下	757	55	7.27	5.52	9.35
5. 22日以上 28日以下	319	22	6.90	4.37	10.26
6. 29日以上 35日以下	161	7	4.35	1.77	8.75
7. 36日以上 42日以下	85	7	8.24	3.38	16.23
8. 43日以上 147日	51	6	11.76	4.44	23.87

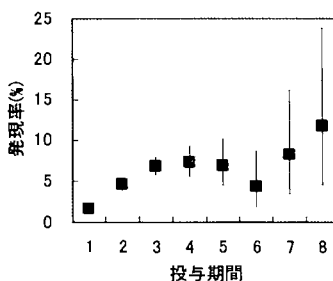


図8.6 F剤

4.2 カテゴリー化された投与期間への指数分布の適合度検定

A剤からF剤について、それぞれ表8.1から表8.6のカテゴリー化された投与期間別の副作用発現率にもとづき、指数分布の適合度検定を行った。ハザードが投与期間カテゴリー間で同じであったか否かを評価したことになる。ただし、一般的にいわれているカテゴリー化された生存時間データとは、2節に述べたように、患者集団の規定の仕方がいくぶん異なっている。

薬剤毎の指数分布の適合度検定におけるp値を表10に示した。A剤、B剤、D剤、F剤の4薬剤において、ハザードが投与期間カテゴリー間で一定であるとの仮説は、有意水準0.05で棄却された。C剤とE剤において、仮説は棄却されなかった。前者の4薬剤において、ハザードは投与期間カテゴリー間で一定でないといえたが、後者の2薬剤において、ハザードは投与期間カテゴリー間で一定でないとはいえなかった。

表10. 指数分布の適合度検定の結果

薬剤	P値		
A剤	p < 0.0001	D剤	p < 0.0001
B剤	p < 0.0001	E剤	p = 0.1667
C剤	p = 0.0522	F剤	p < 0.0001

4.3 原データによる指数分布の適合度検定

上記の 4.1 節での図 8.1 から図 8.6 までの印象と 4.2 節での評価結果が、2 薬剤で同じでなかった。その理由として、①1 節に述べた患者集団の規定の仕方の影響、②カテゴリ化されたことでの情報の損失、および③打ち切り割合が高かったことの影響が考えられた。ここでは、①と②について検討するため、原データを用いて指数分布の適合度検定を行った。原データの使用において、副作用が発現しても投与が継続されていた患者については、副作用発現までの日数を用いた。その結果、適合度検定における p 値は、A 剤, B 剤, C 剤, D 剤, E 剤, F 剤の順に、 $p < 0.0001$, $p < 0.0001$, $p = 0.0002$, $p < 0.0001$, $p < 0.0001$, $p < 0.0001$ であった。6 薬剤すべてにおいて、ハザードが投与期間カテゴリー間で一定であるとの仮説は棄却された。すなわち、ハザードは投与期間カテゴリー間で異なっているといた。すなわち、4.1 節での図表からの印象と同じになった。

4.4 原データへのワイブル分布のあてはめ

上記の 4.3 節において、原データに指数分布をあてはめた結果、4.1 節での図の印象と同じ結論が得られた。そのため、ワイブル分布の形状パラメータの推定値 $\hat{\gamma}$ と 95% 信頼区間から、ハザードと投与期間の間のトレンドの評価を行った。薬剤毎の形状パラメータの推定値($\hat{\gamma}$)とその 95%信頼区間を表 11 に示す。表 11 から、ハザードと投与期間の間で、A 剤, B 剤, E 剤, F 剤については正のトレンド、C 剤と D 剤については負のトレンドがあることが示唆された。A 剤では 1 より大きい $\hat{\gamma}$ が推定されたことより、図 8.1 による解釈と同様に、投与期間が長くなるにつれてハザードは高くなるという結果が得られた。B 剤, E 剤, F 剤についても、投与期間が長くなるにつれてハザードが高くなるという点で、それぞれ図 8.2, 図 8.5, 図 8.6 からの解釈に近い結果が得られた。しかし、C 剤と D 剤については、1 より小さい $\hat{\gamma}$ が推定された。すなわち、投与期間が長くなるにつれてハザードは低くなることが示唆され、それぞれ図 8.3 と図 8.4 からの解釈と異なった。

表 11. 薬剤毎の $\hat{\gamma}$ とその 95%信頼区間の下限と上限

薬剤	$\hat{\gamma}$	下限	上限
A 剤	1.2510	1.1463	1.3652
B 剤	1.4574	1.3779	1.5415
C 剤	0.7728	0.6829	0.8745
D 剤	0.6498	0.6098	0.6924
E 剤	1.3347	1.1758	1.5151
F 剤	1.2887	1.1985	1.3858

5. 考察

薬剤の投与期間が副作用の発現に及ぼす影響の有無を調べるのに、ハザード率が投薬開

始からの時間に関係なく一定かどうかを調べることで扱えることを示した。実践では、SAS の LIFEREG プロシジャを用いて、副作用が発現するまでの時間分布が指数分布に従うか否かを検定することになる。より具体的に、LIFEREG プロシジャでは副作用が発現するまでの時間分布にワイブル分布を考え、尺度パラメータに関する仮説 $H_0: \gamma = 1$ に対して Rao のスコア検定が行われる。薬剤を投与してから副作用が発現するまでの時間分布にパラメトリックな分布型を仮定すれば、尤度原理に基づいた推測が可能になるため、Rao のスコア検定だけでなく尤度比検定や Wald 検定を適用できる。

薬剤の安全性に関する特徴を把握するために、性や年齢といった患者要因が副作用の発現に及ぼす影響を探索するとともに、1 日投与量や投与期間といった治療要因が及ぼす影響も評価することが重要である。とくに投与期間については、その長さに伴う副作用発現率の増加、または減少といった評価が必要になる。実際に、投与期間でハザードが一定でないことが示唆された場合に、投与期間は副作用発現の影響要因かどうかから、ハザードが投与期間とともにどのように変化しているかに関心が移る。この場合、副作用発現までの時間分布にワイブル分布をあてはめ、ワイブル分布の尺度パラメータの推定値を求めることで対応できる。ワイブル分布の尺度パラメータの推定値が 1 より大きければハザード率は投与期間とともに増大し、1 より小さければ投与期間とともに減少することを示唆できる。さらに、時間分布にワイブル分布や対数正規分布などを含むより広い範囲の分布を統一的に表現できる一般化ガンマ分布を想定することで、データに最も適合した分布型を探索することができる (Kalbfleisch and Prentice, 1980)。

本報告では副作用が発現する状況を比較的単純化して、投与期間が副作用の発現に及ぼす影響を評価する方法について検討した。ワイブル分布の尺度パラメータの推定値と図的表示の解釈が整合しない場合があった。したがって、投与期間別の副作用発現率の評価において、図的表示と併用して解釈するなど、注意深く解釈することが必要である。

参考文献

- Goldman, A.I. (1992). EVENTCHARTS: Visualizing survival and other timed-events data. *The American Statistician*, 46, 13-18.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). Inference in parametric models and related topics. *The Statistical Analysis of Failure Time Data*, Chap.3, 21-69, John Wiley and Sons.
- SAS Institute Inc. (1989). *SAS/STAT User's Guide*, Version 6, Fourth Edition, Volume 2. SAS Institute Inc.
- 大橋靖雄・浜田知久馬 (1995). *生存時間解析: SAS による生物統計*. 東京大学出版会.

日本SASユーザー会 (SUGI-J)

SAS Integration Technologies + ASP による解析帳票作成 Web システム構築の試み

○岩本 光司* 矢野 尚**

* 武田薬品工業株式会社医薬開発本部統計解析部統計G

** 株式会社富士通ビー・エス・シー ソリューションビジネス本部

The prototype for Web Statistical Analysis Report System based on SAS Integration Technologies and ASP

Koji Iwamoto* Takashi Yano**

* Biostatistics, Pharmaceutical Development Division, TCI

** 4th System Department, Osaka Branch, Solution Business Division,
Fujitsu Broad Solution & Consulting Inc.

要 旨

SAS 8.2において、Output Delivery Systemにより帳票出力機能が大幅に改良された。この機能を活用するための提案として、SUGI-J2000では「ODSによる総括報告書の電子化」の発表の中で、さらにSUGI-J2002では「PH.Clinical Templateによる解析帳票作成に対するシステム化の試み」のデモンストレーションの中で、ODS機能による解析帳票作成を紹介し、これらを実際の業務システムに導入してきた。一方、ここ数年、Webを活用したシステム化が進んでいる。そこで、新たな挑戦として、ASP (Active Sever Pages)とSAS/IT(Integration Technologies)を利用したWebによるシステム化を試みた。この試みにおいて、他システムとの融合などの有用な点があったので、紹介する。

キーワード： SAS Integration Technologies、ASP、ODS、Web

I. はじめに

臨床試験における報告書などに用いる種々の解析結果帳票作成のためにSASは用いられているが、この業務の効率化、作成した帳票の品質保証のために多大なワークロードを必要とする。たとえば、個々の試験ごとにこれらの帳票をSASプログラムを作成した場合、そのプログラム数は多く、また、それらに対してプログラムが正しく動作していることを保証しなければならない。そこで、各試験間で共通に用いられる帳票について標準化し、帳票作成をシステム化することにより、ワークロード及びコストを軽減することが考えられる。これに関するひとつの提案として、筆者らは、SUGI-J2000でODSによる解析帳票作成の試みを、SUGI-J2002でPH.Clinicalと結合したシステム化の試みを報告した。また、これらを実際の業務システムに活用した。しかし、その開発及び運用する中で、いくつかの課題(後述)が見えてきた。

一方、ここ数年の流れで、Webを利用した(クライアントに特別なアプリケーションを必要としない、IEやNetscapeさえあれば良い)システムが多く見られるようになった。種々の業務がインターネットのブラウザから実行できるようになり、エンドユーザーは各自のPC(端末)から、ブラウザを通して日常の

業務を行うようになってきた。そこで、これらの要求を満たすために、ASP(Active Sever Pages)とSAS Integration Technologiesを利用したWebによる解析帳票作成のシステム化を試みた。

II. PH.Clinicalの使用経験

確かにPH.Clinicalは、①SAS/AFをベースとしたGUI作成機能を用いて独自の帳票テンプレートを容易に作成できる、②ユーザー管理やユーザのすべての処理記録を管理するなど優れたAudit Trail機能を持つ、など優れたアプリケーションといえる。

しかし、総括報告書の各種帳票を作成するための独自帳票作成用のカスタムテンプレートを作成する際や、帳票を生成する処理を実行する際に下記のような課題も挙がった。

- GUI構築に制約がある
 - SAS/AF(6.12)をベースとしているので入力画面上でカット&ペーストできない。
 - SAS/AFの機能をすべて使用できない(例えばオブジェクト間のリンクがとれない)。
- 帳票作成のための定義はテンプレート本体とともに保存されるのでファイルサイズが大きくなる
 - テンプレートのSCLコードや帳票作成用SASプログラムも出力結果と合わせて保存されるので、1帳票あたりのファイルサイズが大きくなる。
- データはPH辞書登録が必須であり、この登録処理が煩雑。また、V6.12の制約を受ける
 - SASデータセット及びその変数に対応する形で辞書登録が必要である。これは、PH.Clinical上で定義しなければならない。
- PC上で動くので、ユーザーは他のアプリの実行などの並行作業は難しい
- 各PC上のアプリケーション管理にコストが掛かる

これらのうち、特に重要な問題は、やはり、SAS V6.12の制約を受けていることである。PH.Clinicalのバージョンアップを望むところではあるが、この点に関してSAS社はV8.2対応版をリリースする予定はなく、次期ソリューションのSDD(SAS Drug Development)において対応するとの見解である。しかし、現状では対応しないことから、これらの課題をクリアすることを含めて今回の提案を試みた。

III. 開発のコンセプト

プログラムの標準化及び共有化を考える上で、アプリケーションに依存せずSASプログラム単体で行うことは、その汎用性という観点からいって重要である。前述のPH.Clinicalと結合したシステム化の試みにおいても、帳票作成プログラムはすべてマクロ化し、そのマクロの引数を与える部分についてPH.Templateを用いた。即ち、PH.Clinicalの特徴としてPH.Template上のSCL内にすべての帳票作成プログラムをコーディン

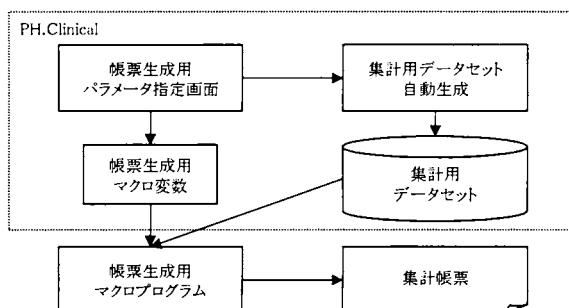


図1 PH.Clinicalでの処理概念図

グすることは可能であるが、SUGI-J2002でも紹介したように、帳票作成のプログラムをマクロ化し、そのマクロの引数をPH.Clinicalから与えることを考えた(図1)。この考えは、システム中枢部である帳票作成部分がアプリケーションに依存しないことを第一に考えたからである。

今回のWebによるシステム化を試みにおいても、帳票作成プログラムはPH.Clinicalと同じものを利用し、マクロの引数を与える部分についてWebによるGUIを作成することを考えた。即ち、パラメータ指定部分と帳票生成用マクロ変数を作成する部分をWebを利用することにした。

また、今回、Web解析システム化する方法を検討した理由のひとつは、解析報告書作成などの定型業務においては上述したようにPH.Clinicalはそのバリデーション機能等優れていることから有用であるが、探索的な解析等、ちょっとした解析を行う際には手続きが多すぎるので、もうすこし簡単に帳票を作成したかったこともある。

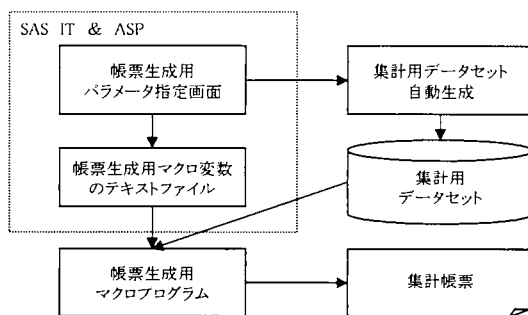


図2 Webシステムでの処理概念図

IV. Webシステムの詳細

1. プロトタイプ作成の前提条件

前述のとおり、Webサーバで集中的に解析処理を行うことのできるシステムについて検討を進めることとしたが、システム開発のために新たなリソースを確保することは考えないこととし、以下の条件のもとでのプロトタイプを作成することとした。

- 単純な環境で構築可能であること
- 高性能マシンが不要(標準的なPCレベルで実現可能)であること
- 高度なWebシステム開発の経験が不要であること
- 既存の資産(データ/プログラム)を利用可能であること

これらの前提条件に従って、プロトタイプシステムはASPを利用して構築することとした。以降で、ASPを選定した理由について述べる。

2. ASPの概要

ASPとは: Microsoft社が提供しているWebサーバ IIS (Internet Information Server/Services)上で動作する「サーバサイドの処理環境」である。ASPは、JavaやVBなどのような固有のプログラミング言語ではなく、「.asp/.aspx/.aspx」ファイルに埋め込まれているスクリプト言語(VBScriptやJavaScriptなど)をサーバサイドで実行処理し、結果(HTML)のみをクライアントに送信する一連の処理環境技術の総称で、多くのWebページでASPが利用されている。IISはサーバ用のOS以外でもWindows 2000 ProfessionalやWindows XP Professionalに添付されており、現有のPCおよびOSで実現可能であるため、今回のプロトタイプはIIS+ASPを利用することとした。

ASPの特性: Webページを動的に表示する代表的な方法として、クライアント上のブラウザでプログラムを動作する方法(クライアントサイドスクリプト DHTML)と、サーバサイドで集中処理を行い、結果のみをHTMLに変換してクライアントに送る方法がある。ASPは後者の方式であり、処理結果のみを返すのでクライアントに負荷をかけない。また、送られたHTMLを表示する機能さえあれば利用可能であり、クライアント環境にほとんど依存しない。

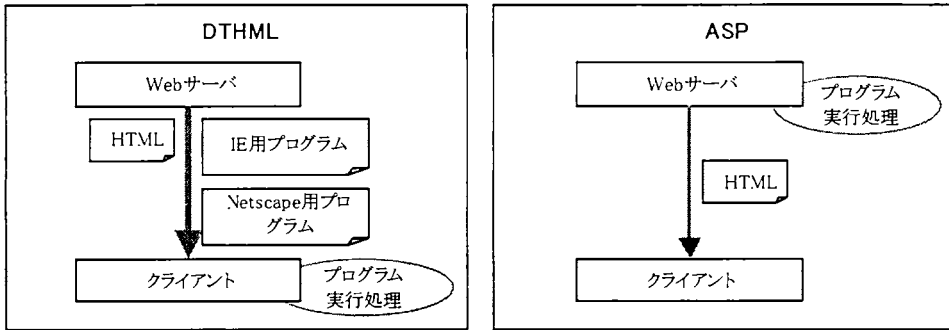


図3 DHTMLとASPの利用イメージ

3. Integration TechnologiesによるSASの利用

VBScriptでの利用: SAS 8.2より追加されたIntegration Technologiesを利用することにより、VBScriptなどの外部プログラムより、SASを利用することができる。

VBScript からの SAS 呼び出しプログラム

```

/* SAS セッションの開始 */
Set objWSMgr = Server.CreateObject("SASWorkspaceManager.WorkspaceManager")
Set objWS=objWSMgr.Workspaces.CreateWorkspaceByServer("My WorkSpace", _
    1, nothing, "", "", Cstr(errString) )
/* SAS プログラムの実行 */
objWS.LanguageService.Submit("data b ; set a ; if d1=' YES' then output; run ;")
/* SAS セッションの終了 */
objWS.Close
    
```

ASPでの問題点: 上記のプログラムは、VBScriptとしてWindows上で実行した場合には正常に動作した。しかし、Webサーバ上に配置し、ASP(Webアプリ)として実行した場合、「書き込みできません」というエラーが通知され実行できなかった。この問題は、ASPで上記プログラムを動作させる場合、Webサーバ上にユーザとしてログインしていないために発生していた。

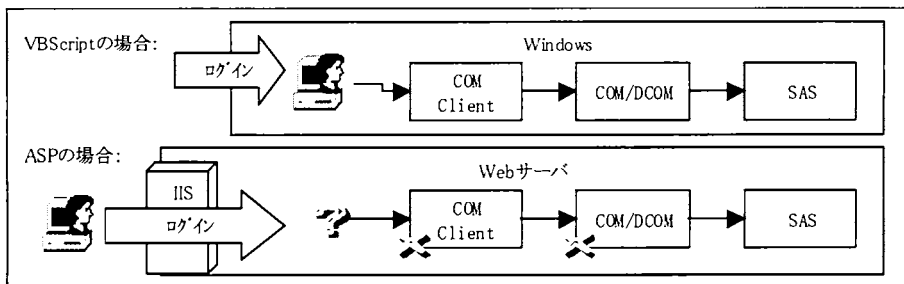


図4 VBScriptとASPの違い

クライアント上のブラウザからWebサーバ(IIS)にアクセスされた場合、デフォルト設定では以下のユーザアカウントが使用される。

- インターネット用ゲストアカウント: IUSR_<Webサーバ・マシン名>
- IIS用プロセスアカウント: IWAM_<Webサーバ・マシン名>

ASP利用に必要な環境定義: 前述の問題を回避するため、Webサーバ上で以下の環境定義を行うことが必要である。“dcomcnfg”コマンドを使用し、SAS:IOM DCOM Serverの環境定義を行うことにより前述の「VBScriptからのSAS呼び出しプログラム」が正常に動作した。ASPを実行するために設定した内容は以下の通りである。

- 分散COMの既定のプロパティで、既定の偽装レベルを「偽装する」に設定する
未知のユーザからの要求を受け取った場合、別ユーザを偽装することによってリソースにアクセスするように定義
- IUSR_<マシン名>およびIWAM_<マシン名>アカウントについて、「アクセス許可」と「起動アクセス許可」を与える
IISで利用されるユーザに対してSAS:IOM DCOM Serverへの「アクセス許可」と「起動アクセス許可」を付与
- アプリケーションの実行時に利用するユーザを設定する
SASの起動時に利用するユーザを設定

4. プロトタイプ

前述の環境定義を行うことによって、ASPからSASを利用することが確認できたため、下図のプロトタイプを作成した。

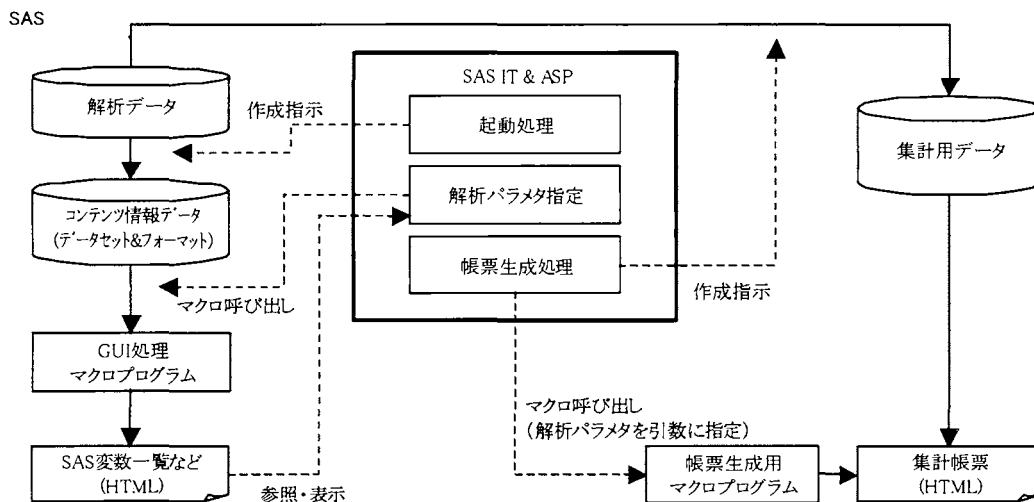


図5 プロトタイプ構成図(背景がグレーの部分にはSASによる処理)

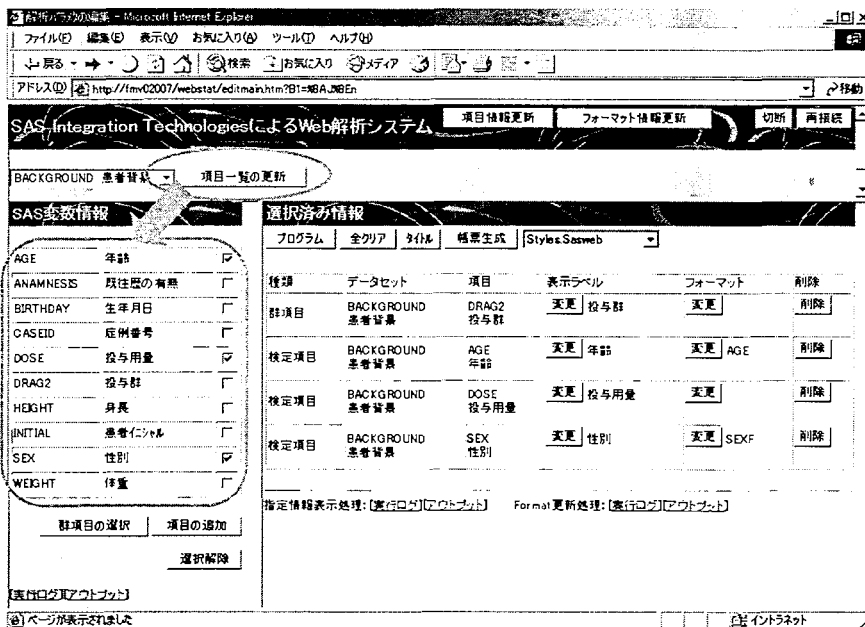


図 6 ASP による Web 解析システムプロトタイプ (パラメタ指定 GUI)

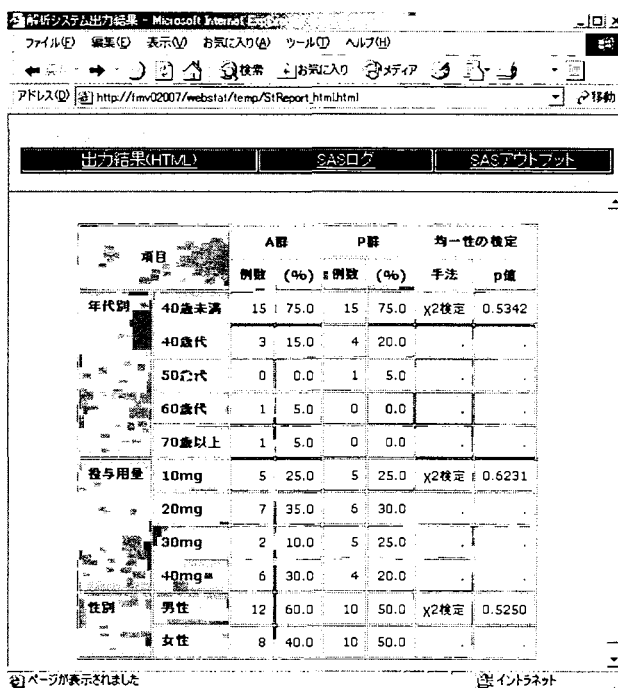


図 7 ASP による Web 解析システムプロトタイプ (帳票出力)

このプロトタイプは、可変表示部分はすべてSASプログラムで実装している。ASPプログラムで行っている処理は、SASの呼び出し、およびSASで出力されたHTMLの取り込みのみである。

「項目一覧の更新」ボタンが押された場合に、左下側のSAS変数一覧を表示するプログラムを以下に示す。

サンプルプログラム

SAS変数一覧表示プログラム(ASP):

```

<%
' --- SAS との連携用ファイル定義 ---'
Const Pbl_SasMacro = "D:\Yinetpub\Ywwroot\YWebStat\Ysasprog\YSellItem.sas"
Const Pbl_CurDir = "D:\Yinetpub\Ywwroot\YWebStat\"
Const Pbl_ASPPFile = "temp\ItemList.asp"
Const Pbl_SasLog = "temp\ItemList_log.txt"
Const Pbl_SasOutput = "temp\ItemList_ist.txt"
画面上側で選択されているデータセット名を取得 ---'
Dim DSname
DSname = Request.Form("DSNAME")
DSname = Left(DSname, Instr(DSname, " "))
' =====
' SAS の起動&処理実行
' =====
Server.Execute "Connect.asp"
objWS = Session.Contents("SASOBJ")
objWS.LanguageService.Submit("%Include '" & Pbl_SasMacro & "' ;")
objWS.LanguageService.Submit("%GenItemList(" & Pbl_CurDir & Pbl_ASPPFile & ", " & _
    Pbl_CurDir & Pbl_SasLog & ", " & Pbl_CurDir & Pbl_SasOutput & ", " & _
    DSname & ") ;")
if err.number = 0 then
' --- 上記 SAS 処理で出力された ASP ファイルをインクルード ---'
If objFile.FileExists(Pbl_CurDir & Pbl_ASPPFile) Then
    Server.Execute Pbl_ASPPFile
else
    Response.Write "<font color=red> SAS の出力ファイルが見つかりません。" & _
        ログを確認してください。<br>"
End If
%>

```

データセット一覧の選択結果を参照

SAS マクロ呼び出し

SAS で出力されたファイル (HTML)を取り込んで表示

SAS変数一覧表示プログラム(SASマクロ):

```

%macro GenItemList (ASPPFILE, LOGFILE, OUTFILE, DSNAME) ;
/* 出力ファイルの制御 */
FILENAME HTML "&ASPPFILE" ;
FILENAME LOGFILE "&LOGFILE" ;
FILENAME OUTPUT "&OUTFILE" ;
PROC PRINTTO LOG=LOGFILE print=output ;run ;

/* データセット・項目リストデータセットを作成 */
DATA ITEMLIST ; SET TBLLIST ;
    IF MEMNAME="&DSNAME" THEN OUTPUT ;
RUN ;

/* 生成されたデータセットから、HTML出力部を作成 */
PROC SORT DATA=ITEMLIST ; BY NAME ; RUN ;
DATA _NULL_ ; SET ITEMLIST END=LASTOBS ;
    FILE HTML ;
    IF _N_=1 THEN DO ;
        PUT '<div align="center">' ;
        PUT '<center>' ;
        PUT '<table border="1" width="100%" bgcolor="white">' ;
    END ;
    _N_2 = ROUND(_N_/2) ;
    IF _N_2 * 2 = _N_ THEN PUT '<tr>' ;
    ELSE PUT '<tr bgcolor="#FFFF99">' ;
    PUT '<td width="40%"><font size="2"> NAME '</font></td>' ;

```

```

PUT ' <td width="50%"><font size="2">' LABEL ' </font></td>' ;
_CHKNAME = "ItemCheck" || TRIM(LEFT(_N_)) || ' ' ;
_DSNAME = "DSNAME" || TRIM(LEFT(_N_)) || ' ' ;
_ITMNAME = "ITEMNAME" || TRIM(LEFT(_N_)) || ' ' ;
PUT ' <td width="10%"><input type="checkbox" name=" ' _CHKNAME ' value="ON">' ;
PUT ' <input type="hidden" name=" ' _DSNAME ' VALUE=" ' "&DSNAME" ' "></td>' ;
PUT ' <input type="hidden" name=" ' _ITMNAME ' VALUE=" ' NAME ' ' "></td>' ;
PUT ' </tr>' ;
IF LASTOBS THEN DO ;
    PUT ' </table>' ;
    PUT ' </center>' ;
    PUT ' </div>' ;
    _Count = LEFT(_N_);
    PUT ' <input type="hidden" name="ITEMCOUNT" VALUE=" ' _Count ' ' "></td>' ;
END ;
RUN ;
/* 出力先をデフォルトに戻す */
PROC PRINTTO ; RUN ;
%MEND ;

```

実行例

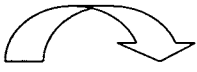
図6のデータセットが選択された場合、SASから以下の文字列(HTML形式であるが、実際には.aspファイルで処理)が出力される。

SASで出力される文字列:

```

<table border="1" width="100%" bgcolor="white">
<tr bgcolor="#FFFF99">
<td width="40%"><font size="2">AGE </font></td>
<td width="50%"><font size="2">年齢 </font></td>
<td width="10%"><input type="checkbox" name="ItemCheck1"
<input type="hidden" name="DSNAME1" VALUE="BACKGROUND"></
<input type="hidden" name="ITEMNAME1" VALUE="AGE " "></td>
</tr>
<tr>
<td width="40%"><font size="2">ANAMNESIS </font></td>
<td width="50%"><font size="2">既往歴の有無 </font></td>
<td width="10%"><input type="checkbox" name="ItemCheck2"
<input type="hidden" name="DSNAME2" VALUE="BACKGROUND"></
<input type="hidden" name="ITEMNAME2" VALUE="ANAMNESIS " ">
</tr>
<tr bgcolor="#FFFF99">
<td width="40%"><font size="2">BIRTHDAY </font></td>
<td width="50%"><font size="2">生年月日 </font></td>
<td width="10%"><input type="checkbox" name="ItemCheck3" value="ON">
<input type="hidden" name="DSNAME3" VALUE="BACKGROUND"></td>
<input type="hidden" name="ITEMNAME3" VALUE="BIRTHDAY " "></td>
</tr>
    ~ (途中省略) ~
</table>

```



AGE	年齢	<input type="checkbox"/>
ANAMNESIS	既往歴の有無	<input type="checkbox"/>
BIRTHDAY	生年月日	<input type="checkbox"/>
CASEID	症例番号	<input type="checkbox"/>
DOSE	投与用量	<input type="checkbox"/>
DRAG2	投与群	<input type="checkbox"/>
HEIGHT	身長	<input type="checkbox"/>
INITIAL	患者イニシャル	<input type="checkbox"/>
SEX	性別	<input type="checkbox"/>
WEIGHT	体重	<input type="checkbox"/>

V. おわりに

今回のWebシステム化の試みにより、以下のことがわかった。

- 本方式で業務システムを構築可能である
- ブラウザが動作する環境であれば、クライアントはOSに依存しない
- サーバ側だけメンテナンスすればよいので、バリデーション・システム管理の負荷を軽減できる
- 他のWebシステムとの融合が容易(ポータルサイト・ドキュメント管理システムなど)である
- ASPによるパラメタ指定は、DBなどを利用する方式の方が効率的である
- 各種操作に対するオーデイトレイル取得も容易に構築できる

また、SASのWebソリューションとしてSAS AppDev Studioなどもあり、これらを用いたアプリケーションの開発も考えられる。さらに、近々、日本においても医薬品開発における統計解析を中心としたトータル的なソリューションであるSAS Drug Developmentがリリースされる予定である。今後はこれらを視野にいれたシステム化も検討していきたい。

2 値および計量値のシグモイド曲線 — 曲線の推定および逆推定と 95%信頼区間 —

○杉山 公仁*、馬場 淳**、天竺桂 裕一郎***、高橋 行雄****

*昭和薬品化工株式会社 開発研究部、**明治製菓株式会社 薬事部

興和株式会社 医薬事業部、*中外製薬株式会社 臨床解析部

Sigmoidal Curve Fitted to Qualitative and Quantitative data
Curve Fitting, Inverse Estimation and its 95% confidence

Kimihito Sugiyama*, Jun Baba**, Yuichiro Tabunoki***, Yukio Takahashi****

*Development and Research Dept. / Showayakuhinkako co., Ltd.

**Registration & Regulatory Affairs Dept. / Meiji Seika Kaisha, Ltd.

***Development Research Dept. / Kowa Company, Ltd.

****Clinical Data Analysis Dept. / Chugai Pharmaceutical Co., Ltd.

要 旨

薬理試験あるいは毒性試験などの分野において、薬剤および化学物質の特徴を示すために、ある反応系で 50%の反応率を示すときの用量 (D_{50}) が求められており、そのために、用量反応曲線を当てはめる方法として、従来からプロビット法が知られているが、ロジスティック回帰がこのようなデータに対して有用であることが知られるようになってきた。そこで、反応率として表わされるような 2 値データに対して、ロジスティック回帰分析を適用して用量反応曲線を推定し、 D_{50} が逆推定できることを示すだけでなく、10%あるいは90%の反応率を示す用量 (D_{10} あるいは D_{90}) を求める拡張法を示した。また、シグモイド型の計量値データとなる摘出組織を用いた *in vitro* 試験や細胞毒性試験では非線形回帰モデルにより D_{50} 及び最大反応量 (E_{max}) を推定する E_{max} 法がデータ解析に用いられている。 E_{max} モデルのパラメータである D_{50} は酵素反応の K_M 、 E_{max} は V_{max} に一致する。 E_{max} モデルでは D_{50} とその 95%信頼区間が直接推定でき、ロジスティック回帰と同様の考え方で、 D_{10} あるいは D_{90} が求められた。

キーワード： シグモイド曲線、 D_{50} 、SAS/PROBIT、SAS/LOGISTIC、SAS/NLIN、JMP

1. 目的

薬理試験あるいは毒性試験などで扱われる反応率として表わされるような 2 値データに対して、薬剤および化学物質の特徴を示すために、 D_{50} が求められている。ロジスティック回帰により、単に D_{50} を逆推定でき、プロビット法と同等であることを示すだけでなく、 D_{10} あるいは D_{90} を推定するために拡張する。また、シグモイド型の計量値データの解析には、非線形回帰モデルにより D_{50} 及び最大反応量 (E_{max}) を推定する E_{max} モデルが用いられ、 D_{50} とその 95%信頼区間が直接推定できる。ロジスティック回帰と同様の考え方で、 D_{10} あるいは D_{90} が求められることを示すとともに、酵素反応やリガンドバイディング試験への応用について示す。

2. 2 値データのシグモイド曲線

2.1. プロビット法（プロビット変換）

薬剤および化学物質の特徴を示すために、薬理試験あるいは毒性試験などで反応率として表わされるようなデータ（2 値データ）に対して、用量との関係を表すとき使用される解析方法の一つが正規分布関数を用いるプロビット法であり、50%の反応率を示すときの用量（50%有効量： ED_{50} あるいは50%致死量： LD_{50} 、以下 $D50$ と呼ぶ）を求めることができる。吉村功編著「毒性・薬効データの統計解析」¹⁾ の「5.4 節 LD_{50} の推定」では、プロビット法による $D50$ の推定とその95%信頼区間の計算法が示されている。

プロビット法により求められたプロビット曲線は反応率が、用量の対数に対して、正規分布の累積確率、すなわち正規分布関数の関係をもつシグモイド（S 字）曲線のことである。式で書けば用量 d と反応率 p の関係が

$$p = \int_{-\infty}^{\log_{10} d} \frac{\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} dx \quad \text{式 1}$$

となるものである。このとき、 $\mu = \log_{10} D50$ である。

計算では次の式で定義するプロビット関数を使う。

$$\text{probit}(p) = \left\{ p = \int_{-\infty}^{y-5} \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}} dx \text{ となる } y \text{ の値} \right\} \quad \text{式 2}$$

y は規準正規分布の累積確率が p となる場所の横軸（正規偏差）に負の数字を嫌って5を加えたものである。数値例として表 1 が示され、 $D50$ の求め方として重み付き最小2乗法の繰り返し計算による最尤法が示されている。

表 1 $D50$ を求める数値例

群 i	投与量 mg/kg (公比 1.35)	群の大きさ	死亡数	死亡率 p_i	プロビット $\text{probit}(p_i) = y_i$
1	101	10	0	0.000	$-\infty$
2	136	10	2	0.200	4.1584
3	183	10	5	0.500	5.0000
4	247	10	8	0.800	5.8416
5	333	10	9	0.900	6.2816
6	450	10	10	1.000	∞

SAS では PROBIT プロシジャ（プロビット変換、ニュートン・ラフソン法）により、求めることができる。プログラミング例を以下に示した。

SAS データセット

```
data d01 ;
  input l dose n y p eta p_hat ;
datalines ;
1 101 10 0 0 3.1931 0.035
2 136 10 2 0.2 4.0463 0.170
3 183 10 5 0.5 4.8975 0.459
4 247 10 8 0.8 5.7575 0.776
5 333 10 9 0.9 6.6142 0.947
6 450 10 10 1 7.4777 0.993
;
proc probit data=d01 log10 inversecl ;
  model y/n = dose / dist=normal itprint covb ;
  output out=out01 p=p std=std xbeta=xbeta ;
run ;
proc print data=out01 ;
run ;
```

SAS の proc PROBIT の反復計算は、初期の回帰係数 $\beta_0 = 0$ 、 $\beta_1 = 0$ からスタートして、6 回のニュートン・ラフソン法による反復の結果 $\beta_0 = -15.0703$ 、 $\beta_1 = 6.6152$ が得られた。 $D50$ は 2.278 であり、95%信頼区間は、フィラーの式²⁾により、常用対数で (2.204, 2.348) と計算され、投与用量に変換して $D50 = 189.7$ 、(160, 223) となった。なお、SAS では、proc LOGISTIC、proc GENMOD でもプロビット法での計算は行えるが、 $D50$ の 95%信頼区間の計算がサポートされていない。

2.2. ロジスティック回帰 (ロジット変換)

シグモイド曲線を得るために正規分布の数値計算は煩雑であることから、数値計算が簡単なロジスティック分布をシグモイド曲線に用いる方法が利用されるようになってきた。

$$f(x) = \frac{\exp\left(\frac{x-\mu}{\tau}\right)}{\tau \left\{1 + \exp\left(\frac{x-\mu}{\tau}\right)\right\}^2}, \quad -\infty < x < \infty \quad \text{式 3}$$

ここで、 $-\infty < \mu < \infty$ 、 $\tau > 0$ であり、平均と分散は、それぞれ μ と $\pi^2\tau^2/3$ である。確率密度関数 $f(x)$ は、正規分布に比べ簡潔とはいえないが、反応率 p 、および用量 d とし、 $\beta_0 = -\mu/\tau$ 、 $\beta_1 = 1/\tau$ とおけば、ロジスティック分布関数は、

$$p_i = \frac{\exp(\beta_0 + \beta_1 \log_{10} x_i)}{1 + \exp(\beta_0 + \beta_1 \log_{10} x_i)} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \log_{10} x_i))} \quad \text{式 4}$$

となる。簡単な式の変形により、ロジスティック回帰式

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \log_{10} x_i \quad \text{式 5}$$

が得られる。式 5 が正規分布を用いた場合とほぼ同様のシグモイド曲線を与える。このシグモイド曲線を用いて $D50$ を推定するのがロジスティック回帰モデルである。JMP による $D50$ の推定は、プロビット法ではなくロジスティック回帰モデルによる推定を行っている。

プロビット法で示したような生物検定法で常用される逆推定、推定値の 95%信頼区間をロジスティック回帰において求めることは SAS でできないわけではないが、JMP では、Inverse Prediction

(逆推定)として対応しており、生物検定法のための統計パッケージとしては JMP が使いやすい。そこで、ロジスティック回帰については JMP により解析を行った。

JMP でロジスティック回帰を行った結果を表 2 に示した。

表 2 ロジスティック回帰モデルの結果

パラメータ推定値				
項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
切片	-26.211489	6.350431	17.04	<.0001
log10(dose)	11.5229099	2.7804107	17.18	<.0001

$p=0.5$ の時の投与量が $D50$ であることから、 $\log_{10}(0.5)=\ln(1)=0$ となる。 $D50$ の時、式 5 は $\beta_0 + \beta_1 \cdot \log_{10}(D50) = 0$ となり、 $\log_{10}(D50) = -\hat{\beta}_0 / \hat{\beta}_1$ が得られる。

したがって、 $D50$ は、求められた回帰係数から、

$$\hat{\mu} = \log_{10}(D50) = -\hat{\beta}_0 / \hat{\beta}_1 = -(-26.2115) / 11.5229 = 2.275$$

となり、 $D50 = 10^{2.275} = 188.2 \text{ mg/kg}$ と推定できる。

プロビット法の 189.7 mg/kg と比べて約 1% の差である。分散は、

$$\hat{\sigma}^2 = \sqrt{\frac{\pi^2 \tau^2}{3}} = \sqrt{\frac{3.14^2 \times (1/11.52)^2}{3}} = 0.157$$

とプロビット法の 0.151 と約 4% の差である。

プロビット法によって得られた $D50 = 2.278$ と $\hat{\sigma} = 0.151$ から推定されるプロビット曲線、 $D50 = 2.278$ と $\hat{\tau} = 0.083$ から推定されるロジスティック曲線を比較した結果を図 1 に示す。違いは、ロジスティック回帰が裾広がりとなるが、図にしてみるとごくわずかである。

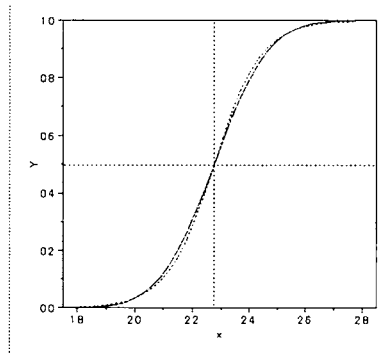


図 1 プロビット曲線とロジット曲線の比較

— プロビット、--- ロジット

$D90$ 値を得るためには、 $p = 0.90$ を、ロジット変換して得られる

$$\log(0.9/0.1) = \beta_0 + \beta_1 \cdot \log_{10}(D90)$$

$$\log_{10}(D90) = \frac{\beta_0 - \log(0.9/0.1)}{\beta_1} = \frac{\beta_0 - 2.1972}{\beta_1}$$

から、 $D90$ 値は、 $D90 = 10^{-(\hat{\beta}_0 - 2.1972) / \hat{\beta}_1}$ と推定される。

同様に、 $D10$ 値は、 $D10 = 10^{-(\hat{\beta}_0 + 2.1972) / \hat{\beta}_1}$ と推定される。

JMP では、逆推定の機能により、任意の死亡率について計算できるので、90%、50%、および10%のそれぞれについて推定した結果を表 3 に示す。D50 の 95%信頼区間は、フィラーの式により、常用対数で (2.196, 2.351) と推定される。

表 3 ロジット法による逆推定

逆推定		確率	予測値 log10(dose)	下限	上限	1-Alpha
	0.90000000	2.46541141	2.38150573	2.66236231	0.9500	
	0.50000000	2.27472827	2.19554355	2.35091729		
	0.10000000	2.08404514	1.87888963	2.17016401		

2.3. フィラーの定理を用いた有効用量の信頼区間

フィラーの定理は2つの正規分布の確率変数の比率の信頼区間によって得ることができる一般的な計算結果である。 $\rho = \beta_0 / \beta_1$ としたとき、 β_0 と β_1 は、 $\hat{\beta}_0$ と $\hat{\beta}_1$ によって推定される。その平均を β_0 と β_1 、分散が v_{00} と v_{11} 、共分散が v_{01} の正規分布になると仮定される関数 $\psi = \hat{\beta}_0 - \rho \hat{\beta}_1$ について考える。このとき、 $\hat{\beta}_0$ と $\hat{\beta}_1$ が β_0 と β_1 の不偏推定量であるので、 $E(\psi) = \beta_0 - \rho \beta_1 = 0$ となり、 ψ の分散は、

$$V = \text{Var}(\psi) = v_{00} + \rho^2 v_{11} - 2\rho v_{01} \quad \text{式 6}$$

で与えられる。 $\hat{\beta}_0$ と $\hat{\beta}_1$ は、正規分布に従うと仮定されるので、 ψ は、同様に正規分布に従い

$$\frac{\hat{\beta}_0 - \rho \hat{\beta}_1}{\sqrt{V}}$$

は、標準正規分布となる。

従って、 $z_{\alpha/2}$ が、標準正規分布の上側 $\alpha/2$ 点であるとしたときに、 ρ の $100(1-\alpha)\%$ 信頼区間は、 $|\hat{\beta}_0 - \rho \hat{\beta}_1| \leq z_{\alpha/2} \sqrt{V}$ で表される。

両辺を2乗し、等式とし、 $\hat{\beta}_0^2 + \rho^2 \hat{\beta}_1^2 - 2\rho \hat{\beta}_0 \hat{\beta}_1 - z_{\alpha/2}^2 V = 0$ を与える。

式6により V を代入した後に式の整理をすると、次のように ρ に関する2次方程式が得られる。

$$(\hat{\beta}_1^2 - \hat{v}_{11} z_{\alpha/2}^2) \rho^2 + (2\hat{v}_{01} z_{\alpha/2}^2 - 2\hat{\beta}_0 \hat{\beta}_1^2) \rho + (\hat{\beta}_0^2 - \hat{v}_{00} z_{\alpha/2}^2) = 0 \quad \text{式 7}$$

この2次方程式の2つの根は、 ρ のための信頼限界を構成する。これが、フィラーの結果である。この結果を $D50 = -\beta_0 / \beta_1$ の信頼区間を得るために、式7の ρ を $-D50$ と置き換える。

D50 による2次方程式を書き換えると、

$$(\hat{\beta}_1^2 - \hat{v}_{11} z_{\alpha/2}^2) D50^2 - (2\hat{v}_{01} z_{\alpha/2}^2 - 2\hat{\beta}_0 \hat{\beta}_1^2) D50 + (\hat{\beta}_0^2 - \hat{v}_{00} z_{\alpha/2}^2) = 0 \quad \text{式 8}$$

が得られ、この2次方程式を標準的な手順により解き、D50 値の $100(1-\alpha)\%$ の信頼限界のために次の式を得る。これも一般的にフィラーの式といわれているものである。

$$D50 = \frac{-\left(\hat{\rho} - g \frac{\hat{v}_{01}}{\hat{v}_{11}}\right) \pm \frac{z_{\alpha/2}}{\hat{\beta}_1} \left\{ \hat{v}_{00} - 2\hat{\rho} \hat{v}_{01} + \hat{\rho}^2 \hat{v}_{11} - g \left(\hat{v}_{00} - \frac{\hat{v}_{01}^2}{\hat{v}_{11}} \right) \right\}^{1/2}}{1 - g} \quad \text{式 9}$$

ここで、 $\hat{\rho} = \hat{\beta}_0 / \hat{\beta}_1$ 、 $g = z_{\alpha/2}^2 \hat{v}_{11} / \hat{\beta}_1^2$ である。

強い用量反応関係があるとき、 $\hat{\beta}_1$ は0に対して高度に有意となり、また、 $\hat{\beta}_1 / \sqrt{\hat{v}_{11}}$ は、 $z_{\alpha/2}$ より極めて大きくなる。この場合に g は、小さくなる。すなわち、より有意となるような関連の場合、 g はより無視できるようになる。

g が式 9 でゼロである場合、 $D50$ 値の信頼限界は、

$$s.e.(D50) \approx \left\{ \frac{\hat{v}_{00} - 2\hat{\rho}\hat{v}_{01} + \hat{\rho}^2\hat{v}_{11}}{\hat{\beta}_1} \right\}^{1/2}$$

で与えられる $D50$ 値の標準誤差の近似に基づくものと一致する。

$\log(dose)$ が説明変数として使用されている場合、 $D50$ 値の信頼区間は、フィラーの定理を用い $\log(D50) = -\beta_0 / \beta_1$ について信頼限界を得ることにより計算でき、次に、その値について指数をとればよい。

2.4. $D50$ の信頼区間の計算事例

対数の $D50$ について 95%信頼区間は、式 9 を用いるのではなく、式 8 の 2 次式の根を求める手順を示す。それぞれの係数は、JMP のロジスティック回帰係数が $\hat{\beta}_0 = -26.2115$ 、 $\hat{\beta}_1 = 11.5229$ で、JMP では標準的に分散共分散行列が出力されないことから下記プログラムに示したように SAS の proc logistic で計算し、得られた分散共分散行列は、

$$\Sigma = \begin{bmatrix} 40.328 & -17.623 \\ -17.623 & 7.731 \end{bmatrix}$$

となるので、

$$a = \hat{\beta}_1^2 - z_{\alpha/2}^2 \hat{v}_{11} = 11.5229^2 - 1.96 \times 7.731 = 103.0801$$

$$b = -\left(2\hat{v}_{01}z_{\alpha/2}^2 - 2\hat{\beta}_0\hat{\beta}_1\right) = -(2 \times (-17.623) \times 1.96^2 - 2 \times (-26.2115) \times 11.5229) \\ = -468.681$$

$$c = \hat{\beta}_0^2 - \hat{v}_{00}z_{\alpha/2}^2 = (-26.2115)^2 - 40.328 \times 1.96^2 = 532.1449$$

となる。これを 2 次式の公式に代入すると 95%信頼区間が得られる。

$$\log(D50) \pm 1.96s.e.(\log(D50)) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = (2.197, 2.350)$$

$D50$ とその信頼区間の計算公式は、ロジット法の場合のみならずプロビット法の場合もそのまま適用できる。JMP による信頼区間の計算は、フィラーの公式によって計算されている。

SAS データセット

```
data d01 ;
  input i dose n y ;
  log_dose=log10(dose) ;
datalines ;
1 101 10 0
2 136 10 2
3 183 10 5
4 247 10 8
5 333 10 9
6 450 10 10
```

```

;
proc logistic data=d01 ;
  model y/n = log_dose / link=logit itprint covb ;
run ;

```

SAS 出力(結果の一部)

Estimated Covariance Matrix

Variable	Intercept	log_dose
Intercept	40.32799	-17.6233
log_dose	-17.6233	7.730687

3. 計量値のシグモイド曲線

3.1. Emax モデル

モルモットから摘出した平滑筋を用いた薬物 - 受容体モデルの実験データから $D50$ の推定について示す。この実験は、摘出した平滑筋を溶液中に懸架して、まずヒスタミン濃度が $0.01 \mu\text{M}$ となるようにヒスタミンを加え、さらに濃度が $\sqrt{10}=3.16$ 倍となるようにヒスタミンを加え、平滑筋の収縮が止まるのを待ち、さらに 3.16 倍の濃度に上げる、といった手順により、平滑筋の収縮量を計測する。

表 4 ヒスタミンによる平滑筋の収縮

ヒスタミン濃度 (μM)	0.01	0.0316	0.10	0.316	1.00	3.16	10.00	31.60	100.00	316.00
平滑筋収縮量 (mm)	1	3	5	23	66	113	158	171	171	165

この収縮反応は、一般的にシグモイド曲線となり、要約統計量として最大反応の 50% の収縮量となるような薬物濃度、いわゆる EC_{50} (以下 $D50$) が要約統計量として用いられている。このシグモイド曲線の当てはめに非線形回帰モデルの一つである Emax モデル³⁾

$$y_i = \frac{Emax \cdot x_i^\gamma}{x_i^\gamma + D50^\gamma} + e_i, \quad i=1,2,\dots,k \quad \text{式 10}$$

y_i : 平滑筋の収縮量

x_i : 収縮の作動薬ヒスタミンの濃度

$Emax$: 最大収縮量

$D50$: 最大収縮量の 1/2 となるヒスタミンの濃度

γ : ロジスティック曲線の傾き

が用いられている。式 10 は分子、分母を、 x_i^γ で割り、一部を指数化すると

$$y_i = \frac{Emax}{1 + \frac{D50^\gamma}{x_i^\gamma}} = \frac{Emax}{1 + \exp(\ln(\frac{D50^\gamma}{x_i^\gamma}))} = \frac{1}{1 + e^{\gamma(\ln(D50) - \ln(x_i))}} Emax \quad \text{式 11}$$

に変形できる。

汎用的な統計ソフトにも非線形回帰モデルを使うための手法も含まれてはいたが、偏微分式の設定、初期値の設定を必要としていた。最近、汎用的な統計ソフトも進化し、SAS ではバージョン 6.12 から、非線形のモデル式からパラメータについての偏微分式を自動的に行う機能が付加されて使い勝手が向上してきた。JMP もバージョン 4 では、偏微分を自動的に行う機能、さらに初

期値をスライダーなどで変化させ当てはまり具合を視覚的に確認できる機能も持っている。JMPでは、さらに、解を求めるための反復計算過程も視覚的に確認できるようになり、非線形回帰モデルを手軽に使えるようになってきた。

3.2. シグモイド曲線のモデル式

表 4 のデータについて X 軸を対数目盛にしてグラフを作成すると図 2 のようにシグモイド状の反応となる。JMP を用いて式 10 の 3 つのパラメータを推定すると $D\hat{5}0 = 1.59$ 、 $\hat{E}max = 171.58$ 、および $\hat{\gamma} = 1.17$ が得られる。図 2 にはこれらのパラメータの推定値を式 10 に代入して得られた反応 y の推定値を図示してある。パラメータ $\hat{E}max = 171.58$ は、ヒスタミンの濃度を無限大にしたときの反応である。パラメータ $D50$ は、 E_{max} が $1/2$ となるような x の値である。 γ に関わらずシグモイド曲線は x が $D50$ のとき、 y は $E_{max} / 2$ を通る。

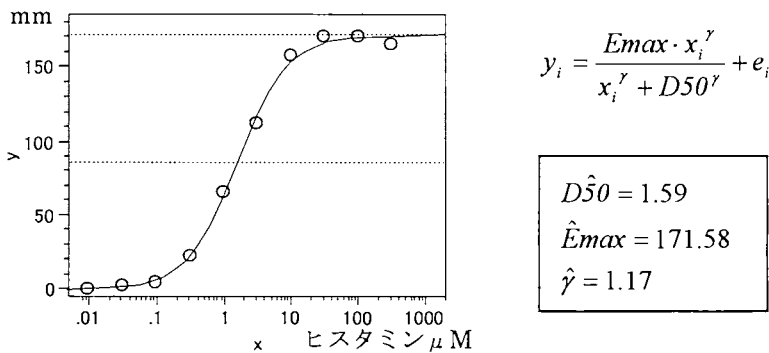


図 2 ヒスタミンによる平滑筋の収縮

3.3. 酵素反応におけるシグモイド曲線

反応が飽和する場合に Michaelis-Menten の酵素反応として知られている式

$$Velocity = \frac{Vmax \cdot Substrate}{Substrate + K_M} \quad \text{式 12}$$

Velocity : 反応速度 : 目的変数 y

Substrate : 酵素の反応で変化する基質の濃度 : 説明変数 x

$Vmax$: 最大反応速度 (データから推定したいパラメータ)

K_M : Michaelis 定数 (データから推定したいパラメータ)

は Hill slope といわれているパラメータ h を用いて一般化して表せる

$$Velocity = \frac{Vmax \cdot Substrate^h}{Substrate^h + K_{0.5}^h}$$

の $h=1.0$ の場合である。酵素反応の場合は、基質の濃度 x を対数ではなく実濃度で表すために直角双曲線として表されているが、軸を対数にするとシグモイド曲線である。

この一般化した式は E_{max} モデルと同じ非線形モデルとなる。

E_{max} モデルにおけるパラメータである $D50$ は酵素反応の K_M 、リガンドバインディング試験の K_D と本質的には同じであり、 $Vmax$ や $Bmax$ は E_{max} に一致する。つまり、酵素反応の場合 $Vmax$

= Emax、Substrate = x、h = γ 、 $K_{0.5} = D50$ である。

Emax モデルでは、直接 D50 の推定、および 95%信頼区間が計算されるので、2 値データのロジスティック回帰のように再計算の手間が要らないのであるが、D10、あるいは D90 などの計算は別途行わなければならない。この場合には、D50 を推定するのではなく、直接 D10 を推定するようにモデル式を変更する。

反応が最大値の 10%となる濃度の推定値は、

$$\ln(0.1/0.9) = \beta_0 + \beta_1 \cdot \ln(D10)$$

なので、

$$\ln(D10) = \frac{-2.197 - \beta_0}{\beta_1}$$

となり、切片 β_0 に求めたい反応のパーセント点 の $\text{logit}(p)$ を加えることにより得られる。

D10、あるいは D90 などの推定は、Emax モデルを次のように変形すればよい。

$$\ln(D10) = \frac{-2.197 - \beta_0}{\beta_1} = -\frac{2.197}{\beta_1} - \frac{\beta_0}{\beta_1}$$

であるので、 $\beta_1 = \gamma$ 、 $\ln(D50) = -\beta_0 / \beta_1$ を代入し、

$$\ln(D50) = \ln(D10) + \frac{2.197}{\gamma}$$

と式を変形して、式 11 に代入すると、

$$y_i = \frac{Emax}{1 + e^{\gamma(\ln(D10) + 2.197/\gamma - \ln(x_i))}} = \frac{Emax}{1 + e^{\gamma(\ln(D10) - \ln(x_i)) + 2.197}}$$

が得られる。同様な手順により D90 は、

$$y_i = \frac{Emax}{1 + e^{\gamma(\ln(D90) - \ln(x_i)) - 2.197}} \quad \text{となる。}$$

下に Emax モデルの SAS プログラムと解析結果を示した。

SAS データセット

```
data d01 ;
  input x ln_x y ;
  datalines ;
  0.01      -2.00      1
  0.0316   -1.50      3
  0.1       -1.00      5
  0.316    -0.50     23
  1         0.00     66
  3.16     0.50    113
  10       1.00    158
  31.6     1.50    171
  100      2.00    171
  316      2.50    165
  ;

Title '<<< logistic, D50 >>>' ;
proc nlin data=d01 list method=gauss;
  parms Emax=170 gamma=1 D50= 5 ;
  model y = Emax / (1 + exp(gamma*(log(D50) - log(x)))) ;
run ;
```

```
Title '<<< logistic, D10 >>>' ;
proc nlin data=d01 list method=gauss;
  parms Emax=170 gamma=1 D10= 1 ;
  model y = Emax / (1 + exp(gamma*(log(D10) - log(x)) + 2.197 )) ;
run ;

Title '<<< logistic, D90 >>>' ;
proc nlin data=d01 list method=gauss;
  parms Emax=170 gamma=1 D90= 10 ;
  model y = Emax / (1 + exp(gamma*(log(D90) - log(x)) - 2.197 )) ;
run ;
```

SAS 出力(結果の一部)

<<< logistic, D50 >>>

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
Emax	171.6	2.6814	165.2	177.9
gamma	1.1678	0.0821	0.9735	1.3620
D50	1.5874	0.1132	1.3197	1.8550

<<< logistic, D10 >>>

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
Emax	171.6	2.6814	165.2	177.9
gamma	1.1678	0.0821	0.9735	1.3620
D10	0.2419	0.0317	0.1669	0.3169

<<< logistic, D90 >>>

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
Emax	171.6	2.6814	165.2	177.9
gamma	1.1678	0.0821	0.9735	1.3620
D90	10.4171	1.7432	6.2951	14.5390

4. まとめ

2 値データに対するロジスティック回帰モデルは、プロビット法のプロビット変換をロジット変換としたモデルであることから、同様の結果を得ることを示した。特定の反応率のロジットからロジスティック回帰の式を変形することにより、その反応率を示す用量を推定でき、フィラーの式によりその 95%信頼区間が求められることを示した。計量値に対しては Emax モデルによる非線形回帰により、推定したいパラメータとして D50 が得られるとともに、95%信頼区間も直接推定できた。Emax モデルを 2 値のロジスティック回帰モデルと同様の考え方でモデル式を変形すると、D10 あるいは D90 など同様に直接求めることができた。

参考文献

- 1)吉村功編著：毒性・薬効データの統計解析－事例研究によるアプローチ，サイエンティスト社（1987）
- 2)D. Collett： Modeling Binary Data, Chapman and Hall/CRC (1991)
- 3)Gabrielsson J, Weiner D. : Pharmacokinetic and pharmacodynamic data analysis: Concepts and Applications, 2nd ed., Swedish Pharmaceutical Press, Stockholm (1997).

日本SASユーザー会 (SUGI-J)

陰性および陽性対照があるシグモイド曲線

ーダミー変数を持つ非線型回帰モデルの応用ー

○山田 雅之*, 吉田 光宏**, 高橋 行雄***

* キッセイ薬品工業株式会社 開発企画部, ** グラクソ・スミスクライン株式会社,

*** 中外製薬株式会社 臨床解析部

Sigmoid Curve with Negative and Positive Control

- Application of Nonlinear Regression Model that have Dummy Variables -

○Masayuki Yamada*, Mitsuhiro Yoshida**, Yukio Takahashi***

* Kissei Pharmaceutical Co., Ltd., ** GlaxoSmithKline K.K.,

*** Chugai Pharmaceutical Co., Ltd.

要 旨

用量反応関係を検討する薬効薬理試験において、検討したい化合物を複数用量設定し処置するほかに、陰性対照物質および陽性対照物質を処置する実験系がしばしば見られる。このような実験系では、陰性対照物質および陽性対照物質の複数のサンプルから得られたデータを平均化したものを、それぞれ最小または最大反応とみなして、各用量の反応を率として、50%有効用量(ED50)を求める解析がよく行われている。しかし、この方法では陰性および陽性対照物質の反応の誤差の考慮や、反応が0%~100%の範囲外となる場合の取扱いに苦慮する点がある。そこで、非線型回帰モデルにダミー変数を用いることで、ED50を直接推定する方法を検討した。この方法では、ED50のみならずそれ以外の推定値も直接推定することが可能であった。

キーワード：シグモイド曲線, ED50, 陰性対照, 陽性対照, SAS/NLIN, JMP

1. 目的

薬理的活性あるいは毒性用量の評価において、用量反応関係をモデル化し、ある反応となるような化学物質の濃度(用量)を求めるための方法として生物検定法が用いられる。生物検定法の代表的な方法としては、プロビット法による50%致死量の推定¹⁾や Emax モデル²⁾が知られている。

上記の実験系では、評価をしたい化学物質を複数用量設定して、用量反応関係を求めるが、最小反応や最大反応を設定するために、陰性対照物質や陽性対照物質を投与して、これらを反応の下限や上限とする場合がある。これらの対照物質は評価したい化合物と同様の取扱いが出来ないため、対照物質の反応の平均値を0%または100%の反応として、それに対する評価したい物質の反応を反応率として示すことで、2値データの用量反応関係から、ある反応となるような化学物質の濃度(用量)を求める方法がしばしば用いられている。

この方法を用いた場合、陰性または陽性対照物質の反応においてもバラツキが生じるにもかかわらずそれらを考慮しない点や、評価したい物質の反応が陰性または陽性対照物質の反応を超えるような場合に、反応の範囲が0%から100%の範囲に収まらないなど、取扱いに苦慮する点がある。

非線形回帰は、プロシ ज्याの中で、偏微分式の入力を必要とするなど、使い方が線形回帰に比べて難しかったが、最近では、自動的に偏微分式を proc NLIN が生成するようになり、使いやすくなってきた。また、JMP の非線形回帰では、収束過程がディスプレイされ、解を求めやすくなり、回帰分析と同様に気楽に使えるようになってきた。

以上より、上記のような実験系の解析に対して、陽性対照、陰性対照などを含むダミー変数を用いた非線形回帰モデルを用いることの有用性について検討した。

2. 非線形回帰モデル^{3,4)}

2.1. 計量値と2値のロジスティック回帰モデルの関連

計量値の反応にシグモイド曲線を当てはめるための関数として、ロジスティック関数、

$$y = \frac{Emax}{1 + \exp(\gamma(\ln(ED50) - \ln(x)))} \quad (1)$$

を用いる²⁾。ここで、 γ はロジスティック曲線の傾きをあらわすパラメータ、 $ED50$ は反応が50%の時の用量 x となるパラメータ、 $Emax$ は最大反応を表すパラメータである。

式(1)を変形すると

$$y = \frac{Emax}{1 + \exp(\gamma(\ln(ED50) - \ln(x)))} = \frac{Emax}{1 + \exp(-(-\gamma \ln(ED50) + \gamma \ln(x)))} \quad (2)$$

となり、 $-\gamma \ln(ED50) = \beta_0$ 、 $\gamma = \beta_1$ とおきかえ、さらに $\eta = \beta_0 + \beta_1 \ln(x)$ とすると、

$$y = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \ln(x)))} Emax = \frac{1}{1 + e^{-\eta}} Emax \quad (3)$$

が得られる。計量値のロジスティック関数は、反応率を p としたときに2値反応のロジスティック回帰、

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \ln(x)))} = \frac{1}{1 + e^{-\eta}} \quad (4)$$

と対応付けられる。ただし、 p は0から1の範囲であるが、 y は0から $Emax$ の範囲でシグモイド曲線を描く。

式(2)で、 $\gamma = \beta_1$ 、 $\ln(ED50) = \beta_2$ とおき、式(4)の2値のシグモイド曲線の形式にあわせると、

$$y = \frac{Emax}{1 + \exp(-\beta_1(-\beta_2 + \ln(x)))} \quad (5)$$

が得られる。以下、式(5)を拡張して用いることにする。

2.2. 陰性対照、および陽性対照がある場合の非線型回帰モデル

反応 y が計量値で、薬物濃度あるいは薬物量の $\ln(x)$ に対してシグモイド曲線となる実験系で、薬物量が0の場合に、これは、溶媒対照、コントロール群、あるいは、陰性対照などと呼ばれるが、薬物濃度0の対数はマイナス無限大となり、このままでは、シグモイド曲線を当てはめるために、このデータを用いることができない。

そこで、陰性対照のデータから推定される β_3 、さらに投与量を無限大まで増やしていった場合に得ら

れる反応を陽性対照のデータから推定される β_4 として、それらを同時に推定するために、以下の 3 つの式を同時に推定する非線型回帰モデルが必要となる。以下の式は、陰性対照に対して、陽性対照が大きい反応性を示す場合の例である。

$$\text{薬物濃度群: } y_i = \beta_3 + \frac{\beta_4 - \beta_3}{1 + \exp(-\beta_1(-\beta_2 + \ln(x)))} + e_i, \quad i = 1, 2, \dots, n_1 \quad (6)$$

$$\text{陰性対照: } y_j = \beta_3 + e_j, \quad j = 1, 2, \dots, n_2 \quad (7)$$

$$\text{陽性対照: } y_k = \beta_4 + e_k, \quad k = 1, 2, \dots, n_3 \quad (8)$$

これら 3 つの回帰式の誤差は、すべて平均 0、分散 σ^2 と共通であるとする。

これら 3 つの回帰式を同時に推定するためにダミー変数を用いる方法は、大森ら⁹⁾により報告されているが、陰性および陽性対照の有無や反応のパターンにより、複数の式が示されているため、更にダミー変数を追加することで、陰性および陽性対照の有無や反応パターンによらず、共通に利用可能な式を設定した。

$$y_i = \beta_3 \cdot d_1 + \left\{ \frac{|\beta_4 - \beta_3|}{1 + \exp(-\beta_1(-\beta_2 + \ln(x)))} + \beta_3 \cdot d_4 + \beta_4 \cdot d_5 \right\} \cdot d_2 + \beta_4 \cdot d_3 + e_i \quad (9)$$

$$i = 1, 2, \dots, (n_1 + n_2 + n_3)$$

$$d_1 \begin{cases} \text{陰性対照} & 1 \\ \text{その他} & 0 \end{cases} \quad d_2 \begin{cases} \text{薬物群} & 1 \\ \text{その他} & 0 \end{cases} \quad d_3 \begin{cases} \text{陽性対照} & 1 \\ \text{その他} & 0 \end{cases}$$

$$d_4 \begin{cases} \text{陰性対照} < \text{陽性対照} & 1 \\ \text{その他} & 0 \end{cases} \quad d_5 \begin{cases} \text{陰性対照} > \text{陽性対照} & 1 \\ \text{その他} & 0 \end{cases}$$

2.3. ED50 以外の推定

ロジスティック回帰の場合、ED50 は $p=0.5$ となり、以下の式で表される。

$$\text{logit}(0.5) = \ln\left(\frac{0.5}{1-0.5}\right) = \beta_0 + \beta_1 \ln(ED50) = 0$$

$$\ln(ED50) = -\frac{\beta_0}{\beta_1} \quad (10)$$

ED10 の場合、上記の式を変形すると、以下のように表される。

$$\text{logit}(0.1) = \ln\left(\frac{0.1}{1-0.1}\right) = \beta_0 + \beta_1 \ln(ED10) = -2.197$$

$$\ln(ED10) = -\frac{\beta_0 + 2.197}{\beta_1} = -\ln(ED50) - \frac{2.197}{\beta_1}$$

$$\ln(ED50) = -\ln(ED10) - \frac{2.197}{\beta_1} \quad (11)$$

ED10を直接推定するためには、(5)式の β_2 (ED50)の代わりに(11)を代入すればよい。

$$y = \frac{E \max}{1 + \exp(-\beta_1(-\ln(ED10) + \ln(x)) + 2.197)} \quad (12)$$

同様に、ED90の場合は、以下のように表される。

$$y = \frac{E \max}{1 + \exp(-\beta_1(-\ln(ED90) + \ln(x)) - 2.197)} \quad (13)$$

3. 解析例

上記の非線形回帰モデルを用いて、以下の2つの事例について解析を行った。

事例1:環境ホルモンEE (ethinyl estradiol) 投与後のラット子宮重量⁶⁾

表1および図1に、解析に用いたデータを示した。

表1 環境ホルモンEE (ethinyl estradiol) 投与後のラット子宮重量

No.	Vehicle	ethinyl estradiol (EE), $\mu\text{g}/\text{kg}$						
		0.01	0.03	0.1	0.3	1	3	10
1	102.35	95	105	112.22	190.45	319.78	373.72	382
2	120.82	115	115	123.47	217.48	351.32	384.72	404.32
3	115.92	115	120	144.42	213.95	326.07	378.37	354.37
7	121.62	120	125	131.25	220.83	317.52	387.43	391.67
8	79.22	90	80	105.08	211.13	287.68	262.2	273.73
9	108.47	115	115	123.6	211.37	357.57	353.82	362.05
11	82.45	100	100	113.38	191.23	297.67	307.6	312.4
18	89.25	90	90	91.8	193.07	334.95	334.48	366.2
19	99.17	100	100	83.17	104.67	135.17	234.17	332.67

投与量 0.01, 0.03 $\mu\text{g}/\text{kg}$ のデータは、グラフより推定した。

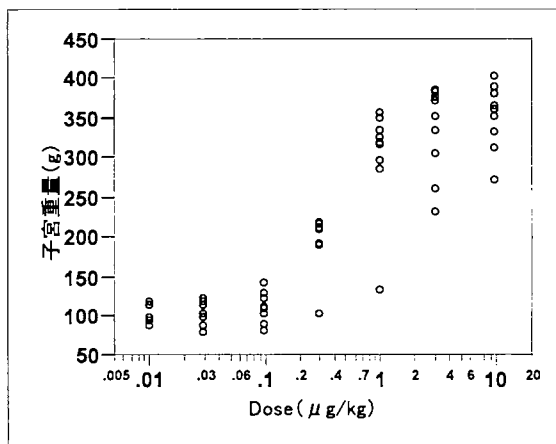


図1 環境ホルモンEE (ethinyl estradiol) 投与後のラット子宮重量
グラフ表示のため、Vehicleは0.001 $\mu\text{g}/\text{kg}$ に表示した。

この実験系では、陽性対照がないが、反応は右肩上がりのシグモイド曲線であることから、ダミー変数 d_4 は1、 d_5 は0となり、解析に用いるデータセットは、DATASET1に示される形式となる。なお、陰性対

照の投与量は、欠測値のままだと SAS の proc NLIN では推定するデータとして使用されないため、他の投与量と識別できる数値(今回は 0.001)を入力した。

DATASET1

dose	uterus	d1	d2	d3	d4	d5
0.001	100	1	0	0	1	0
0.001	102.35	1	0	0	1	0
:	:	:	:	:	:	:
0.01	95	0	1	0	1	0
0.01	115	0	1	0	1	0
:	:	:	:	:	:	:
10	366.2	0	1	0	1	0
10	332.67	0	1	0	1	0

} 陰性対照

} 薬物群

このデータセットを用いて、非線形回帰モデルの式(9)を、SAS の proc NLIN でプログラミングした結果を、PROGRAM1 に示す。なお、パラメータの初期値は、 β_1 および β_2 には 1 を、 β_3 には陰性対照の平均値の 102 を、 β_4 には最高投与量(10 μ g/kg)群の平均値の 353 を用いた。

PROGRAM 1 <<ED50 の直接推定>>

```
proc nlin data=dataset1 method=gauss;
  parms beta1=1 beta2=1 beta3=102 beta4=353;
  model uterus = beta3*d1+((abs(beta4-beta3)/(1+exp(-beta1*(-log(beta2)+log(dose)))))+
    beta3*d4+beta4*d5)*d2+beta4*d3;
run;
```

OUTPUT 1 に、SAS 8.2 で実行したパラメータの推定結果を示す。 β_1 は傾きの推定値、 β_2 は ED50 の推定値、 β_3 は陰性対照の推定値、 β_4 は陽性対照の推定値が得られ、各々の 95%信頼区間も同時に求められている。推定されたシグモイド曲線を重ね書きした結果を図 2 に示す。

OUTPUT 1 <<ED50 の直接推定>>

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	4	3711525	927881	189.74	<.0001
Residual	68	93796.7	1379.4		
Uncorrected Total	72	3805322			
Corrected Total	71	878941			

Parameter	Estimate	Approx		
		Std Error	Approximate 95% Confidence Limits	
beta1	1.7212	0.3392	1.0443	2.3981
beta2	0.4150	0.0560	0.3033	0.5266
beta3	101.5	7.2711	87.0159	116.0
beta4	348.7	10.7247	327.3	370.1

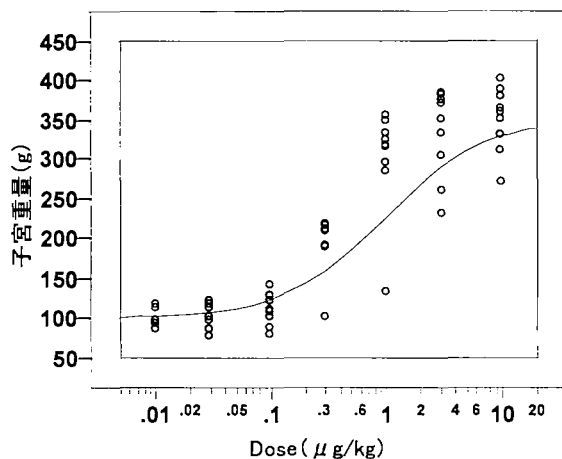


図 2 環境ホルモン EE (ethinyl estradiol) 投与後のラット子宮重量に対するシグモイド曲線

また、PROGRAM 1 の Model 式を PROGRAM 2 のように書き換えることで、ED10 および ED90 の直接推定も可能となる。

PROGRAM 2 <<ED10 および ED90 の直接推定>>

* ED10;

$$\text{model uterus} = \text{beta3} * \text{d1} + \left(\frac{\text{abs}(\text{beta4} - \text{beta3})}{1 + \exp(-\text{beta1} * (-\log(\text{beta2}) + \log(\text{dose}) + \log(0.1 / (1 - 0.1))))} \right) + \text{beta3} * \text{d4} + \text{beta4} * \text{d5} * \text{d2} + \text{beta4} * \text{d3};$$

* ED90;

$$\text{model uterus} = \text{beta3} * \text{d1} + \left(\frac{\text{abs}(\text{beta4} - \text{beta3})}{1 + \exp(-\text{beta1} * (-\log(\text{beta2}) + \log(\text{dose}) + \log(0.9 / (1 - 0.9))))} \right) + \text{beta3} * \text{d4} + \text{beta4} * \text{d5} * \text{d2} + \text{beta4} * \text{d3};$$

OUTPUT 2 にパラメータの推定結果を示す。

OUTPUT 2 <<ED10 および ED90 の直接推定>>

* ED10;

Parameter	Estimate	Approx		
		Std Error	Approximate 95% Confidence Limits	
beta1	1.7212	0.3392	1.0443	2.3981
beta2	0.1158	0.0296	0.0568	0.1748
beta3	101.5	7.2711	87.0160	116.0
beta4	348.7	10.7246	327.3	370.1

* ED90;

Parameter	Estimate	Approx		
		Std Error	Approximate 95% Confidence Limits	
beta1	1.7212	0.3392	1.0443	2.3981
beta2	1.4871	0.4647	0.5597	2.4145
beta3	101.5	7.2711	87.0162	116.0
beta4	348.7	10.7245	327.3	370.1

非線形回帰モデルの当てはめは、JMP を用いても同様に実施可能である。

JMP の場合には、DATASET1 に新たに列を追加し、(9)の Model 式を計算式として設定し、併せてパラメータの初期値を設定する必要がある。設定した計算式を図 3 に示す。なお、JMP 用いる DATASET では、陰性対照の投与量は欠測値とする。

$$B3 * d1 + \left(\frac{B4 - B3}{1 + \exp\left(-B1 * \left(-\log(B2) + \log(Dose)\right)\right)} \right) + B3 * d4 + B4 * d5 * d6 + B4 * d3$$

図 3 JMP における計算式（非線形回帰モデル：ED50 の場合）

非線形回帰を行うには、JMP 5.0.1a の場合、分析—モデル化—非線型回帰を選択し、Y、応答変数に実測値(今回は uterus)を、X、予測式列に計算式を設定した列(今回は uterus_d50)を指定して、アクション—OK を押すと、図 4 に示される「非線形回帰の当てはめ」メニューが表示されるので、実行を押すと収束する場合は、推定値が求められる。

▼ 非線形回帰の当てはめ

▼ 設定パネル

[実行]をクリックして開始。

実行	基準	現在	停止限界
停止	反復	0	60
ステップ	短縮	0	15
リセット	目的関数変化	1.340781e154	0.0000001
	パラメータ変化	1.340781e154	0.0000001
	勾配	1.340781e154	0.000001

パラメータ 現在値 ロック

beta1	1	<input type="checkbox"/>	SSE	.
beta2	1	<input type="checkbox"/>	N	0
beta3	102	<input type="checkbox"/>		
beta4	353	<input type="checkbox"/>		

図 4 JMP の「非線形回帰の当てはめ」の画面

図 5 に求められた推定値を示す。SAS の proc NLIN で実施した結果と推定値はほぼ一致したが、両側 95%信頼区間は SAS と JMP で若干異なった。この原因は、JMP における信頼区間の算出はプロファイル尤度を用いており、SAS の proc NLIN の信頼区間は“Wald based formula”のためである。

解					
	SSE	DFE	MSE	RMSE	
	93796.733008	68	1379.3637	37.139786	
パラメータ	推定値	近似標準誤差	下側信頼限界	上側信頼限界	
beta1	1.7212095531	0.33915151	1.173871	2.79820532	
beta2	0.4149558209	0.05597202	0.31918854	0.55690505	
beta3	101.52540185	7.27142582	86.6744515	115.49179	
beta4	348.71156227	10.7254009	328.174346	373.531549	

図 5 JMP での非線形回帰の結果 (ED50)

なお, JMP においても計算式を PROGRAM 2 のように変更するだけで, ED10 や ED90 の推定も可能となる。

事例 2: NR 法による細胞毒性試験⁵⁾

表 2 に解析に用いたデータを示した。

表 2 NR 法による細胞毒性試験データ

検体	陰性	0.02	0.04	0.05	0.06	0.07	0.08	0.1	0.12	ブランク
1	0.406	0.396	0.318	0.1	0.121	0.086	0.131	0.067	0.047	0
2	0.379	0.318	0.12	0.164	0.086	0.119	0.069	0.044	0.005	0
3	0.417	0.426	0.24	0.23	0.167	0.079	0.113	0.083	0.028	0
4	.	0.376	0.248	0.185	0.142	0.198	0.18	0.074	0.029	0
5	.	0.259	0.197	0.131	0.105	0.148	0.072	0.044	0.029	0
6	.	0.428	0.257	0.209	0.251	0.253	0.176	0.111	0.034	0
7	.	0.426	0.255	0.174	0.251	0.1	0.141	0.149	0.053	.
8	.	0.282	0.2	0.269	0.158	0.19	0.107	0.094	0.027	.
9	.	0.499	0.339	.	0.292	.	0.198	0.116	0.045	.
10	.	0.586	0.473	.	0.234	.	0.175	0.148	0.034	.
11	.	0.298	0.256	.	0.186	.	0.069	0.042	0.004	.

このデータでは, 投与量を増加させていったときの反応の延長にブランクの値があるため, ブランクを陽性対照と見なして解析を行った。

この実験系では, 反応は右肩下りのシグモイド曲線であることから, ダミー変数 d_4 は 0, d_5 は 1 となる。事例 2 についても, 事例 1 と同様の形式で DATASET を作成し, 解析を行った。

OUTPUT 3 に, SAS 8.2 で実行した ED50, ED10 および ED90 のパラメータの推定結果を示す。また, 推定されたシグモイド曲線を重ね書きした結果を図 6 に示す。

OUTPUT 3 <<ED50, ED10 および ED90 の直接推定>>

Source	DF	Squares	Square	F Value	Pr > F
Regression	4	4.3853	1.0963	116.20	<.0001
Residual	93	0.3612	0.00388		
Uncorrected Total	97	4.7464			
Corrected Total	96	1.7149			

* ED50

Parameter	Estimate	Approx		
		Std Error	Approximate	95% Confidence Limits
beta1	-2.1932	0.3679	-2.9237	-1.4626
beta2	0.0519	0.00470	0.0426	0.0613
beta3	0.4280	0.0277	0.3730	0.4830
beta4	-0.00419	0.0233	-0.0504	0.0420

* ED10

Parameter	Estimate	Approx		
		Std Error	Approximate	95% Confidence Limits
beta1	-2.1932	0.3679	-2.9237	-1.4626
beta2	0.1414	0.0244	0.0929	0.1899
beta3	0.4280	0.0277	0.3730	0.4830
beta4	-0.00419	0.0233	-0.0504	0.0420

* ED90

Parameter	Estimate	Approx		
		Std Error	Approximate	95% Confidence Limits
beta1	-2.1932	0.3679	-2.9238	-1.4626
beta2	0.0191	0.00396	0.0112	0.0269
beta3	0.4280	0.0277	0.3730	0.4830
beta4	-0.00419	0.0233	-0.0504	0.0420

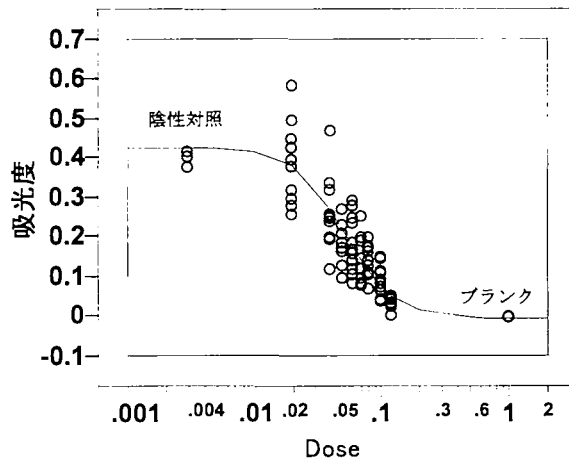


図6 陰性対照とブランクの吸光度を考慮したシグモイド曲線グラフ表示のため、陰性対照は 0.003、ブランクは 1.000 に表示した。

4. まとめ

用量反応関係を検討する薬効薬理試験において、用量反応関係がシグモイド曲線として表された場合に、ロジスティック曲線を当てはめ、ED50のような曲線の形状を示せる値を推定することは、得られた用量反応関係を簡潔に示す上で、有用な方法である。しかし、ロジスティック曲線を当てはめるにあたって、非線形回帰分析を適切に行うためには、妥当な回帰式の選択、収束を可能とする適切な初期値の設定、推定したいパラメータで回帰式を偏微分した式の設定などを行う必要があり、あまり一般的に用いられていないのが現状であった。

近年の SAS や JMP のバージョンアップにより、これらの問題をより簡単にクリアできるだけの周辺環境が整ってきつつあり、本報告で示したようにダミー変数を利用して非線形回帰分析を行う方法についても検討することが可能となった。

ダミー変数を用いた非線形回帰分析を用いることで、本報告のように複数のモデル式を同時に推定することが可能となる。このため、用量反応関係を検討する実験系において見られる、陰性対照物質および陽性対照物質の複数のサンプルから得られたデータを平均化したものを、それぞれ最小または最大反応とみなして、各用量の反応を率として ED50 を求める解析において取扱いに苦慮していた、陰性および陽性対照物質の反応の誤差や、反応率が 0%~100%の範囲外となる場合に対しても、これらを考慮した解析が可能であった。また、Model 式を変形することにより、ED50 のみならず、その他の推定値を直接推定することも可能であった。

これらのことから、ダミー変数を用いた非線形回帰分析は、用量反応関係を表すシグモイド曲線の当てはめを行う場合に、有用な方法であると考えられた。

さらに、ダミー変数を拡張することで、1 薬剤の用量反応曲線を推定するに留まらず、複数の薬剤の用量反応曲線を同時に推定することや、その効力比の推定にも応用が可能と考えられた。

なお、本報告では、SAS と JMP を用いて、非線形回帰分析を行ったが、使用面においてソフトウェアの特徴が現れた。

SAS は、強力なプログラミング言語を有することから、複数の MODEL 式の結果を簡便に得ることが可能であり、非線形回帰分析を繰り返し解くような場合においては、効率よく活用することが可能と考えられた。一方 JMP は、プログラミング言語を有するソフトウェアを使い慣れていない人においても、Template ファイルを準備することによって、GUI ベースで非線形回帰分析を行うことが可能であると考えられた。

文献

1. 吉村功 編著 (1987). 毒性・薬効データの統計解析—事例研究によるアプローチ—, サイエントリスト.
2. 佐久間昭 (1981). 薬効評価 II, 東京大学出版会.
3. Draper, N.R. and Smith, H. (1998). Applied Regression Analysis, 3rd ed., John Wiley & Sons.
4. Bates, D.M., and Watts, D.G. (1988). Nonlinear Regression Analysis and Its Applications. John Wiley & Sons.
5. 大森崇, 加藤麻矢子 (1998). 細胞毒性試験の ED50 推定法—原理, SAS プログラム, 使い方—, サイエントリスト.
6. Kano, J., Onyon, L., Haseman, J., et al. (2001). The OECD program to validate the rat uterotrophic bioassay to screen compounds for *in vivo* estrogenic responses: phase 1, Environmental Health Perspectives, 109(8):(785-784).

連絡先: 東京都文京区小石川 3-1-3, E-mail: masayuki.yamada@pharm.kissei.co.jp

計量値のシグモイド用量反応曲線の同時推定

—効力比とその95%信頼区間—

高橋 行雄
中外製薬株式会社 臨床解析部

Curve Fitting on Dose Response with Sigmoid Quantitative responses
—Estimation of Efficacy ratio and its 95% confidence—

Yukio Takahashi
Clinical Data Analysis Dept. / Chugai Pharmaceutical Co., Ltd.

要旨 新たに合成された化合物と、同じ実験条件下で標準物の対数用量に対する反応に対して平行なシグモイド曲線があてはめられたときに、ある一定の反応を得るための用量の比が定義できる。反応が2値の場合には、ダミー変数を用いたロジスティック回帰モデルにより、効力の差の antilog から効力比とその95%信頼区間が得られる。計量値の反応にシグモイド曲線をあてはめるためには、非線形回帰モデルを用いるのであるが、複数の平行なシグモイド曲線をあてはめるためには、ダミー変数を含む非線形回帰モデルに拡張する必要がある。ダミー変数の与え方により効力の差を直接求めることができ、その antilog から効力比と95%信頼区間が得られる。

キーワード：シグモイド曲線、EC50、効力比、SAS/NLIN、JMP、非線形回帰分析

1. 目的

薬理学的活性あるいは毒性用量の評価において、ある反応となるような化学物質の濃度（用量）を求めるための方法として生物検定法が知られている。反応が2値の *in vivo* の実験の場合にプロビット法による50%致死量の推定が生物検定法の代表的な方法であり、薬理学的活性を標準品に比較して新しい化合物の効力比を求める方法も定式化されている。反応が計量値で、用量反応関係が直線の場合に効力比を求める方法は平行線検定法として知られている^{1,2)}。用量の設定範囲を広く設定する *in vitro* 実験系では、用量反応関係がシグモイド状の曲線となり、反応が直線とならないことがしばしば経験される³⁾。この場合に、効力比とその95%信頼区間を“平行線検定法”と同様の考え方で求められれば、実験結果を簡潔な要約統計量として示すことができる。

2. ヒスタミン誘発収縮反応

G薬のモルモット摘出回腸のヒスタミン誘発収縮反応におよぼす作用についての *in vitro* 実験を取り上げる。実験は、表1に示すように4×4のラテン方格で行われた。

表1 実験デザイン (ラテン方格, G 薬の濃度)

実験日	モルモット 番号	胃側 <-----> 肛門側			
		部位 1	部位 2	部位 3	部位 4
1	1	A: 0 μ M	B: 0.01 μ M	C: 0.1 μ M	D: 1 μ M
1	2	B: 0.01 μ M	C: 0.1 μ M	D: 1 μ M	A: 0 μ M
2	3	C: 0.1 μ M	D: 1 μ M	A: 0 μ M	B: 0.01 μ M
2	4	D: 1 μ M	A: 0 μ M	B: 0.01 μ M	C: 0.1 μ M

実験手順

- 手順 1) 1 匹目のモルモットから回腸を摘出し, 一本の長さが約 20mm となるように 4 本の標本を作製する.
 標本は, 胃側から肛門側へ 1~4 の部位番号を付与する.
- 手順 2) 4 連のマグヌス装置に標本を 1 ずつ懸垂し, それぞれヒスタミン濃度が 300 μ M となるまで累積的に添加し, 懸垂した回腸の最大収縮高を添加前値とする.
- 手順 3) 回腸中のヒスタミンを洗浄する.
- 手順 4) 4 連のマグヌス装置に, それぞれ蒸留水 (G 薬 0 μ M), G 薬の 0.01, 0.1, 1.0 μ M の順にセットする.
- 手順 5) マグヌス装置にヒスタミン濃度が 0.01 μ M となるように添加し, 回腸の収縮が止まってから, 次にヒスタミン濃度が $\sqrt{10}$ 倍, 0.0316 μ M となるように添加する. この累積的添加をスタミン濃度が 316 μ M となるまで繰り返す. この間の回腸の収縮高をキモグラフ (kymograph, 筋肉の運動や心臓の拍動などを記録する装置) に連続的に記録する.
- 手順 6) 2 匹目のモルモットについて手順 1 からの操作を繰り返す. ただし, ヒスタミンの注入順は表 1 に示した手順 4 で, G 薬の 3 用量を先に行い, 蒸留水 (G 薬 0 μ M) は最後とする.

今回は, この実験のモルモット番号 2 について検討した. 実験から得られたデータを表 2 に示す.

表 2 G 薬のモルモット摘出回腸のヒスタミン誘発収縮反応に及ぼす作用

G 薬 用量 μ M	最大 収縮高	ヒスタミンの用量 (μ M)									
		0.01	0.0316	0.1	0.316	1	3.16	10	316	100	316
0	165	1	3	5	23	66	113	158	171	171	165
0.01	158	1	1	3	9	50	98	141	165	170	169
0.1	118	0	1	1	2	25	10	46	96	122	127
1	163	0	0	1	0	1	2	6	54	120	136

生データはキモグラフの目盛りから読み取っているので実際の長さではない.

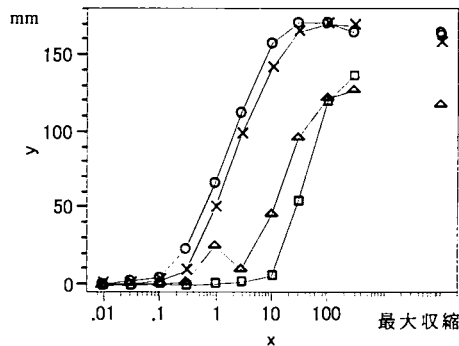


図 1 最大収縮高を含む 4 本の用量反応曲線

○: G 薬 0 μ M, ×: G 薬 0.01 μ M, △: G 薬 0.1 μ M, □: G 薬 1 μ M

3. 非線形回帰モデル^{4,5)}

計量値と2値のロジスティック回帰モデルの関連

計量値の反応にシグモイド曲線をあてはめるための関数として、ロジスティック関数、

$$y = \frac{Emax}{1 + \exp(\gamma(\ln(EC50) - \ln(x)))} \quad (1)$$

を用いる⁶⁾。ここで、 γ はロジスティック曲線の傾きをあらわすパラメータ、 $EC50$ は反応が 50% の時の用量 x となるパラメータである。式 (1) を変形すると

$$y = \frac{Emax}{1 + \exp(\gamma(\ln(EC50) - \ln(x)))} = \frac{Emax}{1 + \exp(-(-\gamma \ln(EC50) + \gamma \ln(x)))} \quad (2)$$

となり、 $-\gamma \ln(EC50) = \beta_0$ 、 $\gamma = \beta_1$ とおきかえると、2 値のロジスティック回帰式に $Emax$ を掛けた式が得られる。

$$y = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \ln(x)))} Emax \quad (3)$$

式 (2) で、 $\gamma = \beta_1$ 、 $\ln(EC50) = \beta_2$ とおくと、

$$y = \frac{Emax}{1 + \exp(-(\beta_1(-\beta_2 + \ln(x))))} \quad (4)$$

が得られる。以下、式 (4) を拡張して用いることにする。

陰性対照、および、陽性対象がある場合の効力比の推定

反応 y が計量値で、薬物濃度あるいは薬物量の $\ln(x)$ に対してシグモイド曲線となる実験系で、薬物量が 0 の場合に、これは、溶媒対照、コントロール群、あるいは、陰性対照などと呼ばれるが、薬物濃度 0 の対数はマイナス無限大となり、このままでは、シグモイド曲線をあてはめるために、このデータを用いることができない。

薬物濃度が低くなった場合に反応 y が、陰性対照に近づくようなシグモイド曲線のあてはめにおいて、下限値として推定される β'_3 と、陰性対照のデータの平均値から下限値として推定される β''_3 のように、推定値が 2 通り存在することになる。これらを 1 つの下限値の推定値とするために、2 つのモデル式を同時に含むような非線型回帰モデルが必要となる。さらに、陽性対照がある場合も同様に、シグモイド曲線のあてはめで推定される上限値 β'_4 と、陽性対照のデータから推定される上限値 β''_4 を 1 つの上限値として同時に推定するようにしたい。

$$\text{薬物濃度群： } y_i^* = \beta'_3 + \frac{\beta'_4 - \beta'_3}{1 + \exp\{-\beta_1(-\beta_2 + \ln(x_i))\}} + e_i^* \quad , \quad i = 1, 2, \dots, n_1 \quad (5)$$

$$\text{陰性対照： } y_j^{\text{陰性}} = \beta''_3 + e_j^{\text{陰性}} \quad , \quad j = 1, 2, \dots, n_2 \quad (6)$$

$$\text{陽性対照： } y_k^{\text{陽性}} = \beta''_4 + e_k^{\text{陽性}} \quad , \quad k = 1, 2, \dots, n_3 \quad (7)$$

これらの 3 つの回帰式の誤差は、すべて平均 0、分散 σ^2 と共通であるとする。ダミー変数 d_i :

(陰性対照の場合に 1, それ以外は 0), ダミー変数 d_2 : (薬物濃度群の場合に 1, それ以外は 0), ダミー変数 d_3 : (陽性対照の場合に 1, それ以外は 0), を考え, 式 (5), (6), (7) を併合した次式を得る. これにより, シグモイド曲線の共通の下限值 β_3 および上限値 β_4 を推定することができる.

$$y_i = \beta_3 \cdot d_1 + \left(\beta_3 + \frac{\beta_4 - \beta_3}{1 + \exp\{-\beta_1(-\beta_2 + \ln(x_i))\}} \right) \cdot d_2 + \beta_4 \cdot d_3 + e_i, \quad i = 1, 2, \dots, (n_1 + n_2 + n_3) \quad (8)$$

複数のシグモイド曲線の同時推定

このモデルをさらに拡張して, 複数のシグモイド曲線の同時あてはめができるように拡張する. その際に, 複数のシグモイド曲線のパラメータ, 傾き β_1 , 左右の位置 β_2 , 下限値 β_3 , 上限値 β_4 のうち, どれが複数のシグモイド曲線に共通で, どれが異なるのかを, 実験前に規定しておく必要がある. 表 2 のデータは, 図 1 から次に示すように,

傾き (β_1):	同じ
左右の位置 (β_2):	異なる
下限値 (β_3):	定数 = 0
上限値 (β_4):	異なる

2 つのパラメータ, 左右の位置 β_2 , および上限値 β_4 が異なるシグモイド曲線のあてはめが必用である. 左右の位置 (β_2) が異なる場合に, β_2 を次に示すように,

$$\beta_2 = \beta_{2,1}z_1 + \beta_{2,2}z_2 + \beta_{2,3}z_3 + \beta_{2,4}z_4 \quad (9)$$

複数のシグモイド曲線を識別するインディケータ型ダミー変数に展開できる. これと同様に, 上限値 β_4 も,

$$\beta_4 = \beta_{4,1}z_1 + \beta_{4,2}z_2 + \beta_{4,3}z_3 + \beta_{4,4}z_4 \quad (10)$$

と展開する. 下限値 β_3 は, この実験系では, 常に 0 であるので, 式 (8) の β_3 に 0 を代入し, 次を次のように簡単化できる.

$$y_i = \frac{\beta_4}{1 + \exp\{-\beta_1(-\beta_2 + \ln(x_i))\}} \cdot d_2 + \beta_4 d_3 + e_i \quad (11)$$

実際の計算に際しては, 式 (11) の, β_2 と β_4 をダミー変数を含む式 (9) と (10) に置き換える.

4. EC50 の直接推定

ダミー変数を 2 種類含む非線型回帰式の作成は, 煩雑なので解析用のデータを作成する SAS プログラムを Program1 に示す. ダミー変数 z_1 , z_2 , z_3 , および, z_4 は, G 薬に対するインディケータ型ダミー変数とし, “切片” z_0 も加えてある. 第 2 番目のインディケータ型ダミー変数は, 陰性対象 d_1 , 薬物濃度群 d_2 , および, 陽性対照 d_3 である.

Program 1 <<SAS データセットの作成>>

```

title 'drug_G 2003-05-06 Y.Takahashi' ;
data d01 ;
  input G_dose @@ ;
  z0=1; z1=(G_dose=1); z2=(G_dose=2); z3=(G_dose=3); z4=(G_dose=4);
  do x = 99999, 0.01, 0.0316 ,0.1, 0.316, 1, 3.16, 10, 31.6, 100, 316 ;
    ln_x = log(x) ;
    input y @@ ;
    output ;
  end ;
datalines ;
1 165 1 3 5 23 66 113 158 171 171 165
2 158 1 1 3 9 50 98 141 165 170 169
3 118 0 1 1 2 25 10 46 96 122 127
4 163 0 0 1 0 1 2 6 54 120 136
;
data d02 ;
  retain top ;
  set d01 ;
  d1=0; d2=0 ; d3=0 ;
  if x=99999. then d2=1 ;
  if x =99999. then do; d3=1; top=y; end;
  y_percent = y / top *100. ;

```

Output 1 <<解析用 SAS データセット>>

OBS	top	G_dose	z0	z1	z2	z3	z4	x	ln_x	y	d1	d2	d3	y_percent
1	165	1	1	1	0	0	0	99999.00	11.5129	165	0	0	1	100.000
2	165	1	1	1	0	0	0	0.01	-4.6052	1	0	1	0	0.606
3	165	1	1	1	0	0	0	0.03	-3.4546	3	0	1	0	1.818
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
12	158	2	1	0	1	0	0	99999.00	11.5129	158	0	0	1	100.000
13	158	2	1	0	1	0	0	0.01	-4.6052	1	0	1	0	0.633
14	158	2	1	0	1	0	0	0.03	-3.4546	1	0	1	0	0.633
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
23	118	3	1	0	0	1	0	99999.00	11.5129	118	0	0	1	100.000
24	118	3	1	0	0	1	0	0.01	-4.6052	0	0	1	0	0.000
25	118	3	1	0	0	1	0	0.03	-3.4546	1	0	1	0	0.847
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
34	163	4	1	0	0	0	1	99999.00	11.5129	163	0	0	1	100.000
35	163	4	1	0	0	0	1	0.01	-4.6052	0	0	1	0	0.000
36	163	4	1	0	0	0	1	0.03	-3.4546	0	0	1	0	0.000
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
44	163	4	1	0	0	0	1	316.00	5.7557	136	0	1	0	83.436

非線型の式 (9), (10), (11) を, SAS の proc NLIN でコーディングした結果を Program2 に示す. 反復計算法は標準的なガウス・ニュートン法を用いている. SAS 6.12 以前は 1 次の導関数をプログラムに含める必要があったのであるが, SAS 6.12 より内部計算されるようになったので指定していない. β_1 および β_{2i} の初期値は, 0 以上 100 未満の y のパーセントデータをロジット変換し単回帰分析した傾きから推定し, β_{4i} の初期値は, 表 2 の最大収縮高とした.

Output2 にパラメータの推定結果を示す. beta1 が傾きの推定値, beta2_1, ..., beta2_4 が左右の位置 EC50 の推定値, beta4_1, ..., beta4_4 が, それぞれの G 薬の濃度群での上限値の推定結果になっている. Output2 で得られた推定値から, G 薬の群ごとにシグモイド曲線を推定し, 生データに重ね書きした結果を図 2 に示す.

Program 2 << EC50の直接推定 >>

```
Title2 '<<< direct estimation >>>';
proc nlin data=d02 method=gauss;
  Parns  beta1=1.1
          beta2_1=0.2 beta2_2=0.8 beta2_3=2.1 beta2_4=4.0
          beta4_1=165 beta4_2=158 beta4_3=118 beta4_4=163;
  beta2 = beta2_1*z1 + beta2_2*z2 + beta2_3*z3 + beta2_4*z4;
  beta4 = beta4_1*z1 + beta4_2*z2 + beta4_3*z3 + beta4_4*z4;
  model y = (beta4 / (1 + exp(-beta1*(-beta2 + log(x)))) ))*d2 + beta4*d3;
run;
```

Output 2 << EC50の直接推定 >>

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	9	415458	46162.0	1265.37	<.0001
Residual	35	1276.8	36.4811		
Uncorrected Total	44	416735			
Corrected Total	43	207113			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta1	1.2843	0.0735	1.1352	1.4335
beta2_1	0.4240	0.0926	0.2361	0.6120
beta2_2	0.8070	0.0947	0.6147	0.9992
beta2_3	2.6386	0.1316	2.3715	2.9058
beta2_4	3.9664	0.1109	3.7414	4.1915
beta4_1	168.8	2.9573	162.8	174.8
beta4_2	166.2	3.0767	160.0	172.4
beta4_3	126.2	3.8739	118.3	134.0
beta4_4	159.7	4.9824	149.6	169.8

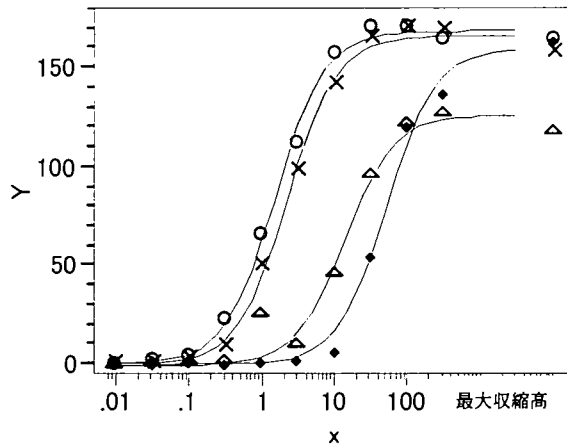


図2 収縮高を考慮したシグモイド曲線

○: 蒸留水, ×: G薬 0.01 μM, △: G薬 0.01 μM, ◆: G薬 0.01 μM

5. 効力比の推定

次に効力比を求める。ダミー変数 z_1 を“切片” z_0 に置き換えることにより、ダミー変数 z_2 , z_3 , および、 z_4 の回帰係数 (パラメータ) が、 $0\mu\text{M}$ 群と各 G 薬との差となり、antilog をとるこ

とにより、効力比が求まる。Program3にSASのプログラムを、Output3に推定されたパラメータを示す。beta2_0が0 μ M群のEC50の推定値、beta2_2, beta2_3, beta2_4がそれぞれの蒸留水とG薬の濃度群でのEC50の差の推定値となっている。beta4_0は、0 μ M群の上限値の推定値となって、beta4_2, beta4_3, beta4_4が、0 μ M群とそれぞれのG薬の濃度群での上限値との差の推定値になっている。95%信頼区間から、beta4_2, および beta4_4 は、それぞれ(-10.96, 5.82), (-20.51, 2.34)と0を含んでいるので、統計的には、差がないことが分かる。beta4_3は、推定値が-42.59, 95%信頼区間は、(-52.06, -33.05)と明らかに差があることが示されている。表3にOutput2とOutput3の結果をまとめ、antilogから元の用量でのEC50および効力比、その95%信頼区間を示した。表4には各群の上限値 β_4 の推定値を示した。こちらは、対数を取っていないので、推定された結果のままである。

表3 収縮高を考慮した効力比

位置	ln(EC50)	EC50(μ M)	差	ln(差)	倍	95%cl L	95%cl U	95%cl 倍
$\beta_{2,1}$	0.424	1.53	$\beta_{2,0}$	-				
$\beta_{2,2}$	0.807	2.24	$\beta_{2,2}$	0.383	1.5	0.116	0.650	(1.1, 1.9)
$\beta_{2,3}$	2.639	13.99	$\beta_{2,3}$	2.215	9.2	1.890	2.539	(6.6, 12.7)
$\beta_{2,4}$	3.966	52.80	$\beta_{2,4}$	3.542	34.6	3.252	3.833	(25.8, 46.2)

表4 収縮高を考慮した上限値の差

	推定値	差	ln(差)	95%cl L	95%cl U
$\beta_{4,1}$	168.8	$\beta_{4,0}$	-		
$\beta_{4,2}$	166.2	$\beta_{4,2}$	-2.6	-10.9	5.8
$\beta_{4,3}$	126.2	$\beta_{4,3}$	-42.6	-52.2	-33.0
$\beta_{4,4}$	159.7	$\beta_{4,4}$	-9.1	-20.5	2.3

Program 3 <<効力比>>

```
Title2 ' <<< difference >>> ' ;
proc nlin data=d02 method=gauss ;
  parms beta1=1.1
        beta2_0=0.2 beta2_2=0.6 beta2_3=1.9 beta2_4=3.8
        beta4_0=165 beta4_2=-7. beta4_3=-47. beta4_4=-2.0 ;
  beta2 = beta2_0*z0 + beta2_2*z2 + beta2_3*z3 + beta2_4*z4 ;
  beta4 = beta4_0*z0 + beta4_2*z2 + beta4_3*z3 + beta4_4*z4 ;
  model y = (beta4 / (1 + exp(-beta1*(-beta2 + log(x)))) ) ) * d2 + beta4*d3 ;
run ;
```

Output 3 <<効力比>>

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	9	415458	46162.0	1265.37	<.0001
Residual	35	1276.8	36.4811		
Uncorrected Total	44	416735			
Corrected Total	43	207113			

Output 3 続き

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta1	1.2843	0.0735	1.1352	1.4335
beta2_0	0.4240	0.0926	0.2361	0.6120
beta2_2	0.3829	0.1316	0.1158	0.6501
beta2_3	2.2146	0.1599	1.8901	2.5392
beta2_4	3.5424	0.1431	3.2519	3.8330
beta4_0	168.8	2.9573	162.8	174.8
beta4_2	-2.5750	4.1336	-10.9666	5.8165
beta4_3	-42.5950	4.7416	-52.2208	-32.9691
beta4_4	-9.0820	5.6268	-20.5050	2.3409

6. 収縮量を変化率にした場合

上限値が群によって異なるシグモイド曲線のあてはめを前節で示したのであるが、上限が群によって同じ場合もある。表2のデータで、陽性対象としての最大収縮高を100%とし、それぞれの収縮量を収縮率に変換した場合は、推定したいパラメータは、次に示すように、左右の位置 β_2 のみが変化することになる。

- 傾き(β_1): 同じ
- 左右の位置(β_2): 異なる
- 下限値(β_3): 定数 = 0
- 上限値(β_4): 定数 = 100

さらに、陰性および陽性対象もないので式(8)は、さらに簡単化でき、式(12)となる。SASによるプログラムをProgram4に、結果をOutput4に示す。

$$y = \frac{100}{1 + \exp(-(\beta_1(-(\beta_{2,0}z_0 + \beta_{2,2}z_2 + \beta_{2,3}z_3 + \beta_{2,4}z_4) + \ln x)))} \quad (12)$$

Program 4 << 収縮率を用いた場合の効力比 >>

```
Title2 '<< percent, difference >>';
proc nlin data=d02 method=gauss;
  where d2=1;
  Parms beta1=1.1
         beta2_0=0.2 beta2_2=0.6 beta2_3=1.9 beta2_4=3.8;
  beta2 = beta2_0*z0 + beta2_2*z2 + beta2_3*z3 + beta2_4*z4;
  model y_percent = 100 / (1 + exp(-beta1*(-beta2 + log(x)))));
run;
```

Output 4 << 収縮率を用いた場合の効力比 >>

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	5	137420	27484.0	1032.67	<.0001
Residual	35	931.5	26.6145		
Uncorrected Total	40	138352			
Corrected Total	39	74159.9			

Output 4 続き

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits
beta1	1.3740	0.0963	1.1784 1.5696
beta2_0	0.3728	0.1157	0.1378 0.6077
beta2_2	0.3173	0.1637	-0.0151 0.6497
beta2_3	2.1352	0.1636	1.8031 2.4673
beta2_4	3.6265	0.1639	3.2938 3.9593

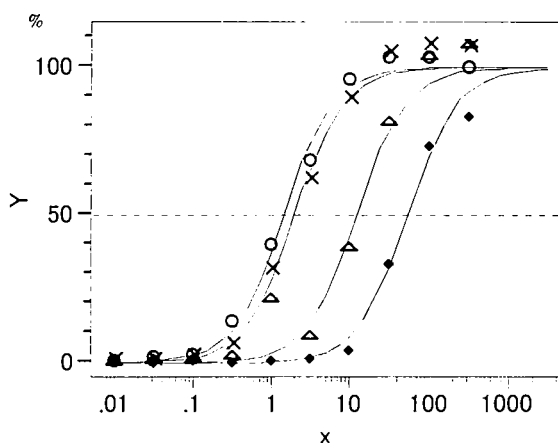


図3 傾きが共通でEC50が異なるのシグモイド曲線

○: 蒸留水, ×: G薬 0.01 μM, △: G薬 0.01 μM, ◆: G薬 0.01 μM

Output4のパラメータ(回帰係数)を整理し, antilogの計算を追加し, 元の濃度でのEC50, G薬の0μMを基準にした場合のEC50の比(倍)を表5に示す. G薬の最小用量0.01μMで95%信頼区間の対数の下限がわずかに0を下回っている. 統計的には, 「有意差なし」であるが, シグモイド曲線のわずかなずれを反映している.

表5 効力比

	ln(差)	倍	95%cl L	95%cl U	95%cl 倍
$\beta_{2,0}$	-				
$\beta_{2,2}$	0.317	1.4	-0.015	0.650	(0.90, 1.9)
$\beta_{2,3}$	2.135	8.5	1.803	2.467	(6.1, 11.8)
$\beta_{2,4}$	3.627	37.6	3.294	3.959	(26.9, 52.4)

7. 考察

シグモイド状の用量反応が計量値として得られた場合に, ロジスティック曲線をあてはめ, EC50 (ED50) あるいは EC10 (ED10) などの要約統計量を推定することは, 実験結果を簡潔に示すために有用な方法である. 応用分野は, 多岐にわたっている. 最近の問題では, 各種の環境ホルモンに対して in vivo での実験が国際的な共同研究⁷⁾として行われ, この実験データを要約するため

の統計量として ED_{10} および ED_{90} が用いられている。しかし、本質的に非線形の問題であり、一般的に使用されているとはいえない。

非線形回帰分析を適切に行うためには、妥当な回帰式の選択、初期値の設定、推定したいパラメータで回帰式を微分した式などを設定する必要があるが、線形回帰分析のように手軽に使うことができにくい。このために、手軽にシグモイド曲線のあてはめを行うために SigmaPlot, GraphPad, WinNonlin などの専用の統計ソフトが整備されてきた。しかしながら、これらの専用統計ソフトはシグモイド曲線を一本ごとにあてはめることを基本としていて、2つのシグモイド曲線の同時あてはめは標準機能に含まれていないので効力比の算出、陰性対照あるいは陽性対照があるような場合には手軽に使うことができない。

そこで、SAS あるいは JMP などの汎用ソフトの非線形回帰分析を用いて、これらの問題を定式化することにした。本報告では SAS/NLIN のプログラムと結果を示したが、シグモイド曲線の作図のためには、JMP の非線形回帰分析による推定結果を用いて行った。汎用ソフトを使い慣れていない人達が、非線形の問題を解くためには、GUI ベースの JMP が薦められ、SAS/NLIN は、非線形の問題を繰り返し解くための計算エンジンとして使うような場合に薦められる。

今回は、1 個体に割り付けられた G 葉の 4 用量間の比較を扱ったが、実際の実験データの解析は 4 個体分あり、これを同時に解析する場合には、個体を変量効果とするような非線形混合モデル SAS/NLMIXED を使う問題となり、更なる検討が必要である。

反応が直線的な場合は、線形回帰分析の問題であるが、線形回帰分析は一般的には効力比を求めるための基本である逆推定（ある Y の反応を得るための X の値を推定する）の問題を取り扱っていないために、線形回帰分析のパラメータの推定値および分散共分散行列から、デルタ法あるいはフィラーの方法などで別途計算する必要がある。しかし、反応が直線的な場合であっても、 $y = \beta_0 + \beta_1 x$ を、 $y = \beta_1(\beta_0/\beta_1 + x)$ と傾き β_1 を共通項とし、 $\beta_2 = -(\beta_0/\beta_1)$ とおくと式 (4) の指数項と同様の形 $y = \beta_1(-\beta_2 + x)$ となる。式 (9) のように、 β_2 をダミー変数を含めた式に置き換えることにより、複数の反応直線間の効力比とその信頼区間が、非線形回帰分析の標準機能で求めることができる。

文献

1. 佐久間昭 (1977). 薬効評価 I, 東京大学出版会.
2. Finney, D.J (1978). Statistical Method in Biological Assay, 3rd ed., Charles Griffin.
3. 大森崇, 加藤麻矢子 (1998). 細胞毒性試験の ED_{50} 推定法—原理, SAS プログラム, 使い方—, サイエнтиスト.
4. Draper, N.R, and Smith, H. (1998). Applied Regression Analysis, 3rd ed., John Wiley & Sons.
5. Bates, D.M., and Watts, D.G. (1988). Nonlinear Regression Analysis and Its Applications. John Wiley & Sons.
6. 佐久間昭 (1981). 薬効評価 II, 東京大学出版会.
7. Kano, J., Onyon, L., Haseman, J., et al. (2001). The OECD program to validate the rat uterotrophic bioassay to screen compounds for *in vivo* estrogenic responses: phase 1, Environmental Health Perspectives, 109(8):(785-784).

口頭論文発表
チュートリアル

生存時間解析における症例数設計

○ 浜田知久馬* 藤井陽介*

* 東京理科大学工学部経営工学科

Sample size design for survival analysis

Chikuma Hamada and Yosuke Fujii

Tokyo University of Science

1-3,Kagurazaka,Shinjyuku-ku, Tokyo, 162-8601

要旨

最近では癌の臨床試験以外でも、あるイベントが起きるまでの時間を主要な解析対象とした臨床研究が増えている。このような試験の症例数設計は、多くの場合ログランク検定に基づいて行われる。このための公式として Schoenfeld 式, Freedman 式が有名であり、実際に多く用いられている。

本稿では、チュートリアルとして、計量データに基づく症例数設計の一般論を示し、これと対比して生存時間解析の場合の特徴を説明する。次に、Schoenfeld 式, Freedman 式を理解するための理論的背景を解説し、2つの式の違い、ログランク検定との関連、SAS のプログラム・コーディング例を示す。また実際の臨床研究の例数設計では、様々な拡張が必要である。非劣性試験、患者登録期間が存在する、患者のリクルートが一定でない、途中脱落が存在する、比例ハザード性が成り立たない、多群で行われる、プライマリーな解析方法がログランク検定でなくウイルコクソン検定等である等の場合が生じ得る。このような場合の症例数設計についても方針を解説する。

キーワード：症例数設計, Schoenfeld 式, Freedman 式, ログランク検定, LIFETEST

論文概略

生存時間解析の症例数設計の原理, 実例, SAS のコーディング例, 適用上の注意について解説する。特に Schoenfeld, Freedman 式について詳述する。また比例ハザード性が成り立たない場合や、登録期間が存在する場合の拡張について述べる。

1. はじめに

生存時間解析の手法は大きく 3 種類に分類できる。特定の分布を仮定せずに生存時間分布の記述・検定を行うノンパラメトリック手法、Cox の比例ハザードモデルに基づき、生存時間分布とは独立に共変量の影響を評価するセミパラメトリック手法、ワイブル分布等の特定の生存時間分布を仮定したパラメトリック手法である。SAS ではそれぞれの解析用に、LIFETEST、PHREG、LIFEREG が用意されており、これらの手法は、現在では医薬統計の標準的な手法として定着している。最近では癌以外の領域でも、脳疾患や心疾患等のイベントが起きるまでの時間をエンドポイントとする臨床研究が増えている。これらの研究をデザインする際には、統計学的な症例数設計が行われる。SAS のバージョン 9 からは、計量データや二値データについて、例数設計を行うための POWER プロシジャ、計量データを対象に対比や交互作用項等のより複雑な仮説で例数設計を行うための GLMPOWER プロシジャが加わった。残念ながらこれらのプロシジャでは生存時間解析の例数設計を行うことはできない^{1),2)}。生存時間解析では、症例数そのものが直接、精度に影響を与えるわけではない。いくら症例数が多くてもフォローアップ期間が短い場合は、イベント数は少なく情報量はあまり大きくない。生存時間解析では統計的な精度を保証するために必要なイベント数を求め、フォローアップ期間からイベントを起こす割合を見積もって、必要な症例数を算定する。

2. 例数設計の原理

2 群の並行群試験で正規分布型の計量データについて、t 検定を行う場合の症例数設計では、次の 4 種類の条件を決める必要がある。

表 1 : 例数設計を行う際に必要な条件

α	:	検定の有意水準 (通常は 5%)
β	:	差を見逃す確率 (通常は 20%)
SD	:	個体間のばらつきの大きさ
Δ (デルタ)	:	予想される 2 群間の平均値の差 (生物学的に検出する価値がある差)

α , β は適用する検定の精度, SD は研究デザインによって規定されるバラツキの大きさ, Δ は比較したい治療群間の実力の違いを表す指標である。 α については、通常は 0.05 (片側検定の場合は 0.025), β については 0.10~0.20 が用いられることが多い。このとき 1 群あたり必要な例数は(1)式で与えられる。

$$N = \frac{2 \{z_\alpha + z_\beta\}^2 SD^2}{\Delta^2} \quad (1)$$

ここで z_α と z_β は、それぞれ標準正規分布の上側 α 点と、 β 点を表している。片側検定を $\alpha = 0.025$ (片側検定), $\beta = 0.20$ を行う場合、正規分布の数値表を調べてみると、 $z_{0.025} = 1.96$, $z_{0.20} = 0.84$ となる。ちなみに両側検定の場合は (1) 式で z_α を $z_{\alpha/2}$ に置き換えればよい。例えば、SD が 20、 Δ が 10 のときは

$$\begin{aligned} N &= 2\{1.96 + 0.84\}^2 \times 20^2 / 10^2 \\ &= 31.4 \end{aligned}$$

切り上げると 1 群あたり 32 例、2 群合わせると 64 例必要になる。この式は対応のない t 検定を有意水準 α で行ったとき、平均値の差 $\Delta = 10$ が見逃される確率が β になるように例数設計を行ったものである。厳密にいうと、t 分布を正規分布で近似していることになるが、通常の第 III 相の臨床試験のように、全体で数百例以上になれば、正規分布で十分精度よく近似できる。t 検定では、2 つの群の平均値の差をその標準誤差で除したものが検定統計量になる。t 統計量に対立仮説の下での Δ と SD を代入すると次のようになる。

$$t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\hat{\sigma}^2/N + \hat{\sigma}^2/N}} = \frac{\Delta}{\sqrt{SD^2/N + SD^2/N}} = \frac{\Delta}{SD\sqrt{2/N}} \quad (2)$$

t 統計量は対立仮説が正しいときこの値を中心に分布することになる。ある程度例数が多くなれば t 分布は正規分布で近似できるので、この t が正規分布の上側 α 点を上回れば有意になる。

$$\frac{\Delta}{SD\sqrt{2/N}} = z_\alpha \quad (3)$$

とにおいて、 Δ , SD, α を決めれば、未知数は N だけになり、t 検定でちょうど有意にするために必要な例数を求めることができる。ただしこのように N を設定しても、t 統計量は確率変数でありバラツキを伴うので、いつも z_α を越えるとは限らない。実は (3) 式に基づいて N を定めると、t 統計量は z_α を中心に分布するので、 z_α を越えるか越えないかは五分五分の確率になる。したがって差を見逃す β エラーの確率は 50% となる。これでは見逃しの確率が大きすぎてしまう。見逃しのエラーの大きさを β に抑えたい場合は、 z_α に z_β だけ保険を加えて、t 統計量が $z_\alpha + z_\beta$ を中心に分布するように N を設定する。すなわち、(4) 式を

$$\frac{\Delta}{SD\sqrt{2/N}} = z_\alpha + z_\beta \quad (4)$$

Nについて解くと、(1)式が導かれる。

標準正規分布は0を中心に分布するため、分布の50%点 $z_{0.50}$ は0になる。 $\beta=0.50$ に設定すれば、 β に由来する項はなくなる。したがって先ほどの例数設計は、 $\beta=0.50$ の特殊な場合と考えることもできる。

実は用いる検定の種類 (t 検定やカイ 2 乗検定等) によって、症例数設計の式は、形式上は微妙に異なることになるが、本質的には、(1)式と同じ形をとる。特に後述する、各条件の影響については、検定に基づいたすべての例数設計の問題で原則的に成り立つ。一見して、式の形が似つかぬように見えるのは、精度を少し改善するために工夫を行っているためである。それぞれの検定を行う場合の精密な例数の設計式については、成書を参照されたい³⁾。

- 4つの条件のNに対する影響

Nに対して ①～④ の条件がどのような影響を与えるかを解説する。

- ① SD^2 に比例する

例) SD が倍になれば必要なサンプルサイズは4倍になる。

データのバラツキが小さいほど、必要なNは小さくてすむ。しかもSDは2乗のオーダーで効いてくるので、その影響はたいへん大きい。例えば降圧薬の試験で血圧のSDが10mmHgのところを12mmHgと見積もっても、大きな違いはないように感じられるが、必要なサンプルサイズは1.44倍に増えてしまう。エンドポイント(評価項目)を選択する際にも、なるべくSDが小さいものを選ぶ必要がある。拡張期血圧と収縮期血圧では、一般に後者の方の変動が大きいので、SDの大きさという観点からいえば、拡張期血圧の方がエンドポイントとして望ましいといえる。(より厳密には、 Δ/SD が大きいエンドポイントが好ましい。) また測定条件は出来る限り標準化して、実験のバラツキを小さく抑える必要がある。また同一の患者について、2つの薬剤の投与時期を変えて投与するクロスオーバー実験では、患者内で2つの薬剤の効果を比較することによって、薬剤の効果を推定する際に、個体間変動ではなく個体内変動を誤差として用いることができる。このため同時に2つの群を比較する並行群試験と比べて、SDが小さくなり、症例数が少なくてすむ。

- ② Δ^2 に反比例する

例) 効果が倍になれば必要なサンプルサイズは1/4になる。

Δ についても2乗のオーダーで効いてくるので、Nに対して非常に大きな影響を与える。例えば効圧効果の Δ が25mmHgのところを、少し小さめに20mmHgと見積もれば、必要なサンプルサイズは1.56倍に増加する。臨床試験の場合、強い効果のある薬剤は必要な症例数が少なく、したがって開発費用も安価で、しかも開発期間も短くて済み、更に、医師も患者も強い効果を期待して、臨床試験に積極的に参加するだろうから、全ての面で開発が容易になる。これに対し弱い効果の薬剤は、仮に開発できたとしても、費用や

期間が余計にかかることになる。

③ 検定の有意水準 (α)を厳しくすると必要なサンプルサイズは大きくなる

α エラー, β エラーは, 検定の精度を表すパラメータで, エラーの大きさを低く抑えようとする, 大きな N が必要になる. α は通常, 両側 5%に設定されることが多いわけだが, これを小さくすると, 必要な症例数は増大する.

表 2 : α と N の関係 ($\beta=0.20$)

両側有意水準 (α)	$z_{\alpha/2}$	$2\{z_{\alpha/2}+z_{\beta}\}^2$ $\Delta=SD$ のとき必要な N	5%水準の N を 100 とした場合の N
1.0%	2.576	23.34	149
2.0%	2.326	20.05	128
5.0%	1.960	15.68	100
10.0%	1.645	12.35	79
20.0%	1.282	9.01	57

例えば有意水準を 5%から 1%に厳しくすると, 必要な症例数は, 5%水準の場合と比べ約 1.5 倍に増加する. 両側検定を片側検定に変更する場合, 有意水準は両側の倍になる. 例えば片側 5%の場合, 両側 10%点(1.645)をみればよく, 必要な症例数は両側検定の約 0.8 倍で済む. また多群で実験を行って, 多重性を考慮した検定, 例えば Dunnett 検定を行う場合には, (3)式の $z_{\alpha/2}$ を Dunnett 検定の棄却限界値で置き換えればよい. 当該の比較を $1-\beta$ の検出力で検定できる. 対照群を含めた 4 群の試験で, 自由度無限大のとき Dunnett 検定の両側 5%棄却限界値は 2.349 となり, Dunnett 検定で必要な例数は 1 群あたり $2 \times \{2.35+0.84\}^2 = 20.4$ で約 1.3 倍に増大する. 実際には 4 群で行うので, 全例数は $21 \times 4 = 84$ となる.

表 3 : β と N の関係 (両側 $\alpha=0.05$)

β エラー	z_{β}	$2\{z_{\alpha/2}+z_{\beta}\}^2$ $\Delta=SD$ のとき必要な N	β エラー20%の N を 100 とした場合の N
50%	0.000	7.68	49
30%	0.524	12.34	79
20%	0.840	15.68	100
10%	1.282	21.02	134
5%	1.645	25.99	166

④ β を小さくすると必要なサンプルサイズは大きくなる

見逃しを防ぐため, β エラーを小さく設定すると必要な N は増大する. 通常 β エラーの大きさは 20% (検出力 80%) に設定することが多いが, 少し厳しめに 10% (検出力 90%)

に設定すると、 z_β は0.84から1.28と大きくなり、必要なNは約1.34倍に増大する。逆に β エラーが30%（検出力70%）に増大することを許容すると、症例数は0.79倍に減少する。

3. 生存時間解析の例数設定

生存時間解析の例数設計を行うために、Freedman式⁴⁾とShoenfeld式⁵⁾が知られている。この二つの式では、症例の登録期間は考慮されず（全ての症例でフォローアップ期間が一定であることを想定）また途中脱落による打ち切りも考慮されてない。それぞれ(6)、(7)式にしたがって、1群あたり必要なイベント数dを算出し、仮説の下で想定される2群を平均したイベントを起こした症例の割合で除すことによって、1群あたり必要な症例数Nを(5)式にしたがって求める。

$$N = \frac{d}{\text{2群を併せたイベントの割合}} = \frac{2d}{2 - \pi_1 - \pi_2}$$

π_1 : 群1の最終時点の生存率 π_2 : 群2の最終時点の生存率 (5)

$$\text{2群を併せたイベントの割合} = \frac{1 - \pi_1 + 1 - \pi_2}{2} = \frac{2 - \pi_1 - \pi_2}{2}$$

Freedman 式

$$d = \frac{\{z_\alpha + z_\beta\}^2 (HR + 1)^2}{2(HR - 1)^2} \tag{6}$$

Shoenfeld 式

$$d = \frac{\{z_\alpha + z_\beta\}^2 \cdot 2}{(\log(HR))^2} \tag{7}$$

λ_1 : 群1のハザード λ_2 : 群2のハザード

HR = λ_1/λ_2 : 2群のハザード比

どちらの式でも、2つの群のハザード比HRを見積もる必要がある。ハザード比は、臨床家にとって直感的なイメージをつかむのが困難な指標であるが、いくつかの試算法がある。

① 時点tでの2群の生存率S(t)からの推定法

指数分布では時点tの生存関数S(t)は次のように表される。

$$S(t) = \exp(-\lambda t)$$

また確率密度関数は

$$f(t) = \lambda \cdot \exp(-\lambda t)$$

となり、ハザード関数は

$$h(t) = f(t)/S(t) = \lambda$$

となる。指数分布ではハザードが時点にかかわらず一定で λ となる。 λ は時点 t における生存割合 $S(t)$ が求まれば

$$\begin{aligned} \log S(t) &= -\lambda t \\ \lambda &= -\frac{\log S(t)}{t} \end{aligned}$$

として求めることができる。

したがってある時点 t における対照群の生存率を $S_1(t)$ 、処置群の生存率を $S_2(t)$ とすると、2つの群のハザード比は(8)式のようになる。

$$HR = \frac{\log S_2(t)}{\log S_1(t)} \quad (8)$$

② メディアン生存時間 (M) に基づく方法

メディアン生存時間（生存率が50%に低下する時点）が判明している場合は、これからハザード比を求めることができる。指数分布の場合は、メディアン生存時間の比の逆数そのままハザード比となる。またワイブル分布の場合は、メディアン生存時間の比の逆数を γ 乗することにより、ハザード比が求められる。ただし γ はワイブル分布の形状母数である

$$HR = \left(\frac{M_1}{M_2} \right): \text{指数分布} \quad (9)$$

$$HR = \left(\frac{M_1}{M_2} \right)^\gamma: \text{ワイブル分布} \quad (10)$$

③ 人年法によるハザードの推定

イベントを起こした症例についてはイベント発生までの時間、打ち切り症例については打ち切りまでの時間を足し合わせた総観察時間と、イベントの総数が判明していれば、人年法によるハザードの推定値を(11)式で求めることができる。

$$\text{ハザード} = \text{総イベント数} / \text{総観察時間} \quad (11)$$

これを人年法によるハザードとよぶ。このようにして求めたハザードは、4節で示す生存時間分布に指数分布を仮定したときのパラメータ λ の最尤推定量となっている。

4. Freedman 式 と Shoenfeld 式の数理的背景

4.1. 指数分布に基づく推定と検定

Freedman 式 と Shoenfeld 式を説明するための準備として、指数分布の母数 λ の最尤推

定について説明する。

最尤推定を行う場合、死亡した個体の尤度関数への寄与は、死亡する確率を表す確率密度関数 $f(t)$ になる。これに対し打切りを受けた個体については、まだ死亡が起きてないわけであるから、いつ死亡したかについては情報は得られていない。しかし打切りを受けた時点までは生存していた（死亡は起きるとすれば、この時点より後で起きた）ことはわかるので、打切りを受けた個体の寄与は、時点 t まで生存する確率 $S(t)$ となる。個体 i が死亡であれば 1、打切りであれば 0 をとるような変数を c_i とする。打切り症例を含めた尤度 L は、

$$L = \prod_{i=1}^n [f(t_i)^{c_i} S(t_i)^{1-c_i}] \quad (12)$$

となる。(12)式では $c_i=1$ のときは $f(t_i)$ 、 $c_i=0$ のときは $S(t_i)$ をかけることになる。

生存時間分布に指数分布を仮定した場合

$$\begin{aligned} &= \prod \{ \lambda \cdot \exp(-\lambda t_i) \}^{c_i} \cdot \{ \exp(-\lambda t_i) \}^{1-c_i} \\ &= \lambda^d \cdot \exp(-\lambda \sum t_i) \end{aligned} \quad (13)$$

となる。ここで d は総観測数から打切りを受けた個体の数を除いた総イベント数である。最尤推定では L が最大になるような λ を求める。このため通常は対数尤度を母数で微分した有効スコア関数が 0 となるような λ を求める。このようにして求めた λ の推定値は前述のハザードの年法による推定値と一致する。最尤法では求めたパラメータの精度を評価するために情報量が用いられる。情報量は対数尤度を母数で 2 階微分したものにマイナスの符号を付けたものである。この情報量の逆数が、推定値の分散になる。(パラメータが複数ある場合は、対数尤度をパラメータベクトルで 2 階微分したものにマイナスの符号を付けたものが情報行列で、この逆行列が分散・共分散行列になる。)

指数分布の場合の対数尤度と関連した統計量

対数尤度関数	:	$\log L = d \cdot \log \lambda - \lambda \sum t_i$
有効スコア関数	:	$\frac{\partial \log L}{\partial \lambda} = \frac{d}{\lambda} - \sum t_i$
最尤推定量	:	$h = \hat{\lambda} = \frac{d}{\sum t_i}$
観測情報量	:	$I = - \left. \frac{\partial^2 \log L}{\partial \lambda^2} \right _{\lambda=h} = - \frac{d}{h^2}$
最尤推定量の分散	:	$V[h] = - \frac{1}{I} = \frac{h^2}{d}$

情報量の期待値をとったものを Fisher の情報量とよぶが、ここでは期待値をとらずに最

尤推定値を代入している。この統計量を観測情報量とよぶが、サンプルサイズが大きい場合には、大数の法則により Fisher の情報量の近似とみなすことができる。

指数分布の場合は母数は一つであり、生存時間分布の違いはハザードを表す λ に縮約される。指数分布を前提に検定を行う場合、比較する2群のハザードをそれぞれ λ_1, λ_2 と表すと帰無仮説 H_0 と両側検定の対立仮説 H_1 は次のようになる。

$$H_0 : \lambda_1 = \lambda_2 \quad H_1 : \lambda_1 \neq \lambda_2$$

λ_1, λ_2 の最尤推定値を h_1 と h_2 と置くと、帰無仮説を検定するために、 h_1 と h_2 の差をその標準誤差と比較する。すなわち(14)式の Z 統計量を検定等計量として用いる。

$$Z = \frac{h_2 - h_1}{\sqrt{V[h_2 - h_1]}} = \frac{h_2 - h_1}{\sqrt{\frac{h_2^2}{d_2} + \frac{h_1^2}{d_1}}} \quad (14)$$

上式で d_1, d_2 は各群の死亡数を表す。最尤推定量の分布は漸近的に正規分布にしたがうので、Z 統計量の帰無仮説の下での分布は正規分布で近似できる。

さて、帰無仮説の下では $\lambda_1 = \lambda_2 = \lambda$ なので、2つの群でほぼ

$$h_1 \doteq h_2 \doteq \frac{(h_1 + h_2)}{2} = h \quad d_1 \doteq d_2 \doteq \frac{(d_1 + d_2)}{2} = d$$

が成り立つ。

(14)式の分母を2つの群の平均の h と d で置き換えると

$$Z = \frac{h_2 - h_1}{\sqrt{\frac{h_2^2}{d_2} + \frac{h_1^2}{d_1}}} \doteq \frac{h_2 - h_1}{\sqrt{\frac{h^2}{d} + \frac{h^2}{d}}} = \frac{h_2 - h_1}{\sqrt{2h^2}} = \frac{h_2 - h_1}{2} \sqrt{\frac{2}{d}}$$

となる。

$$Z = z_\alpha + z_\beta$$

等式は、各群のハザード h_1, h_2 を与えると、 d のみが未知数であり、この式を d について解くことにより、群当たりに必要な死亡数を求めることができる。実は等式を d について解くと Freedman 式が導かれる。

$$\begin{aligned} Z^2 &= \frac{(h_2 - h_1)^2}{\frac{(h_2 + h_1)^2}{2d}} = (z_\alpha + z_\beta)^2 \\ &= \frac{(1 - h_1/h_2)^2}{(1 + h_1/h_2)^2} = \frac{(1 - HR)^2}{(1 + HR)^2} \end{aligned}$$

$$d = \frac{\{z_\alpha + z_\beta\}^2 (HR + 1)^2}{2(HR - 1)^2}$$

さて、2群間でハザードの差が0であるとして検定統計量を導いたが、次のように帰無仮説を設定することもできる。

$$H_0 : HR = \lambda_2 / \lambda_1 = 1$$

すなわち帰無仮説として2群のハザード比が1とおいてもよい。この式の対数をとると

$$H_0 : \log HR = \log \lambda_2 - \log \lambda_1 = 0$$

となる。対数変換したハザードの差が0であることを示している。(14)式では、変換前のハザードが0であるかどうかを検定したが、ハザードは正の値しかとらず、このためハザードの推定値は歪んだ分布となるので、サンプルサイズが小さいときは正規分布による近似はあまりよくない。これに対して、対数変換した場合、ハザードが1未満のときは負、1を越えるときは正の値をとり、正規分布の定義域ともあうようになる。このため正規近似の精度が改善される。ハザードの推定値を対数変換した $\log h$ の分散はデルタ法により近似的に(15)式ようになる。

$$\begin{aligned} V[\log h] &\doteq \left[\frac{\partial \log h}{\partial h} \right]^2 \cdot V[h] \\ &= \frac{1}{h^2} \cdot \frac{h^2}{d} = \frac{1}{d} \end{aligned} \quad (15)$$

したがって、対数ハザードの差が0であるかを検定する場合のZ統計量は

$$Z = \frac{\log(h_2) - \log(h_1)}{\sqrt{V[\log(h_2) - \log(h_1)]}} \doteq \frac{\log(h_2) - \log(h_1)}{\sqrt{\frac{1}{d_2} + \frac{1}{d_1}}} \quad (16)$$

となる。やはり 帰無仮説の下では $d_1 \doteq d_2 \doteq (d_1 + d_2) / 2 = d$ が成り立つので、 d_1 , d_2 を d で置き換えると

$$Z = \frac{\log(h_2) - \log(h_1)}{\sqrt{\frac{1}{d} + \frac{1}{d}}} \doteq \frac{\log(h_2) - \log(h_1)}{\sqrt{\frac{2}{d}}} \quad (17)$$

となる。等式

$$Z = \frac{\log(h_2) - \log(h_1)}{\sqrt{\frac{2}{d}}} = z_\alpha + z_\beta$$

は、各群のハザード h_1, h_2 を与えると、 d のみが未知数であり、この式を d について解くことにより、一つの群当たりに必要な死亡数を求めることができる。実はこのようにして求めた 1 群当たりの死亡数は Shoenfeld 式に完全に一致する。

以上示したように、Freedman 式がハザードの差が 0 であるかの検定に対応するのに対し、Shoenfeld 式は、対数変換後のハザードの差が 0 (ハザード比 1) であるかの検定に対応する。二つの式を比較してみると

Freedman 式

$$d = \frac{\{z_\alpha + z_\beta\}^2 (HR + 1)^2}{2(HR - 1)^2}$$

Shoenfeld 式

$$d = \frac{\{z_\alpha + z_\beta\}^2 2}{(\log(HR))^2}$$

Freedman 式と Shoenfeld 式は、 $\{z_\alpha + z_\beta\}^2$ の項を共通して持つ。またどちらもハザード比 HR に対する非線形関数となっている。そこで両式の違いを明らかにするため、非線形関数を多項式で近似するテーラー展開を適用してみる。関数 $f(x)$ を a の周りでテーラー展開して 2 次式で近似すると

$$f(x) \doteq f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2} \quad (18)$$

となる。 $\log(HR)$ を 1 の周りでテーラー展開して 2 次式で近似すると

$$\log(HR) \doteq \log(1) + (HR-1) - \frac{(HR-1)^2}{2} = (HR-1) - \frac{(HR-1)^2}{2} \quad (19)$$

となる。これに対し、 $f(HR) = (HR \cdot 1) / (HR + 1)$ を $HR=1$ の周りでテーラー展開すると

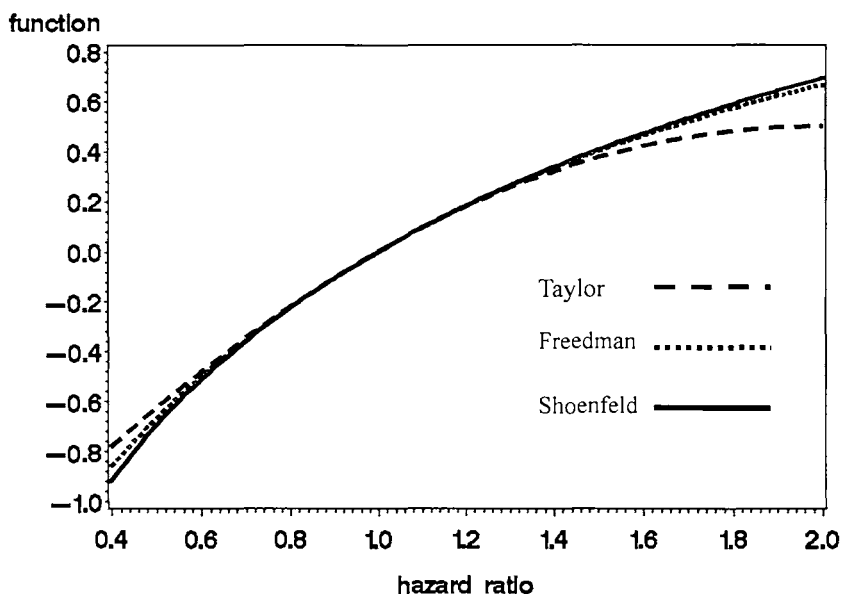
となる。したがって

$$\frac{(HR-1)}{(HR+1)} \doteq \frac{(HR-1)}{2} - \frac{(HR-1)^2}{4}$$

$$2 \cdot \frac{(HR-1)}{(HR+1)} \doteq \log(HR)$$

が近似的に成り立ち、帰無仮説 ($HR=1$) の近傍では、両式はほぼ等しくなることがわかる。実際には $\log(HR)$ の方が $2 \{ (HR \cdot 1) / (HR + 1) \}$ より若干大きめの値をとる。これは症例数設計式の分母の方なので、Freedman 式と比べて Shoenfeld の方が必要な例数は少なくすむ。図 1 に、Shoenfeld: $\log(HR)$, Freedman: $2 (HR \cdot 1) / (HR + 1)$, テーラー展開: $(HR - 1) - (HR - 1)^2 / 2$ の 3 つの関数を、 HR を 0.4 から 2.0 まで変化させて比較した結果を示

した。ハザード比が1に近いときは3つの関数ともほぼ等しい値をとるが、1から離れるにつれ絶対値は Shoenfeld > Freedman > テーラー展開 の順になる。



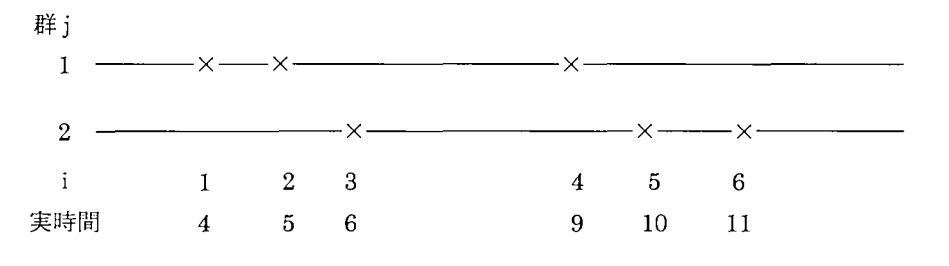
Shoenfeld: $\log(h)$ Freedman: $2 \times (h-1)/(h+1)$ Taylor: $(h-1) - (h-1)^2/2$

図 1: 3 つの関数形の比較

以上、Freedman 式と Shoenfeld 式が指数分布の母数の検定に基づいた方法であることを示した。しかし、これらの方法は一般に特定の分布を仮定しないノンパラメトリックなログランク検定ベースの方法と呼ばれている。これは何故だろうか。実はノンパラメトリック検定でありながらログランク検定はハザード比の検定と密接な関連を有する。

4.2. ログランク検定の結果に基づくハザード比の推定

図 2 : 仮想例



ログランク検定とハザード比の関連を、図2の仮想的な2群試験の例を用いて説明する。このデータについてLIFETESTプロシジャを用いて、ログランク検定を実施するためのプログラムは次のようになる。

表4 : ログランク検定のためのLIFETESTのプログラム

```
DATA WORK;
INPUT GROUP TIME CENSOR @@;
CARDS:
1 4 1 1 9 1 1 5 1
2 10 1 2 6 1 2 11 1
PROC LIFETEST DATA=WORK NOTABLE;
TIME TIME*CENSOR(0);STRATA GROUP;RUN;
```

プログラムを実行すると、次のような出力が得られ、ログランク検定のカイ2乗統計量は2.5567となる。もちろん、1群の症例数が3と少ないので有意にはならない。

表5 : LIFETESTの出力結果

Rank Statistics			
GROUP	Log-Rank	Wilcoxon	
1	1.5167	7.0000	
2	-1.5167	-7.0000	
Covariance Matrix for the Log-Rank Statistics			
GROUP	1	2	
1	0.899722	-.899722	
2	-.899722	0.899722	
Test of Equality over Strata			
Test	Chi-Square	DF	Pr >
Log-Rank	2.5567	1	0.1098
Wilcoxon	2.4500	1	0.1175
-2Log(LR)	0.7067	1	0.4005

$$\chi^2 = \frac{\{\sum(O_{i2} - E_{i2})\}^2}{V[\sum(O_{i2} - E_{i2})]} = \frac{U^2}{I} = \frac{(-1.5167)^2}{0.899722} = 2.5567 \quad (20)$$

このカイ2乗統計量は、(20)式に示したように群2について各時点ごとの観測死亡数 O_{i2} と期待死亡数 E_{i2} の差を足し合わせてから(Rank Statistics)2乗したもの(-1.5167²)を、対応する分散(0.899722)で除したものである(群1について同様の操作を行っても結果は同じになる)。LIFETESTでは、2つの群の観測死亡数と期待死亡数の差の分散共分散行列がCovariance Matrix for the Log-Rank Statisticsとラベルされて出力されているので、その2行2列目の要素が対応する分散になる。分散については、一般に超幾何分布に基づいて計

算されるが、死亡に同順位がない場合は、2項分布に基づく分散と等しくなり、分散は(21)式で与えられる。

$$V[\sum(O_{i2} - E_{i2})] = \sum \frac{n_{i1}n_{i2}}{n_i^2} = \sum p_{i1}p_{i2} \quad (21)$$

ここで、 n_i は時点*i*において2群を併せたリスク集合(時点*i*の直前で死亡も打ち切りも起こしてない個体の数)の大きさ、 n_{ij} は、時点*i*の群*j*のリスク集合の大きさを表し、 $p_{ij} = n_{ij} / n_i$ である。

$$U = \sum(O_{i2} - E_{i2})$$

$$I = V[\sum(O_{i2} - E_{i2})]$$

とおくと、Peto法では次のようにハザード比を推定する。

$$HR = \exp(U/I) \quad (22)$$

見方を変えれば、 $b = U/I$ は対数ハザード比(Cox回帰の係数*b*)の推定値となっている。また*b*の分散は、近似的に

$$V[b] = 1/I$$

であたえられる。先の例では

$$b = U/I = -1.5167/0.899722 = -1.6857$$

$$HR = \exp(U/I) = \exp(-1.6857) = 0.1853$$

$$V[b] = 1/I = 1/0.899722 = 1.05426$$

となる。

通常、ハザード比を推定するためにはCox回帰を行うが、Peto法で求めたハザード比は、Cox回帰でニュートン・ラプソン法による反復計算を1回しか行わない場合の推定値と一致する。このことを次のプログラムにより確認する。

表6：反復計算を一回に制限するPHREGのプログラム

```
DATA WORK;
INPUT GROUP TIME CENSOR @@;
CARDS;
1 4 1 1 9 1 1 5 1
2 10 1 2 6 1 2 11 1
PROC PHREG DATA=WORK;
MODEL TIME*CENSOR(0)=GROUP/ITPRINT MAXITER=1;
```

Cox 回帰分析を行うための、PHREG プロシジャでは、MODEL 文で、MAXITER=反復回数オプションを指定することにより、反復計算の数を制限でき、表 6 の指定では反復計算は一回しか行わない。PHREG プロシジャの出力は表 7 のようになる。

表 7 : 反復計算を一回だけ行った場合の PHREG の出力

Testing Global Null Hypothesis: BETA=0						
Test		Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio		2.4720	1	0.1159		
Score		2.5567	1	0.1098		
Wald		2.0716	1	0.1501		
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
GROUP	1	-1.68571	1.17120	2.0716	0.1501	0.185

確かに Peto 法で求めた対数ハザード比 -1.68571 と一致していることが確認できる。

ちなみに MAXITER=オプションを指定しないと二回で収束し、そのときの対数ハザード比は -1.68785 となる。この例が示すように、一般に Peto 法は Cox 回帰によるハザード比の推定値をよく近似できるが、推定値の絶対値が低めに評価されてしまう。

対数ハザード比とその分散から、対数ハザード比が 0 (ハザード比が 1) かの検定は(23)式のように構成できる。

$$Z = \frac{b}{\sqrt{V[b]}} = (U/I) \cdot I^{0.5} = U/I^{0.5} = \frac{\sum(O_{i2} - E_{i2})}{\sqrt{V[\sum(O_{i2} - E_{i2})]}} \quad (23)$$

この Z 統計量を 2 乗したものがログランク検定のカイ 2 乗統計量に等しくなる。以上示したように、ログランク検定は対数ハザード比が 0、すなわち 2 群間で対数ハザードの差が 0 かを検定していると考えられる。したがってログランク検定で例数設計を行う場合は、 $Z = z_{\alpha} + z_{\beta}$ になるようにすればよい。

帰無仮説の近傍では、分散 I は

$$I = V[\sum(O_{i2} - E_{i2})] = \sum p_{i1} p_{i2} \doteq \sum 0.5 \cdot 0.5 = 2 \frac{d}{4} = \frac{d}{2} \quad (24)$$

近似できる。すなわち、ハザード比が 1 に近づけば、どちらの群でも同様に確率 0.5 で死亡が発生するので、リスク集合の大きさはどの時点でも等しく、したがって $p_{i1} p_{i2} \doteq 0.5 \cdot 0.5 = 1/4$ と近似できる。ここで d は 1 群当たりの平均死亡数である。

したがって、

$$Z = b \cdot I^{0.5} = b \cdot \left(\frac{d}{2}\right)^{0.5} = z_{\alpha} + z_{\beta}$$

上の式を d について解くと、

$$d = \frac{\{z_\alpha + z_\beta\}^2 \cdot 2}{b^2}$$

b=logHR であることに注意すると、Shoenfeld 式に一致することがわかる。すなわちログランク検定に基づいた症例数設計は Shoenfeld 式、あるいは HR が 1 に近いときは、Shoenfeld 式とほぼ等しくなる Freedman 式を用いて近似することができる。またログランク検定ベースのハザード比の推定は Cox 回帰の係数(対数ハザード比)が 0 かを検定する場合も精度よく近似できる。以上のように、少し奇妙ではあるが、パラメトリックに指数分布を仮定した場合のハザード比の検定、セミパラメトリックな Cox 回帰、ノンパラメトリックなログランク検定の 3 つのアプローチのいずれを採用しても、例数設計については、ほぼ同様になる。

5. SAS での症例数設計のプログラム

5.1. 症例数設計

有意水準両側 $\alpha=0.05$, $\beta=0.20$ で、

対照群 手術単独群の 5 年生存率 : 0.65

薬剤群 補助化学療法群の 5 年生存率 : 0.80

として、生存率からハザード比を求めて、Freedman 式と Shoenfeld 式に基づいて症例数設計を行うプログラム例を次に示す。

表 8 : Freedman 式と Shoenfeld 式に基づく症例数設計プログラム

```
data samplesize;
  alpha=0.05;beta=0.20;
  t=5;pc=0.65;pd=0.80;
  h2=-log(pd)/t;h1=-log(pc)/t;hr=h2/h1;
  za=probit(1-alpha/2);zb=probit(1-beta);
  ef=(za+zb)**2*(hr+1)**2/(2*(hr-1)**2);
  nf=2*ef/(2-pd-pc);
  es=2*(za+zb)**2/((log(hr))**2);
  ns=2*es/(2-pd-pc);
proc print;run;
```

プログラム中で変数EF, NFがFreedman式で必要な1群あたりのイベント数と症例数, ES, NSがShoenfeld式で必要な1群あたりのイベント数と症例数である。結果は次のようになる。

表 9 : Freedman 式と Schoenfeld 式に基づく症例数設計の結果

OBS	alpha	beta	t	pc	pd	h2	h1	hr	za	zb	ef	nf	es	ns
1	0.05	0.2	5	0.65	0.8	0.045	0.086	0.52	1.96	0.84	38.92	141.54	36.28	131.93

5年生存率から、対照群のハザード(h1)は0.086となる。ハザードは1/年という単位を持つことに注意する必要がある。生存時間の単位が年ではなく、月で測られる場合は、5年の代わりに60ヶ月で割ることになり、このときの単位は1/月となり、値は1/12に低下する。これに対し薬剤群ではハザード(h2)は0.045となり、ハザード比(HR)は0.52となる。Freedman式では、1群あたり必要な症例数は142例、Schoenfeld式では132例になる。5年時点の生存率65%と80%を生存・死亡の2値データに基づいて検定するために必要なNを、参考のために示すと、1群151例になる。生存時間解析では、単に生存の有無だけではなく、生存時間の長さを評価し、情報量が増えるため、必要な症例数はより少なくなる事が確認できる。

5.2. 検出力の評価

前節までは、 α , β エラーの大きさ、対数ハザード比(logHR)を与えて1群あたり必要なイベント数dを求めたが、logHR, α , dを決めれば、 β エラーの大きさを求めることができる。1から β エラーの大きさを引くと、検定の検出力になる。

Freedman式とSchoenfeld式を z_β について解くと、

Freedman式

$$z_\beta = \sqrt{2d} \cdot \left| \frac{HR-1}{HR+1} \right| - z_\alpha \quad (25)$$

Schoenfeld式

$$z_\beta = \sqrt{\frac{d}{2}} \cdot |\log(HR)| - z_\alpha \quad (26)$$

となる。この z_β を正規分布の分布関数と比較することにより、検出力を計算できる。SASでは、PROBNORM関数を利用して検出力が計算できる。

有意水準両側 $\alpha=0.05$, $N=150$

対照群 手術単独群の5年生存率 : 0.65

薬剤群 補助化学療法群の5年生存率 : 0.80

として、生存率からイベント数を求めて、Freedman式とSchoenfeld式に基づいて検出力を算出するプログラム例を表10に示す。

表 10 : Freedman 式と Shoenfeld 式に基づく検出力の計算プログラム

```

data power;
  alpha=0.05;beta=0.20;
  t=5;pc=0.65;pd=0.80;n1=150;n2=150;
  h2=-log(pd)/t;h1=-log(pc)/t;hr=h2/h1;
  za=probit(1-alpha/2);
  e=(n1*(1-pc)+n2*(1-pd))/2;
  zbf=(e*2)**.5*abs((hr-1)/(hr+1))-za;
  zbs=(e/2)**.5*abs(log(hr))-za;
  pf=probnorm(zbf);
  ps=probnorm(zbs);
proc print;run;

```

プログラムでは 1 群あたりの平均イベント数(e)を予測し、それから z_{β} を求め、検出力を計算している。実行結果は次のようになる。

表 11 : Freedman 式と Shoenfeld 式に基づく検出力の計算結果

OBS	alpha	beta	t	pc	pd	n1	n2	h2	h1	hr	za	e	zbf	zbs	pf	ps
1	0.05	0.2	5	0.65	0.8	150	150	0.04	0.09	0.52	1.96	41.25	0.92	1.03	0.82	0.85

各群 150 例のとき 1 群当たりの平均イベント数は 41.25 となる。 Z_{β} は Freedman 式では、0.92、Shoenfeld 式では 1.03 となる。標準正規分布で、 z_{β} を越えない確率が検出力で Freedman 式(82%)、Shoenfeld 式では(85%)となる。例示はしないが、 α 、 β 、 d を決めて、 $\log HR$ について式を解けば、イベント数(症例数)と検出力を固定した上で、検出可能な効果の大きさ(対数ハザード比)を求めることもできる。

Freedman 式と Shoenfeld 式では、途中脱落や、患者登録期間が考慮されていない。これらの要素を考慮する場合は、群当たりの平均イベント数がどのようなかを評価をすればよい。これらの症例数設計では比例ハザード性を前提とし、ハザード比を固定できれば有意水準 α が一定の下では、イベント数だけが問題になる。中間解析を行う段階の検出力を評価することは重要であるが、この場合、中間段階での予測イベント数を求め、(25)式または(26)式に代入するだけで検出力を評価できる。

6. より複雑な問題における症例数設計

6.1. 登録期間を考慮した検出力の評価

通常の臨床試験では、患者は逐次的に研究に登録され、一度に全症例が試験に組み入れられるわけではない。例えば、登録期間を2年で、その後、5年間のフォローアップ期間を設定した試験では、終わりの方に登録された症例のフォローアップ期間は5年間だが、初期の登録例は7年近く追跡され、全員のフォローアップ期間を5年とした場合と比べて、期待イベント数は増え、若干、検出力も増大する。このとき登録期間を考慮した検出力の計算方法を次に示す。研究登録期間をR年、フォローアップ期間をT年とする。患者が一定の速度で(範囲0~Rを確率密度1/Rで一様分布にしたがう)登録されるものとする。このとき1年当たりの2群を合わせた患者登録数をnperyearとすると、期待される総イベント数は

$$R \times nperyear \times \text{イベントの割合}$$

となる。

登録時点tの個体のR+T年経過後のイベントの割合をP(event|t)とすると、登録時点tは0~Rを1/Rの確率で一様に分布するので、

$$\begin{aligned} \text{イベントの割合} &= \int_0^R P(\text{event} | t) \cdot \frac{1}{R} \cdot dt \\ &= 1 - \int_0^R S(\text{event} | t) \cdot \frac{1}{R} \cdot dt \\ &= 1 - \int_0^R S(R+T-t) \cdot \frac{1}{R} \cdot dt \end{aligned} \quad (27)$$

となる。ここでS(t)は2群を合わせた生存関数であり、時点tで登録された患者が試験期間R+T年中、生き残る確率がS(R+T-t)になる。生存時間分布に指数分布を仮定すると、

$$\text{イベントの割合} = 1 - \int_0^R -\exp\{\lambda \cdot (R+T-t)\} \cdot \frac{1}{R} \cdot dt \quad (28)$$

となる。ただしλはハザードを表す指数分布の母数である。この式をtが0~Rの範囲で積分すると、

$$\text{イベントの割合} = 1 - \frac{\exp\{\lambda \cdot (R+T)\} \cdot (\exp\{\lambda R\} - 1)}{R\lambda}$$

となる。イベントの割合から、総イベント数が求まれば、(25)式または(26)式を適用することにより、簡単に検出力を計算することができる。先の例で、年当たりの登録例数nperyearを150例、登録期間R=2年として、Freedman式ベースの検出力を求めるプログラムを表12に示す。

表 12 : 登録期間を考慮した検出力の計算プログラム

```

data powerwithrt;
  r=2;nperyear=150;alpha=0.05;
  pc=0.65;pd=0.80;t=5;
  za=probit(1-alpha/2);
  lambda=-log((pd+pc)/2)/t;
  h2=-log(pd)/t;h1=-log(pc)/t;hr=h2/h1;
  n=nperyear*2;
  pevent=1-(1/r)*exp(-lambda*(r+t))*(exp(lambda*r)-1)/lambda;
  e=n*pevent/2;
  zbf=(e*2)**.5*abs((hr-1)/(hr+1))-za;
  pf=probnorm(zbf);
proc print;run;

```

結果は次のようになる。

表 13 : 登録期間を考慮した検出力の計算結果

OBS	r	nperyear	alpha	pc	pd	t	za	lambda	h2	h1	hr	n	pevent	e	zbf	pf
1	2	150	0.05	0.65	0.8	5	1.96	0.06	0.04	0.09	0.52	300	0.32	47.95	1.15	0.87

総症例数は 300 例と表 11 と同じであるが、登録期間を考慮することによりイベントの割合(pevent)は、 $(0.35+0.20)/2=0.275$ から 0.32 に増大し、1 群当たりの平均イベント数(e)は 41.25 から 47.95 に増大する。このため検出力(pf)は 0.82 から 0.87 に増大する。また各時点で、どの程度の検出力があるかを評価することも可能である。図 3 に登録期間 2 年終了後、フォローアップ期間 0~6 年目までの各時点の検出力と 2 つの群を併せた期待イベント数を示した。登録期間終了直後でも、最大 2 年間フォローアップされている患者がいるため 30%弱の検出力はある。フォローアップ期間中期待イベント数はほぼ直線的に増加し、イベント数の増大に伴い検出力は増大する。フォローアップ 4 年で検出力は 80%を越え、登録期間 2 年と合わせて、全試験期間が 6 年あれば、検出力を 80%以上にすることができる。このような検討はフォローアップ期間、中間解析の時点を決めたり、患者登録の速度を考慮して、参加施設数を検討するために有用である。

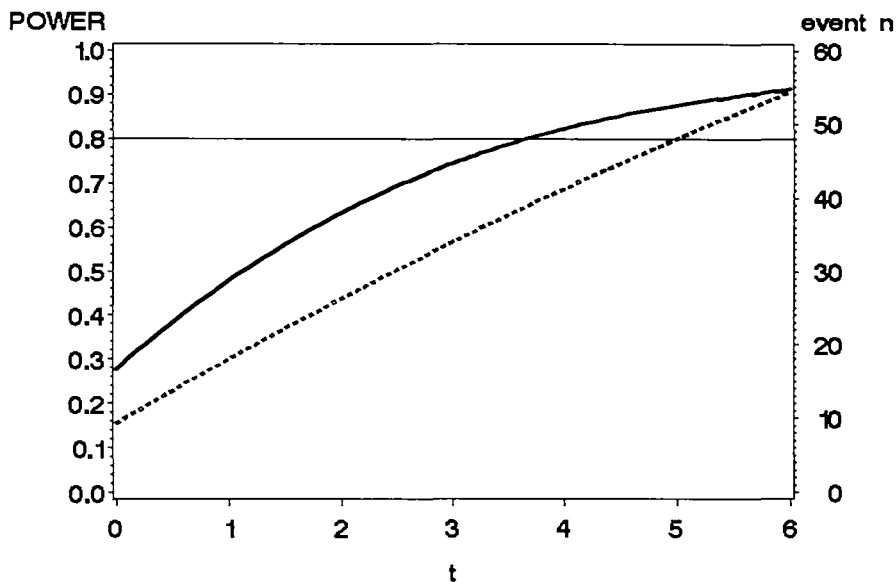


図3：フォローアップ期間による検出力の変化
 実線：検出力 点線：イベント数

6.2. 非劣性試験の例数設計

比例ハザード性を前提として、対照群のハザードを λ_1 、実験群のハザードを λ_2 とすると、ハザード比の優越性について片側検定で検討するためには、帰無仮説 H_0 と対立仮説 H_1 は次のように設定される。

$$H_0 : \lambda_1 \leq \lambda_2 \quad H_1 : \lambda_1 > \lambda_2$$

これに対してハザード比の非劣性について検討する場合は、帰無仮説と対立仮説を次のように設定する。

$$H_0 : \lambda_1 + \Delta \leq \lambda_2 \quad H_1 : \lambda_1 + \Delta > \lambda_2$$

ここで、 Δ はいわゆる非劣性マージンで、臨床的に許容できる差である。すなわち Δ のハンデをつければ、実験群のハザードは対照群のハザードより有意に低くなる。裏を返せば、実験群のハザードは対照群と比べて、 Δ 以上高い可能性は棄却できる。

このときFreedman式に基づいて、非劣性仮説を検討する場合、1群当たりに要求される死亡数 d は次の式で与えられる。

$$d = \frac{[z_\alpha + z_\beta]^2 (HR' + 1)^2}{2(HR' - 1)^2} \quad (29)$$

ただし $HR' = (\lambda_1 + \Delta) / \lambda_2 = HR + \Delta / \lambda_2$ である。

前述の例について、 Δ / λ_2 を0～-0.3まで0.05刻みで変化させたときのFreedman式に基づく例数設計のプログラムを次に示す。

表 14 : 非劣性試験の症例数設計のプログラム

```

data inferiority;
  alpha=0.05;beta=0.20;
  t=5;pc=0.65;pd=0.80;
  h2=-log(pd)/t;h1=-log(pc)/t;hr=h2/h1;
  za=probit(1-alpha/2);zb=probit(1-beta);
  do delta=0 to -0.3 by -0.05;
    ef=(za+zb)**2*(hr+delta+1)**2/(2*(hr+delta-1)**2);
    nf=2*ef/(2-pd-pc);
    output;
  end;
proc print;run;

```

結果は次のようになる。

表 15 : 非劣性試験の症例数設計の結果

OBS	alpha	beta	t	pc	pd	h2	h1	hr	za	zb	delta	ef	nf
1	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	0.00	38.92	141.54
2	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.05	29.88	108.66
3	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.10	23.30	84.71
4	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.15	18.39	66.86
5	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.20	14.66	53.30
6	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.25	11.78	42.82
7	0.05	0.2	5	0.65	0.8	0.04	0.09	0.52	1.96	0.84	-0.30	9.52	34.62

非劣性マージンを設定することにより、優越性試験 ($\Delta=0$) と比べて必要な例数は、大幅に減少することが確認できる。

6.3. シミュレーションによる例数設計

表 16 に示したような、より複雑な問題については、例数設計の公式は教科書レベルのテキストには少なくとも記載されていない。このような場合、乱数を利用したシミュレーションによって検出力を検討するのが簡便である。

表 16 : シミュレーションによる例数設計が必要な状況

- 1) プライマリーな解析として、一般化ウイルコクソン検定を用いる。
- 2) 生存時間分布としてワイブル分布などの指数分布以外の分布が想定される。
- 3) 2群で例数をアンバランスにしたい(プラセボの割合を減らしたい)。
- 4) 3群以上でデザインしたい。
- 5) 比例ハザード的ではない効果を想定する。

シミュレーションの手順は次のようになる。(計算公式等が知られていれば、それを用いて n を試算する。)

- ① 想定する対立仮説の下で n を決めて乱数を発生させる。
- ② ①のデータについて検定を行い、 α 水準で有意かどうかを評価する。
- ③ ①, ②の過程を数百～数千回くり返し、有意になる割合を調べる。これが検出力の推定値になる。
- ④ 検出力が不適切であれば、 n を変えて①～③の過程をやり直す。

①のステップには、SAS の乱数関数を用いることができる。次によく用いられる SAS の乱数関数を紹介する。

- 正規分布 rannor 関数
 - 文法 rannor(seed)
 - 例) $x = 100 + 20 * \text{rannor}(4989)$;
平均が 100 で標準偏差が 20 の正規分布
- 2項分布 ranbin 関数
 - 文法 ranbin(seed, n, p)
 - 例) $x = \text{ranbin}(5963, 50, 0.3)$;
 $n = 50, p = 0.30$ の 2項分布
- 指数分布 ranexp 関数
 - 文法 ranexp(seed)
 - 例) $x = \text{ranexp}(4649) / 0.2$;
ハザードが 0.2(期待値が 1/0.2)の指数分布

生存時間解析の例数設計では、指数分布にしたがう乱数を発生させることが多い。指数分布を発生させるためには RANEXP 関数が用意されている。期待値が 1 の指数分布にしたがう乱数が発生するので、これを定数倍することにより、任意の期待値の指数乱数を発生させることができる。またワイブル分布にしたがう乱数を発生させるためには、RAND 関数で WEIBULL オプションを指定すればよい。

一般化ウイルコクソン検定の検出力を検討するためのプログラムを次に示す。前述の条件で Freedman 式にしたがい例数設計を行うと 1 群 142 例となったので、2 群合わせて 284

の乱数を発生させる。ただし、生存時間が5年を越えたものは、この時点で打ち切り扱いとする。2つの群で5年生存率がそれぞれ65%と80%になるように指数分布のハザードを調整する。

合計1000組のデータを発生させ、それぞれについてLIFETESTプロシジャを用いて、ログランク検定、一般化ウイルクソン検定、尤度比検定を行い、その結果をODS(Output Delivery System)の機能を利用してSASデータセット化し、FREQプロシジャで集計する。

表 17 : シミュレーションによる検出力の評価

```
data data;
r1=-log(0.650)/5;r2=-log(0.800)/5;
do n=142;do i=1 to 1000;
dose=0;do j=1 to n;
      t=ranexp(4989)/r1; censor=2;
      if t gt 5 then do t=5; censor=0;end;output;end;
dose=1;do j=1 to n;
      t=ranexp(4989)/r2; censor=2;
      if t gt 5 then do t=5; censor=0;end;output;end;end;
end;
ods listing close;
proc lifetest data=data ;
time T*censor(0);strata dose;
by n i;
ods output HomTests=out;run;
ods listing;
data out;set out;
if 0<Probchisq<0.05 then sign=1;else sign=0;
proc freq;tables sign*test/nopercent norow;
```

結果は次のようになる。

表18 : シミュレーションによって評価した検出力

表 : sign * test				
度数 列のパーセント	-2Log(LR)	Log-Rank	Wilcoxon	合計
0	186 18.60	190 19.00	194 19.40	570
1	814 81.40	810 81.00	806 80.60	2430
合計	1000	1000	1000	3000

ログランク検定の検出力 81.4%に対し、一般化ウイルコクソン検定では 80.6%となり、この例では一般化ウイルコクソン検定でも十分な検出力があることがわかる。

7. おわりに

SAS では現在のところ生存時間解析用のプロシジャは用意されておらず、本稿で示したようにデータステップでプログラムを作成するしかないが、NQUERY 等の標準的な例数設計のソフトウェアでは様々な状況での例数設計が可能である⁹⁾。例えば比例ハザード性が成り立たず、区分ごとで異なったハザード比を持った区分指数モデルを用いて、症例数設計を行うことができる。

また世界最大の癌の臨床試験のグループ SWOG(South West Oncology Group)のホームページでは、生存時間解析を含めた様々な例数設計のプログラムが無償で提供されており⁷⁾、これを用いれば、生存時間解析で非劣性仮説を検証する場合の例数設計も可能である。

ICH 以後、治験で患者一人当たりに必要なコストは激増した。無駄のない試験を行うためには、試験の計画段階で適切な症例数を統計的に見積もることが不可欠である。

■ 参考文献

- 1)Castelloe, J.M. (2000), "Sample Size Computations and Power Analysis with the SAS ® System," *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, Paper 265-25, Cary, NC: SAS Institute Inc.
- 2)Castelloe, J.M. and O'Brien, R.G. (2001), "Power and Sample Size Determination for Linear Models," *Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference*, Paper 240-26, Cary, NC: SAS Institute Inc.

- 3) 宮原英夫・丹後俊郎(1995) 医学統計学ハンドブック. 朝倉出版
- 4) Freedman, L. S. (1982), "Tables of the number of patients required in clinical trials using the log-rank test," *Statistics in Medicine*, 1, 121–129.
- 5) Schoenfeld, D. (1981), "The asymptotic properties of nonparametric tests for comparing survival distributions," *Biometrika*, 68, 316–319.
- 6) <http://www.statsol.ie/nquery/features.htm>
- 7) <http://www.swogstat.org/stat/public/default.htm>
- 8) 大橋靖雄・浜田知久馬(1995) 生存時間解析 東大出版会
- 9) Collett, D. (1994) *Modelling survival data in Medical Research*. Chapman and Hall.

口頭論文発表
統計解析

日本SASユーザー会 (SUGI-J)

区間打ち切り生存時間データのセミパラメトリックな解析法 のSASプログラムの紹介

～ギブス・サンプラーを利用した周辺尤度アプローチ～

○西山智^{1,2} 吉村健一³

¹ アベンティス ファーマ株式会社 ² 東京理科大学 ³ 東京大学

The semi-parametric model for the analysis of time-to-event data with interval censoring
using SAS

Hiroshi Nishiyama^{1,2} Kenichi Yoshimura³

¹ Aventis Pharma Ltd. ² Tokyo University of Science ³ University of Tokyo

要 旨

区間打ち切り生存時間データに対して、区間打ち切りを無視した解析が行われることが多い。本稿では、ギブス・サンプラーを利用した比例ハザードモデルのパラメータ推定とそれを行うプログラムを紹介する。

キーワード： 区間打ち切り生存時間データ, 比例ハザードモデル, ギブス・サンプラー

1 はじめに

区間打ち切り生存時間データは、研究者が対象とするイベントの正確な発生時間が観察できず、ある期間中に発生したことのみのみが情報として得られる状況で生じる。たとえば、臨床検査の結果のみに基づいてイベントを定義する様な場合であれば、ある期間ごとに実施される臨床検査で陰性判定された最後の検査日から陽性判定された最初の検査日までの期間がイベントの発生情報として観察される。イベントがこの観察された区間のどこかで発生している事は確かであるが、観察者には正確な発生時間が分からない。図1は、対象 a, b, d についてそれぞれ $(t_1, t_6]$, $(t_2, t_7 = \infty)$, $(t_3, t_4]$ という区間打ち切り生存時間、対象 c については t_5 という生存時間が観察されたとする仮想データである。このような区間打ち切りを受けることにより、観察対象ごとに区間の長さが異なる可能性があり、さらに特殊な状況ではタイデータが多く存在しうるため、通常の統計解析手法は一般に用いる事ができない。

ところが、現実の臨床試験においてこのような状況はしばしば生じ、多くの場合には補完 (imputation) を行うなどした上でこの区間打ち切りを無視したナイーブな解析法が用いられている。たとえば、よ

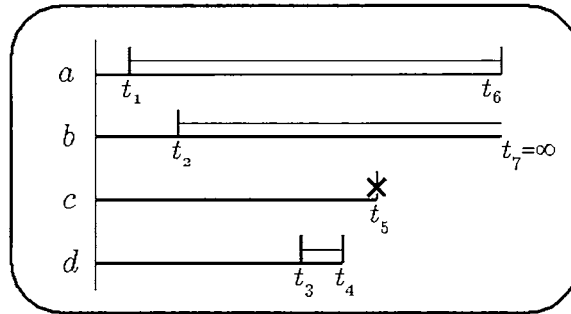


図 1: 4 例の仮想データ

く行われる陽性判定された最初の検査日をイベント発生時間とするのも補完の一種である。骨粗鬆症例の新規脊椎骨折の観察研究では、骨折診断方法に Clinical Fracture および Radiological Fracture の両者を共に用いる事が一般的であるが、骨折というイベントを、前者は一般的に症状を訴えて受診した患者に対する X 線撮影による一方で、後者は定期来院時の X 線撮影により判定する。前者であればほぼ正確なイベント発生時間が観察される一方、後者であれば区間打ち切りを受けた上で観察される。全ての症例が前者の状況で観察されたのであれば SAS システム (PHREG プロシジャなど) で提供されている通常の解析法の適用も妥当であるが、後者の様な状況で観察される症例が混在してしまっている場合において解析を妥当に行うためのアルゴリズムは現在の SAS システムでは提供されていない。

本稿では、区間打ち切りデータ解析のための Satten (1996) の周辺尤度アプローチに基づくプログラムを作成したので紹介する。この方法の基本的な考え方は次のとおりである。生存時間解析においてよく知られている観察データの順位統計量を用いた Cox (1972, 1975) の部分尤度に基づくパラメータ推定では、発生したイベントの順序が特定される必要がある。しかし、区間打ち切りデータでは順序が一意に特定されないため、観察データに矛盾しない全てのイベント発生順序の組を考えた組ごとの部分尤度を足し合わせた周辺尤度を構成するという Cox の部分尤度を拡張した形を考えることになる。しかし、データ数が極端に少ない場合を除いて、一般に全ての可能な発生順序の組を考えて尤度を構成する事は容易ではない。そこで Satten (1996) はマルコフ連鎖に基づくモンテカルロ法 (Markov chain Monte Carlo 法: MCMC 法) の一つであるギブス・サンプラーを利用する事で、イベント発生順序の組を比例ハザード性を仮定した下での分布から発生させることにより、推定したいパラメータの事後分布を得るアルゴリズムを提案した。なお、ここで用いるギブス・サンプラーを始めとする MCMC 法はベイズ統計学で頻繁に利用される統計的推測法であるが、現在の SAS システムではベイズ統計学的手法は提供されていない。

本稿で紹介するプログラムは、上述の骨折事例のような区間打ち切りを含む任意の打ち切りを含む一般的な生存時間データに適用可能である。また、特に治験データへの適用を考えたとき、データ解析に利用する統計手法の妥当性およびモデルの仮定に対する結果の頑健性が強く求められる。これに関して、採用した方法は区間打ち切りでない正確なイベント発生時間が得られた場合には一般に性能の良さが認知されている Cox 回帰分析に一致する。また、尤度の構成に関しても Cox 回帰分析の自

然な拡張であるため受け入れられ易いと思われ、母数の推定においても SAS/PHREG プロシジャを利用しているのでプログラムの妥当性の保証も容易であり、著者は紹介する方法およびプログラムが広く利用されることを期待している。

本稿では、まず、第2章において Satten の周辺尤度アプローチを概説すると共に作成した SAS マクロプログラムにおける推定方法を示す。続く第3章で作成したプログラムの仕様および実際のデータへの適用例を示し、第4章では第3章の実例の状況および区間打ち切りデータが観察される具体的ないくつかの状況を設定してモンテカルロ・シミュレーションにより紹介する推定方法およびマクロプログラムの性能評価を行い、最後に第5章で考察する。

2 ギブス・サンプラーを利用した周辺尤度アプローチ

2.1 Satten (1996) の周辺尤度アプローチ

Satten (1996) は、Kalbelesch and Prentice (1973) および Prentice (1978) と同様に、潜在 (結果) 変数である真の生存時間 $T_i (i = 1, 2, \dots, n, n : \text{対象者数})$ の順位に関する周辺分布に基づき推論を行った。

Kalbelsch and Prentice (1973) および Prentice (1978) は、データの背後に存在する (潜在的な) 打ち切りを受けていない生存時間が、打ち切りが存在したことによって一部だけ観察されたと仮定した。いま、4人 (A, B, C, D) の生存時間が 112, 69⁺, 32, 112⁺ とそれぞれ観察されていたとする。数字の右肩の⁺は、その対象の観察が右側打ち切りを受けていることを示す。

この場合、観察データに矛盾しない順位統計量ベクトルは、

$$\{(C, B, A, D), (C, A, B, D), (C, A, D, B)\}$$

の成分となる。たとえば対象 B は生存時間が 2 番目に大きい、あるいは 3 番目に大きい、あるいは最も大きい可能性がある。ここで、区間打ち切りデータに対する周辺尤度を、観察されたデータに矛盾しない $T_i (i = 1, 2, \dots, n)$ に対応する可能な順位ベクトル $R (= (r_1, \dots, r_n)' \in \mathcal{R}, \mathcal{R}$ は全ての可能な順位) それぞれが生じる確率の和として以下のように構成する。

$$\text{周辺尤度: } \mathcal{L} = \sum_{R \in \mathcal{R}} pr(R|\beta, \mathbf{x}_i) \quad (1)$$

この $pr(R|\beta, \mathbf{x}_i)$ は R が与えられた下での Cox (1972, 1975) の部分尤度、

$$L(\beta) = \prod_i \left[\frac{\exp(\beta' \mathbf{x}_i)}{\sum_{u \in R_i} \exp(\beta' \mathbf{x}_u)} \right]^{\delta_i} \quad (2)$$

に対応する。ただし、 β はパラメータベクトル、 \mathbf{x}_i は共変量ベクトル、 δ_i はイベントを観察すれば 1、それ以外は 0 をとる指示変数、および R_i は T_i でのリスクセットである。

2.2 パラメータ β の (点推定) 事後分布の推定方法

前節で構成した (区間打ち切りデータに対する) 周辺尤度 (式 (1)) は, データ数あるいは区間打ち切りを受けた対象が少ないために可能な順位ベクトル R の個数が少ないような特別な場合を除いて, 一般にはデータとパラメータの関係が複雑で簡単にはパラメータ推定を行うことができない. そこで Satten (1996) はギブス・サンプラーを用いて, R の系列を \emptyset の下で確率的に発生させる事によりパラメータの事後分布を得る手順を考えた. 著者は, Satten (1996) の考えをアレンジして以下のパラメータ (点推定) 事後分布を得るアルゴリズム *Step 1*~*Step 4* のプログラムを作成した.

Step 1. 任意の比例ハザード族に属する (生存時間) 分布を決める. 本稿では, Satten (1996) の方法とは若干異なり観察された区間の左側・右側時間による切断分布とする. なお, この切断はデータに矛盾しない順序の生存時間データを発生させる事に対応する. 通常, 簡単のために期待値 $1/\lambda$ の指数分布を用いる.

Step 2. 前回 (第 $s-1$ 回) で得たパラメータ推定値 $\hat{\beta}^{(s-1)}$ および共変量 $\mathbf{x}_i (= (x_{i1}, x_{i2}, \dots, x_{ip})', i = 1, 2, \dots, n, p : \text{共変量の個数})$ を与えた下で, 時間 $t_i (t_i = x/\lambda, x \sim \text{Exp}(1), \lambda = \lambda_0 \exp(\hat{\beta}^{(s-1)'}, \mathbf{x}_i))$ を観察された各対象の区間内に入る乱数を発生させ生存時間とする. ただし $\beta^{(s)} = (\beta_1^{(s)}, \beta_2^{(s)}, \dots, \beta_p^{(s)})'$, 第 1 回目では適当な初期値 $\beta^{(0)}$ を与える. なお, 本稿では初期値 $\beta^{(0)} = \mathbf{0}$ とした.

Step 3. *Step 2.* で発生させた生存時間に対して, SAS/PHREG プロシジャを用いて Cox の比例ハザードモデルを適用する. ここで推定された推定値を第 s 個目の推定値 $\hat{\beta}^{(s)}$ とする.

Step 4. *Step 2.* から *Step 3.* を適当な数 $G+K$ の $\hat{\beta}^{(s)}$ ($s = 1, 2, \dots, G, \dots, G+K$) が得られるまで繰り返し, 第 $G+1$ 回から第 $G+K$ 回までの平均値 $\hat{\beta} = \frac{1}{K} \sum_{s=G+1}^{G+K} \hat{\beta}^{(s)}$ を母数の最終点推定値とする. また, 反復それぞれにおける母分散の推定値を $\hat{V}^{(s)}$ とする. G および K については後述する.

2.3 パラメータ β の母分散の推定方法

前節の *Step 3.* の反復ごとに得られる母数分散の推定量は実際の観察データの母数分散推定量としてはバイアスが存在し不適切である. それは, 各々の反復で発生させた順位統計量は完全データであり, 実際に観察された順位情報 (不完全データ) より多くの情報を持っていることに起因する. したがって, 母数分散の事後分布から直接得られる分散の推定量の期待値は, 母数の真の分散に対して必ず等しいかあるいは過小評価となる. これと同様の問題は, EM アルゴリズム (Dempster *et al.*; 1977) を用いた推測においても生じる. EM アルゴリズムの枠組みにおいては, Louis (1980) がこの過小評価の大きさを導き, その推定法を提案している. Louis (1980) によると, 母数の情報量 I_Y は以下のように得られる.

$$I_Y = I(\hat{\beta}) = I_X - I_{X|Y} \quad (3)$$

ただし, X : 完全データ, Y : 不完全データである. すなわち, いま知りたい母数分散 V_Y は情報量 I_Y の逆数として得られる.

ここで、式 (3) の右辺第二項は母数の最終点推定値を与えた下でのスコア関数の分散に置き換えることで推定できる。紹介するプログラムにおける母数分散の推定アルゴリズム Step 5.~Step 6. を以下に示す。

Step 5. 前回 (第 $G + K$ 回) のパラメータ推定値 $\hat{\beta}^{(G+K)}$ を用いて、Step 2. から Step 3. を K 回繰り返す。ここで、反復ごとの通常のスコア推定値ではなく、反復それぞれのスコア関数の母数に最終点推定値 $\hat{\beta}$ を代入したスコアを $\hat{U}^{(s)}$ ($s = G + K + 1, G + K + 2, \dots, G + 2K$) とする。ここで、スコア関数 U は Cox の部分尤度 (式 (2)) の対数を一階微分した下式より得られる。

$$U = U(\beta) = \frac{\partial \ln L(\beta)}{\partial \beta} = \sum_i \delta_i \left(x_i - \frac{\sum_{u \in R_i} x_u \exp(\beta' x_u)}{\sum_{u \in R_i} \exp(\beta' x_u)} \right)$$

Step 6. 式 (3) の右辺の第一項、第二項の推定値を、順に $\hat{I}_X = \frac{1}{K} \sum_{s=G+1}^{G+K} (1/\hat{V}^{(s)})$, $\hat{I}_{X|Y} = \hat{V}ar(U) = \frac{1}{K-1} \sum_{s=G+K+1}^{G+2K} (\hat{U}^{(s)} - \bar{U})^2$ とし、母数分散の推定値 $\hat{V}_Y = 1/\hat{I}_Y = 1/(\hat{I}_X - \hat{I}_{X|Y})$ を得る。ただし、 $\bar{U} = \frac{1}{K} \sum_{s=G+K+1}^{G+2K} \hat{U}^{(s)}$ である。

なお、Step 5. および Step 6. のスコアの分散推定のためのプログラムは、SAS/IML プロシジャを用いて作成した。このアルゴリズムでは、Step 1.~Step 6. においてギブス・サンプリングの反復を合計 $G+2K$ 回行う。

3 作成したマクロプログラム (%phregintcens)

3.1 仕様

本マクロプログラムはごく簡単な指定のみで実行できるように工夫した。事前準備として、打ち切りの左側、右側時間それぞれに対応する変数および共変量を含む解析用データセットを用意する。打ち切り時間に対応する変数の入力方法を表 1 に示す。これは、SAS/LIFEREG プロシジャにおける打ち切り生存時間を含むデータ解析に使用するデータセットと同様である。

表 1: 打ち切り時間変数

左側時間	右側時間	比較	解釈
非欠測	非欠測	=	非打ち切り (正確なイベント発生時間)
非欠測	非欠測	左側時間 < 右側時間	区間打ち切り
欠測	非欠測		左側打ち切り
非欠測	欠測		右側打ち切り
非欠測	非欠測	左側時間 > 右側時間	解析に用いない
欠測	欠測		解析に用いない

次に、作成したマクロ (%phregintcens) のヘッダーから入出力パラメータの説明部分を抜粋したものを表 2 に示す。

表 2: マクロ (%phregintcens) の入出力パラメータの仕様

```

Input : DATA.sas7bdat(dataset to analyze)
        LOWER      : left censoring time variable
        UPPER      : right censoring time variabe
        COVARIATES: explanatory variables(numerical, delimited by space)
        BETA0      : initial values of parameters(delimited by space)
        NiteG      : number of iteration of Gibbs sampling until stationary
        NiteK      : number of iteration of Gibbs sampling for estimating the
                    posterior distributions of parameters
        SEED       : seed of exponential randam variable
Output: Point estimate, SE, Hazard ratio, 95%CI of hazard ratio
        and Fig of change of parameter estimate
        %phregintcens(DATA, LOWER, UPPER, COVARIATES, BETA0, NiteG, NiteK, SEED);
    
```

このマクロプログラムは、任意の数の共変量の推定を入力パラメータ COVARIATES にスペースを空けて入力することにより実行する。ただし、共変量の変数のタイプは数値とし、交互作用についてはダミー変数を作成して入力する (PHREG プロシジャの class ステートメントに対応していない)。

本稿で紹介したギブス・サンプラーを利用した方法では、サンプル数 (イベント数) およびサンプリング回数が十分に大きい下で、パラメータ β の推測に対して基準分布は比例ハザード族であれば任意でよい。しかし作成したプログラムでは推定精度の向上を期待して、区間打ち切りデータの右側時間に対して、指数分布を SAS/LIFEREG プロシジャより推定し、この推定値を第 2.2 節 *Step 1* の基準分布とした。

本稿では、パラメータの推定事後分布が定常となるのに必要なギブス・サンプラーの反復数 G を 1000 回とし、事後分布の推定に使用するサンプリング数 K も同様に 1000 回とした。この妥当性については後で考察する。

またタイデータの処理には Efron 法を用いた。正確法としなかったのは第 2.3 節の *Step 5* および *Step 6* におけるスコア関数の SAS/IML プロシジャの記述に関して、正確法の尤度に基づいた記述をすることが難しいからである。点推定については、本プログラム中の SAS/PHREG プロシジャ Model 文のオプションを ties=exact と変更することで正確法に基づいた推定を行うことができる。実用上は、オプションを ties=exact に変更して分散の過小推定分のみを Efron 法で推定しても推定値に対する影響は小さいと考えられる。

3.2 適用例

Whitehead (1989) に掲載されている胃潰瘍および胃癌再発をイベントとした 301 症例の臨床試験データを用いる。このデータには、ランダム割付された 2 治療群の試験開始・6ヶ月・12ヶ月時の内視鏡検査によるイベント発生および 4 つの共変量の情報を含む。ただし、患者の症状の訴えによりイ

イベント発生の情報が得られている場合もある。ここでは、共変量 G(2 治療群) および Age を用いて、定期的内視鏡検査によりイベント発生が観察された場合は区間打ち切りデータ、また患者の症状の訴えに基づく内視鏡検査でイベントが観察された場合は正確なイベント発生時間が得られたものとして解析した。イベントタイプの内訳は、正確発生時間、区間打ち切り時間、および 12ヶ月時点で観察打ち切りの順に 49 例、21 例、231 例である。なお、このデータは Collett (1994, Chapter. 8) にも区間打ち切り生存データ解析の例として取り上げられている。

SAS データセット ULCER には、上記事例の打ち切り時間を表 1 に倣って変数 LOWER および UPPER に、また共変量の群と年齢をそれぞれ変数 G と AGE に入力されている。以下に示す表 3 および図 3 は下枠を実行することにより得られる。ただし、作成したマクロプログラムが実際に出力するパラメータの推移図は 2000 回 (G+K 回) までである。

```
%phregintcens(DATA=ULCER, LOWER=LOWER, UPPER=UPPER, COVARIATES=G age,
              BETA0=0 0, NiteG=1000, NiteK=1000, SEED=4649) ;
```

この場合の計算時間は、Intel® Pentium® 4 CPU 2.80GHz メモリ 1.5GB の PC で 440 秒であった。ギブス・サンプラーによるパラメータ推定値の推移を図 3 に示す。この推移は、反復 1 万回の推定値の 10 回ごとの平均値である。パラメータの推定事後分布が速やかに定常となることが分かる。

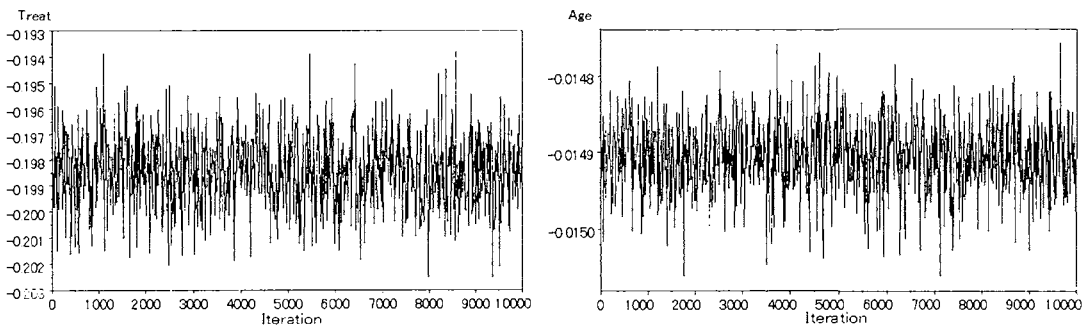


図 2: ギブス・サンプラーによるパラメータ推定の推移

ギブス・サンプラー (G=K=1000) による推定結果、および右側時間をイベント発生時間とした Cox 回帰分析による推定結果を表 3 に示した。この事例では、Cox 回帰分析とギブス・サンプラーの推定結果が点推定値が若干異なり、信頼区間幅はギブス・サンプラーのほうが多少広いことがわかる。しかし、この状況においてギブス・サンプラーの方が正確な推定法であることを次章でモンテカルロ・シミュレーションを用いて示す。

4 性能評価

本章では、紹介するプログラムの性能を具体的な状況を複数設定してモンテカルロ・シミュレーションにより評価する。同時比較対象として、一般に良く行われている初めてのイベント観察時間をイベント発生時間とした (右側補完) 通常の Cox 回帰分析による推定結果を示す。

表 3: 推定結果

推定法	パラメータ	推定値	標準誤差	ハザード比 [95%信頼区間]
ギブス・サンプラー	Treat	-0.1982	0.2394	0.820 [0.513~1.311]
	Age	-0.0149	0.0078	0.985 [0.970~1.000]
右側補完	Treat	-0.2079	0.2305	0.812 [0.517~1.276]
	Age	-0.0152	0.0068	0.985 [0.972~0.998]

4.1 状況設定

各群 100 例の 2 群比較を考えて、生存時間分布に指数分布および共変量の効果に比例ハザード性を仮定する。ここで、観察スケジュールを次の 5 通り設定する。

スケジュール A：基準群の生存分布を $\lambda_0 = 1$ の指数分布、および観察間隔は 2 群の全観察対象で各々独立な $\lambda = 2$ 指数分布に従うとする。

スケジュール A'：スケジュール A において $t = 0.4$ で観察打ち切りとする。たとえば、骨粗鬆症における新規骨折をイベントとするような長期に渡る観察研究においては、観察期間中のイベント発生率が 2~4 割であることも珍しくない。この場合、イベント発生率は約 33% である。

スケジュール B：群間で区間打ち切りの間隔幅が異なる場合として、基準群の生存分布を $\lambda_0 = 1$ の指数分布、基準群および比較群の対象の観察間隔をそれぞれ $\lambda = 1, 1.5$ の各々独立な指数分布に従うとする。

スケジュール C：抗がん剤のランダム化臨床試験の例として、基準群の憎悪までの期間の分布を $\lambda_0 = 1$ の指数分布とする。比較する 2 群に次のレジメンを考える。標準レジメン (基準) 群：各コース Day1 に抗がん剤の投与を行う。1 コースは 28 日間、検査日は各コース Day1 のみとする。強化レジメン群：各コース Day1 および Day15 に抗がん剤の投与を行う。1 コースは 28 日間、入院を必要とする治療と想定し、この場合は憎悪までの時間は正確に測定できるものとする。

スケジュール D：第 3.2 節の事例に沿って共変量を群および年齢の 2 つとした状況を考える。対象それぞれの年齢は互いに独立に平均 50、分散 15^2 の正規分布に従うものとする。基準生存分布を $\lambda_0 = 0.05$ の指数分布、および時刻 0, 6, 12 で観察が実施されるとし、群および年齢に対応する真のパラメータをそれぞれ $\beta_1 = -0.2$ および $\beta_2 = -0.015$ とする。ただし、事例は正確なイベント時間を含む生存時間データが観察されているが、ここではイベント発生時間はすべて区間打ち切り生存時間とした。

スケジュール A, A', B, C については、推定するパラメータ β_1 (基準群に対する他方の群の効果) の真値を 0, -0.5, -1.0 の 3 通りで検討する。

表 4: シミュレーション結果 1

観察スケジュール	推定法	真値 (β_1)	推定値 ($\hat{\beta}_1$)	95%信頼区間の被覆割合 (%)
A	ギブス・サンプラー	0	0.0032	96.4
		-0.5	-0.4676	94.0* ²
		-1.0	-1.0090	95.6* ¹
	右側補完	0	0.0000	95.3
		-0.5	-0.4973	95.2* ²
		-1.0	-0.9593	93.9* ¹
A'	ギブス・サンプラー	0	0.0124	95.5
		-0.5	-0.4994	95.3
		-1.0	-1.1748	97.9
	右側補完	0	0.0225	95.7
		-0.5	-0.4807	95.7
		-1.0	-1.1404	97.6
B	ギブス・サンプラー	0	0.0500	96.0
		-0.5	-0.4041	92.7
		-1.0	-0.8819	91.2
	右側補完	0	0.2533	58.2
		-0.5	-0.2043	46.0
		-1.0	-0.7018	50.0
C	ギブス・サンプラー	0	-0.0075	94.7* ³
		-0.5	-0.5030	95.5* ³
		-1.0	-1.0039	95.1* ³
	右側補完	0	-0.0910	92.1* ³
		-0.5	-0.5682	94.1* ³
		-1.0	-1.0520	95.0* ³

*1 : 10000 回, *2: 5000 回, *3: 2000 回, その他 : 1000 回

4.2 結果

モンテカルロ・シミュレーション 1000 回による推定値 (最終点推定値の平均値) および 95%信頼区間の被覆割合 (coverage probability) を表 4 および表 5 に示す。

スケジュール A, A' では, ギブス・サンプラーと右側補完で同様に安定した推定結果が得られている。しかし, スケジュール B のように群間で区間打ち切りの間隔幅が大きく異なるような場合では, 通常の推定方法 (右側補完) では点推定値および区間推定ともにまったく妥当でないことがわかる。これに対して, ギブス・サンプラーは群間で区間打ち切り幅が異なっている場合でも比較的安定した推定が行われている。またスケジュール C は, 群間で観察されるイベントデータのタイプが異なっている状況であり, 通常の推定法では点推定および区間推定ともに推定値にバイアスがあるが, ギブス・サンプラーでは点推定の精度が良く, 区間推定の精度も安定している。さらに, 共変量が 2 つ (群・

年齢) の場合のスケジュール D でも、通常の推定法は区間推定の精度が悪いのに対してギブス・サンプラーは安定した推定結果が得られている。

全体として、通常の推定法 (右側補完) は点推定の精度が悪く、区間推定はリベラルな傾向 (被服割合が 95% より小さい) であり有意水準が保たれないという観点から妥当でない。これに対して、ギブス・サンプラーは常に精度良く点推定が行われており、区間推定は正確あるいは保守的 (被服割合が 95% 以上) な傾向となり妥当な推定法と言える。

表 5: シミュレーション結果 2(2500 回)

観察スケジュール	推定法	パラメータ	真値 (β)	推定値 ($\hat{\beta}$)	95%信頼区間の被覆割合 (%)
D	ギブス・サンプラー	β_1 (群)	-0.2	-0.2082	95.2
		β_2 (年齢)	-0.015	-0.0154	94.6
	右側補完	β_1 (群)	-0.2	-0.2080	93.2
		β_2 (年齢)	-0.015	-0.0154	92.0

5 考察

5.1 ギブス・サンプラーについて

ギブス・サンプラーを適用することの問題点は、ギブス・サンプラーの反復数 $G+K$ 、および推定するパラメータの初期値 $\beta^{(0)}$ の選択に解析実施者の自由度が入ることである。前者は、パラメータの推定事後分布が定常であるとするための反復数および事後分布の推定のための反復数のことであるが、著者の検討では、第 3.2 節の事例 (図 2) のように推定するパラメータが少ない場合であれば反復 1000 回で十分、あるいはもっと少なくても良さそうである。これは、今回の状況でサンプリングに用いる指数分布の性質が良いからであると思われる。ただし、反復 1 万回でも通常に使用されている PC で実行可能なので、十分多くの反復を行いその推移を図示することが重要であろう。また、十分な検討は行っていないものの母数推定のためのサンプリング数 K を増やすと精度の向上が期待できる。後者の初期値については、十分な反復の下で収束先は同じであり、また通常は 0 で良いと考えられるが、収束速度が遅くなるような場合があるかどうかは未検討である。

5.2 加速モデルへの拡張

SAS/LIFEREG プロシジャでは、区間打ち切り生存時間データに対応した解析法が用意されているが、著者が予備的に前章と同様の検討を行ったところ推定精度が悪かった。本稿で紹介したギブス・サンプラーの考え方は加速モデルにも適用可能であり、これは今回作成したプログラムの簡単な修正により実行できる。

また、著者の検討では、スケジュール A', B, C のように観察期間が定められておりイベント発生数が少ないデータに対する推測は、そもそも区間打ち切りでない正確なイベント発生時間が得られる場

合であってもパラメータ推定にバイアスが入る。すなわち、現状考える推定法ではこのような状況において加速モデル(パラメトリックモデル)の正確なパラメータ推定は難しいと思われる。

6 おわりに

Yoshimura *et al.* (2003) は、本稿で紹介した方法の理論的側面の妥当性および多変量の区間打ち切り生存時間データに拡張した推定方法について今後報告する予定である。また、紹介したマクロプログラム(%phregintcens)および事例データを作成するプログラム(Ulcer.sas)は、東京理科大学「医薬統計コース」ホームページ(www.rs.kagu.tus.ac.jp/yoshilab/iyaku/top.html)に公開する予定である。

参考文献

1. Collett, D. (1994). *Modelling survival data in medical research*. Chapman & Hall.
2. Cox, DR. (1972). Regression models and life table (with discussion). *JRSSB* **34**: 187-220.
3. Cox, DR. (1975). Partial Likelihood. *Biometrika* **62**: 269-276.
4. Dempster, AP., Larid, NM. and Rubin, DB. (1977). Maximum Likelihood from incomplete data via EM algorithm (with discussion). *JRSSB* **39**:1-38.
5. 岩崎学 (2002). 不完全データの統計解析. エコノミスト社.
6. Kalbelesch, JD. and Prentice, RL. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**: 267-278.
7. Kim, MY. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Stat. Med.* **21**: 3715-3726.
8. Louis, TA. (1982). Finding the observed information matrix when using the EM algorithm. *JRSSB* **44**: 226-233.
9. Prentice, RL. (1978). Linear rank tests with right-censored data. *Biometrika* **65**: 167-179.
10. Satten, GA. (1996). Rank-based inference in the proportional hazards model for interval-censored data. *Biometrika* **83**: 355-370.
11. Whitehead, J. (1989). The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Stat. Med.* **8**: 1439-1454.
12. Yoshimura, K, Nishiyama, H. and Ohashi, Y. The semi-parametric model for the analysis of multivariate time-to-event data with interval censoring. (to be presented at Joint Stat. Mtg. in Aug. 2003),

イベント発生確率推定時における連続変数のカテゴリー化、

およびカテゴリ変数の実数化

○上條 史夫、川崎 章弘

株式会社 数理技研

金融工学センター

Categorization of continuous variable and translation of categorical variable values to real numbers in estimation of probability for event occurrence

Fumio KAMIJO, Akihiro KAWASAKI

Financial Engineering Center, SURIGIKEN CO., LTD.

要 旨

十分なサンプル数が与えられた時に、多変量解析においてその推定精度を上げる為の、連続変数のイベント発生確率を用いたグループ化(カテゴリ分割)、及び、その後のカテゴリ分割されたグループに対する単変量解析をベースとした実数化方法を提示し、これを多変量ロジスティック回帰モデル、cox 回帰モデルに適用することで推定モデルの高精度化を図る方法について紹介する。またこの手法を用いたコックス回帰モデルを消費者金融の例に適用し、貸し倒れ確率の推定精度が向上したことを報告する。

キーワード：

コックス回帰モデル、カプランマイヤー法、ロジスティック回帰、生存時間関数、ロジット,phreg

1. はじめに

我々が個人金融における貸し倒れ確率の推定システムを開始するにあたって、当初単純にコックス回帰モデルを適用する、という方針で進めたが、確かに期間構造をもった生存確率はえられたが、正誤判別率、デフォルト補足率を他の手法と比較検討したとき、多種法を超えるものは得られなかった。この原因は線形性を前提とした回帰モデルの所にあった。

一般にロジスティック回帰や生存時間解析の cox 回帰モデルを作成するにあたって使用される説明変数は、その意味の分類を表す、カテゴリ変数と年齢、金額、利率等の連続変数とに大別される。通常回帰モデル作成では、これらの変数の値を直接用いて確率モデルの係数の決定がなされている。しかしこれらの値の大小関係が直接確率に反映するものではない為、多くの場合作成されたモデルの推定精度は劣化している。これは上記回帰モデルが確率の対数または二重対数に対して変数値が線型関係にあることを前提としているからである。

本論分では、連続変数をイベント発生確率を元にグループ化(カテゴリ分割)し、その後カテゴリ分割されたグループに対し Kaplan-Meier 法を用いて実数化する方法を提示し、この結果としてロジスティック回帰モデル、cox 回帰モデルの高精度化を図る方法について紹介する。またこの手法を消費者金融の例に適用し、貸し倒れ確率の推定精度が向上したことを報告する。

1.1. 既存方式の問題点

Cox 回帰モデル自体は、周知のアルゴリズムであり SAS では PHREG プロシジャーとして実装されている。

ロジスティック回帰であれ、cox 回帰であれある変数値に対する統計量(確率を元に算出された量)は異なる値間で異なる値を示さなければならないことは自明である。一方カテゴリ変数を例に取れば、その値に属するイベントの発生確率が同じでも、異なる値を複数とる場合が多々ある。このような場合直接カテゴリ値を用いると、得られた統計量は必ず異なる値を示すことになり同一確率であるということに対し矛盾してしまう。

我々が扱う対象は、目的となる従属変数がデフォルトする、しないの 2 値状態で、これを複数の変数を用いて、より精度よく分離することが最終目標である。このための基本となる誤差要因は単変量解析時の 1 変数毎の分離精度の優劣である。つまり単一変数での従属変数評価を行なった時の評価誤差を最小にすることである。

1.2. 単変量解析における考え方

1 変数で見た時は、各個人の特性を個別に評価しているのではなく、各人が属している層のデフォルト割合が評価の対象となっているのである。カテゴリ変数の変数値による i 層と j 層の間の相対距離をロジスティック回帰モデル、及び cox 回帰モデルをベースにした場合以下のように定義する。

$$x_i - x_j \equiv \log p_i / (1 - p_i) - \log p_j / (1 - p_j)$$

$$x_i - x_j \equiv \log(-\log(S(t | i層))) - \log(-\log(S(t | j層)))$$

p_i : i 層における死亡確率, $S(t|i)$ 層): i 層における生存時間確率

これをもとに各層間の相対距離を定義し、これから絶対距離を作成、この値を該当カテゴリ変数の独立変数値とする。

1.3. 従来のカテゴリデータ処理とどのように違うのか

従来のカテゴリ変数処理方式(CATMOD)では、デフォルト確率のような 2 値問題に対し、従属変数のロジット化を行い、これを独立変数(カテゴリ変数値)に対し回帰式へのフィッティングを行っていた。このようにすると当然ながら独立変数に対し、多峰性や飽和性を持つ場合の近似精度は悪化する。このために高次項を持ち込んで、何とか近似精度をあげようとしている。この近似モデルを用いて外挿予測する場合は、これでよいが貸し倒れ確率のように、母集団のサンプル数が十分確保されていて、なおかつ母集団間で確率定常性が前提とされている場合は、母集団の説明能力が高い回帰モデルが要求される。本方式は単変量解析レベルでは従属変数をロジットとした場合モデル誤差をほとんどゼロにする変換方式である。

2. カテゴリ変数値の実数化手順

以下に上記層間相対距離を元にして、カテゴリ変数値に実数値を割り当てる際の計算手順を示す。

- ① 1つの整数型変数に対しその値(x_i)を用いて全ケースを分割する(G_i)。
- ② 分割したそれぞれのグループ G を対象にカプランマイヤー法を適用し各グループ毎の生存時間関数($S(t|x_i)$)を計算する。
- ③ それぞれの対数累積ハザード関数($\log(-\log(S(t|x_i)))$)を計算する
- ④ ケース数最大のグループの対数累積ハザード関数を基準として、他のグループとの相対対数累積ハザード関数($\Delta lh(t|x_i)$)を計算する、これは 2 層間の相対距離

$x'_i - x'_b : x'_b$:(基準値)に相当する。

上式に見るように、完全に比例ハザード性が成立していれば、上記相対累積ハザード関数は時間依存性のない定数になるが、実際は若干の時間依存性をもつ、しかしこの関数が相互に並行状態にあれば比例ハザード性が成立していると考えてよい。

- ⑤ 相対累積ハザードの時間平均($x'_i - x'_b = \int_0^T \Delta lh(t|x_i) \delta t / T$)を計算する。

⑥ 最後に基準値である x'_b を決定すれば、すべての x'_i は決定される。現在この決定方法と

してはケース数最大の G_i グループでの貸し倒れ確率 $p(x_i)$ をもとに

$x'_b \equiv \log p(x_i)/(1 - p(x_i))$ として計算している。(相対距離をベースに単変量

ロジスティック回帰を適用してバイアス値を求める方法もある)

尚、本方式は期間指定データがなければ対数オッズ比をベースにした実数化も可能である。

3. 貸し倒れ確率計算への応用

本方式を個人金融分野へ適応した例を以下に示す。

個人金融分野においては、各社数年分、数百万件に及ぶデータを保持しており、十分な精度で貸し倒れ確率を計算できる状態になっている。今回の報告では、当社で 2 年間分、3 万件の実験用データを作成し、これを用いて期間構造分析を行なった。

実際の分析を行なうにあたって、本方式の拡張として連続変数のカテゴリ化も実施した。

これは、連続変数の区分場所をイベント発生割合を用いて決定し、区分化された連続区間をカテゴリとして扱い、実数化を施す手法である。以下に実際に解析を行なうにあたっての手順を示す。

- ① 連続変数の区分化(カテゴリ化)
- ② カテゴリ変数のそのカテゴリ値を上記方法で実数化
- ③ 上記方法で得られた test データの実数共変量を独立変数としてコックス回帰モデルを作成
- ④ 上記方法で得られた score データの実共変量にコックスモデルを当てはめ、生存確率の下界を 99.5,99.、98.5,98,97.5,97.、96,95,93,90,70,50,0 の 13 段階にリスクランク分割を行い各ランクに属する貸し倒れ件数を元に精度検証を行なった。

4. 結果考察

本方式は、決定木、ロジスティック回帰モデルへの前処理としても有効であるが、今回は、生存時間確率の推定精度に的を絞って検討を行なった。

以下に生データ、カテゴリ変数の実数化、及び連続変数のカテゴリ化・カテゴリ変数の実数化の 3 例の結果を示し、精度評価を行なう。

4.1. 3例結果紹介

4.1.1. 生データ

生データ

	下限値	総数	管理移行	完了	総残	総金	管理移行の完了金	総残金	生存確率	標準偏差	指数のみ	予想管理移行	予想管理移行金	
RiskRank[0]	0.995	3789	10	2621	1158	307983.00	773.00	207934.00	98676.00	0.997457	0.000002	0.187618	9.64	782.18
RiskRank[1]	0.990	6021	11	4301	1709	343818.00	904.00	238995.00	103919.00	0.992435	0.000002	0.559353	45.55	2600.98
RiskRank[2]	0.985	5326	27	3822	1477	257543.00	1655.00	183881.00	72007.00	0.987625	0.000002	0.917231	65.91	3187.09
RiskRank[3]	0.980	3605	27	2594	984	164115.00	1594.00	114952.00	47569.00	0.982653	0.000002	1.288906	62.54	2846.90
RiskRank[4]	0.975	2462	20	1834	608	103897.00	1129.00	75947.00	26821.00	0.977671	0.000002	1.663249	54.97	2319.92
RiskRank[5]	0.970	1684	26	1301	357	66652.00	1339.00	50351.00	14962.00	0.972678	0.000002	2.040375	46.01	1821.07
RiskRank[6]	0.960	2109	48	1570	491	77469.00	2442.00	55955.00	19072.00	0.965608	0.000008	2.577954	57.47	2664.31
RiskRank[7]	0.950	1289	47	960	282	41502.00	1948.00	29507.00	10047.00	0.955418	0.000008	3.359322	57.47	1850.24
RiskRank[8]	0.930	1194	56	894	244	35115.00	2476.00	24680.00	8009.00	0.941465	0.000034	4.443924	69.89	2055.46
RiskRank[9]	0.900	749	55	546	148	20197.00	2085.00	13749.00	4363.00	0.917098	0.000075	6.3771	62.09	1674.37
RiskRank[10]	0.800	530	74	375	81	14200.00	2845.00	8880.00	2475.00	0.867494	0.000723	10.505274	70.23	1881.59
RiskRank[11]	0.500	239	89	134	16	8660.00	2868.00	4936.00	856.00	0.682549	0.00613	28.801942	75.87	2749.13
RiskRank[12]	0.000	539	437	99	3	17631.00	14213.00	3069.00	349.00	0.126443	0.022224	310.7181	470.85	15401.68
合計		29536	927	21051	7558	1458382	36221	1012836	409325			1163.54	41834.92	

件数

分類実績		予測値	
		0	1
実績	0	28357	252
実績	1	401	526
値		136	178
		28758	778
		9737	263

デフォルト予測後の分類		予測値	
		0	1
実績	0	28141	231
実績	1	9528	547
値		617	185
		209	185
		28758	778
		9737	263

金額

分類実績		予測値	
		0	1
実績	0	1412951	9210
実績	1	9688	663
値		19140	17081
		131	117
		1432091	26291
		9820	180

デフォルト予測後の分類		予測値	
		0	1
実績	0	1408407	8140
実績	1	9657	556
値		23684	18151
		162	124
		1432091	26291
		9820	180

4.1.2. カテゴリ変数の実数化

CREDIT SAVER 方式(カテゴリ変数実数化)

	下限値	総数	管理移行	完了	総残	総金	管理移行の完了金	総残金	生存確率	標準偏差	指数のみ	予想管理移行	予想管理移行金	
RiskRank[0]	0.995	7804	2	5494	2308	525668.00	272.00	359449.00	165947.00	0.996374	0.000001	0.431384	28.30	1906.07
RiskRank[1]	0.990	14601	64	10639	3898	682805.00	3777.00	487241.00	191847.00	0.992894	0.000002	0.833426	104.54	4889.31
RiskRank[2]	0.985	4031	52	2976	1003	147810.00	2253.00	106467.00	39090.00	0.988815	0.000002	1.415872	47.77	1751.55
RiskRank[3]	0.980	918	29	685	204	23307.00	963.00	16244.00	6100.00	0.983044	0.000002	2.03078	15.57	395.19
RiskRank[4]	0.975	282	10	199	73	7107.00	326.00	4688.00	2093.00	0.977767	0.000002	2.669838	6.27	158.01
RiskRank[5]	0.970	94	1	73	20	3645.00	49.00	2756.00	840.00	0.972917	0.000002	3.260422	2.55	98.72
RiskRank[6]	0.960	92	2	74	16	4583.00	78.00	3597.00	908.00	0.965846	0.000009	4.126702	3.14	156.52
RiskRank[7]	0.950	46	1	41	4	2314.00	50.00	2040.00	224.00	0.955534	0.000008	5.401514	2.05	102.89
RiskRank[8]	0.930	111	2	98	11	6038.00	128.00	4721.00	1189.00	0.940839	0.000032	7.243373	6.57	357.21
RiskRank[9]	0.900	96	9	85	2	4649.00	527.00	4030.00	92.00	0.915932	0.000071	10.432124	8.07	390.83
RiskRank[10]	0.800	182	13	166	3	6238.00	618.00	5289.00	331.00	0.852464	0.000893	19.027388	26.85	920.33
RiskRank[11]	0.500	422	136	279	7	15199.00	6153.00	8779.00	267.00	0.64363	0.007551	53.415775	150.39	5416.47
RiskRank[12]	0.000	857	606	242	9	28959.00	21027.00	7535.00	397.00	0.252351	0.019048	199.20668	640.74	21651.17
合計		29536	927	21051	7558	1458382	36221	1012836	409325			1042.79	38194.28	

件数

分類実績		予測値	
		0	1
実績	0	28072	537
実績	1	9504	182
値		185	742
		63	251
		28257	1279
		9567	433

デフォルト予測後の分類		予測値	
		0	1
実績	0	28005	488
実績	1	9482	165
値		252	791
		685	268
		28257	1279
		9567	433

金額

分類実績		予測値	
		0	1
実績	0	1405183	16978
実績	1	9635	116
値		9041	27180
		62	186
		1414224	44158
		9697	303

デフォルト予測後の分類		予測値	
		0	1
実績	0	1403097	17090
実績	1	9621	117
値		11127	27068
		676	186
		1414224	44158
		9697	303

4.1.3. 連続変数のカテゴリー化・カテゴリ変数の実数化

	下限値	総数	管理移行	完済	継続	総金	管理移行金	完済金	継続金	生存確率	標準偏差	指数・重み	予想管理移行	予想管理
RiskRank[0]	0.995	11842	4	8437	3401	704913.00	255.00	489357.00	215301.00	0.997593	0.000002	0.330479	28.50	1696.73
RiskRank[1]	0.990	6349	9	4616	1724	300382.00	632.00	212898.00	86852.00	0.992774	0.000002	0.994258	45.88	2170.56
RiskRank[2]	0.985	3388	14	2462	912	150011.00	860.00	105210.00	43941.00	0.987747	0.000002	1.690069	41.51	1838.08
RiskRank[3]	0.980	2025	17	1496	512	86117.00	955.00	61912.00	23250.00	0.982732	0.000002	2.387828	34.97	1487.07
RiskRank[4]	0.975	1311	11	981	319	53661.00	449.00	39181.00	14031.00	0.977726	0.000002	3.08781	29.20	1195.25
RiskRank[5]	0.970	920	19	689	212	37648.00	1063.00	28101.00	8484.00	0.972621	0.000002	3.805478	25.19	1030.76
RiskRank[6]	0.960	1016	26	766	224	35805.00	1142.00	26157.00	8506.00	0.965626	0.000008	4.795182	34.92	1230.76
RiskRank[7]	0.950	505	13	399	93	18555.00	898.00	14280.00	3377.00	0.955553	0.000008	6.232681	22.45	824.71
RiskRank[8]	0.930	521	30	407	84	16986.00	1322.00	12603.00	3061.00	0.94132	0.000035	8.291802	30.57	996.74
RiskRank[9]	0.900	313	19	253	41	8743.00	523.00	6952.00	1268.00	0.918049	0.000072	11.726345	25.65	716.50
RiskRank[10]	0.700	561	160	367	34	19235.00	7066.00	10963.00	1206.00	0.821871	0.003271	27.230225	99.93	3426.31
RiskRank[11]	0.500	269	174	94	1	9948.00	6748.00	3158.00	42.00	0.605907	0.003154	69.278368	106.01	3920.44
RiskRank[12]	0.000	516	431	84	1	16378.00	14308.00	2064.00	6.00	0.237269	0.026772	679.1374	393.57	12492.01
合計		29536	927	21051	7558	1458382	36221	1012836	409325				918.36	33025.92

分類実績		予測値	
件数		0	1
実績	0	28429	180
実績	1	96.25	0.61
値	1	322	605
		1.09	2.05
		28751	785
		97.34	2.66

デフォルト予測後の分類		予測値	
件数		0	1
実績	0	28332	285
実績	1	95.92	0.97
値	1	419	508
		1.42	1.69
		28751	785
		97.34	2.66

金額		予測値	
金額		0	1
実績	0	1416891	5270
実績	1	97.15	0.36
値	1	15165	21058
		1.04	1.44
		1432056	26326
		98.19	1.81

デフォルト予測後の金額		予測値	
金額		0	1
実績	0	1415443	9914
実績	1	97.06	0.68
値	1	16613	16412
		1.14	1.13
		1432056	26326
		98.19	1.81

4.2. 各種評価指標を用いた結果まとめ

4.2.1. 誤差率による精度評価

生存率 70%を基準とした判別率を以下に示す。

件数	生データ	正誤率
	生データ	97.79
	カテゴリ実数化	97.55
	連続変数処理後	98.3

4.2.2. デフォルト補足率による精度評価

生存率 70%でみたデフォルト補足率

件数	生データ	デフォルト補足率
	生データ	56.74%
	カテゴリ実数化	80.04%
	連続変数処理後	65.26%

4.2.3. 機会損失件数による精度評価

生存率 70%でみた機会損失件数

件数	生データ	機会損失件数
	生データ	252
	カテゴリ実数化	537

4.3. 考察

正誤判別率、機会損失件数で見た場合、カテゴリ変数の実数化を施したケースが最悪値を示しているが、管理移行件数のリスクランク毎の分布をみればわかるように、生データに比べて格段の分離精度が得られていることがわかる、また連続変数処理を施すことで判別率、機械損失件数が更に向上したことがわかる。

以上のことから、単変量解析のレベルで、カテゴリ変数に対し、本方式の実数値化変換をすることでデフォルト確率の推定精度を向上させることができること、及び連続変数に対してカテゴリ化することで更に精度向上が図れることが示された。

日本SASユーザー会 (SUGI-J)

SASによる生存時間の多重イベントの解析～糖尿病合併症を例に～

○ 広本 篤・金子 徹治・大橋靖雄
(東京大学大学院医学系研究科健康科学・看護学専攻)

Survival analysis for multiple events featuring Diabetes Mellitus complications
Atsushi Kohmoto , Tetsuharu Kaneko , Yasuo Ohashi

School of Health Sciences and Nursing, Graduate School of Medicine,
The Univ.of Tokyo

要 旨

生存時間の適用場面では、それぞれの対象個体について相関のあるイベントが複数観測・解析される場合がある。このような多重イベントの解析は、SAS System 8からのPHREGプロシジャによって実行が可能となった。多重イベントに対する解析手法はいくつか存在するが、本論文では糖尿病網膜症の左右眼の発症と糖尿病合併症の発症とを例に、PHREGプロシジャで周辺モデルの当てはめを行った式例を紹介する。

キーワード： PHREG プロシジャ、Multiple Failure Outcomes、周辺モデル

1. はじめに

生存時間解析は、元々は非再起的な事象を対象としたものであるが、臨床研究では1人の対象者について複数の相関のあるイベントが起こる場合がある。例えば糖尿病網膜症の左右眼の発症や糖尿病合併症(腎症、神経症、大血管症など)の発症がこれにあたる。

前者については、左右眼のうち早く網膜症が発症・進展した時点をその患者のイベント発生時点とした解析が行われることがいままでは一般的であった。しかし、これは情報を全て利用した解析とはなっていない。両眼のデータを利用して解析を行うことにより推定効率が高くなる可能性がある。

後者については、糖尿病合併症は全身の血管病変に由来する類似した作用機序を持つ疾患と推測されており、各合併症はおのおの独立に発生するとは考えにくい。つまり、1つの合併症の発症が他の合併症の発症に関する間接的な情報を持っており、合併症の発症・進展因子を検討する際に、合併症ごとに検討するだけでなく、複数の合併症の発症を同時に

検討することも必要であると考えられる。

しかしながら、1人の対象者から得られたイベントの間には相関があり、このようなデータに対して観測データの独立性を仮定する解析方法を適用することは妥当でない。1 対象者内で高い相関のあるイベントが起こっている場合の解析方法として周辺モデルの適用が提案されている。本研究では、糖尿病網膜症の左右眼の発症と糖尿病合併症の発症を例に、複数のイベント間の相関を考慮した周辺モデルによる解析を行う。

多重イベントに対して周辺モデルの当てはめを行う方法は、SAS バージョン 8 のマニュアルから紹介されている。

2. SAS の PROC PHREG の概説

今回用いる PROC PHREG のステートメントを以下に概説する。

```
PROC PHREG < options >;
  MODEL response < *censor(list) > = variables < /options >;
  < programming statements >
  STRATA variable < (list) > < ...variable < (list) >< /option >;
  < label: > TEST equation1 < ,...,equationk > < /option >;
  ID variables;
  OUTPUT < OUT=SAS-data-set >
  < keyword=name... keyword=name > < /options >;
```

MODEL ステートメントは生存時間を表す変数、打ち切り変数、説明変数を表す変数を特定する。STRATA ステートメントは層を表す変数を特定する。ID ステートメントはアウトプットされるデータセット中のオブザベーションにつけるラベルの値の変数を特定する。OUTPUT ステートメントにより様々なデータセットを作成する。

3. 周辺モデルを用いた解析事例

3.1 糖尿病網膜症の左右眼での発症

今回用いたデータは糖尿病に対する生活指導が糖尿病網膜症の発症を抑制するかを検討したランダム化臨床研究のデータの一部である。共変量として糖尿病罹病期間、割り付け群、ヘモグロビン A1c 値、性別、BMI 値、収縮期血圧が測定され、糖尿病網膜症発症までの生存時間が測定された。データの一部を以下に示す。

OBS	ID	survR	survL	cancel	RIBYOU	GUN	HBA1C	SEX	BMI	SBP
1	1	2146	151	1	7.4	2	7.0	1	25.9	138
2	4	1803	1803	1	4.1	1	7.7	1	21.5	124
3	5	1948	1948	1	2.8	1	8.9	1	18.3	128
4	6	2172	2172	1	14.1	1	7.6	1	23.9	164
5	10	2137	2137	1	10.1	1	7.0	1	22.6	136
6	11	700	700	1	4.1	2	8.0	1	22.9	126

ただし、OBS=オブザベーション数、ID=ID 番号、survR=右眼の発症までの時間、survL=左目の発症までの時間 cancel=打ち切り変数(0:イベント発生、1:打ち切り)、RIBYOU=罹病期間、GUN=割り付け群(1:対照群、2:介入群)、HBA1C=ヘモグロビン A1c 値、SEX=性別(1:男、2:女)、BMI=BMI 値、SBP=収縮期血圧をそれぞれ表す。

糖尿病網膜症の解析事例に用いた SAS プログラムを紹介する。以下にプログラムを示す。WLW モデルを適用するには、データセットに次のような加工をしなければならない。まず、層を表す変数を特定する。次に、1人について右眼のデータセットと左眼のデータセットを作り、各々に右眼の共変量「xxxR」と左眼の共変量「xxxL」を用意する。右眼のデータセットについては「xxxR」にその対象者の共変量の値を入力し、「xxxL」は全て 0 とする。左眼のデータセットについても左右逆にして同様の操作を行う。

```
*****右眼のデータを加工する*****;
```

```
data righteye;
```

```
* 層を指定する変数 右眼=1, 左眼=2;
```

```
type=1;
```

```
* 周辺モデルのためのダミー変数の作成;
```

```
* 右眼の共変量データは残し、左眼の共変量データは全て0にする;
```

```
surv=survR;
```

```
RIBYOUR=RIBYOU; GUNR=GUN; HBA1CR=HBA1C; SEXR=SEX; BMIR=BMI; SBPR=SBP;
```

```
RIBYOUUL=0; GUNL=0; HBA1CL=0; SEXL=0; BMIL=0; SBPL=0;
```

```
*****左眼のデータを加工する*****;
```

```
data lefteye;
```

```
* 層を指定する変数 右眼=1, 左眼=2;
```

```
type=2;
```

```
* 周辺モデルのためのダミー変数の作成;
```

```
* 左眼の共変量データは残し、右眼の共変量データは全て0にする;
```

```
surv=survL;
```

```
RIBYOUL=RIBYOU; GUNL=GUN; HBA1CL=HBA1C; SEXL=SEX; BMIL=BMI; SBPL=SBP;
RIBYOUR=0; GUNR=0; HBA1CR=0; SEXR=0; BMIR=0; SBPR=0;
```

```
*****左右眼のデータをSetして解析用データセットを作成する*****;
```

```
data eye;
  set righteye lefteye;
run;
```

この結果作成されたSASデータセットは次のようになる。

OBS	ID	TYPE	surv	censor	RIBYOUR	RIBYOUL	GUNR	GUNL	HBA1CR	HBA1CL
1	1	1	2146	1	7.4	0	2	0	7.0	0
2	1	2	151	1	0	7.4	0	2	0	7.0
3	4	1	1803	1	4.1	0	1	0	7.7	0
4	4	2	1803	1	0	4.1	0	1	0	7.7
5	5	1	1948	1	2.8	0	1	0	8.9	0
6	5	2	1948	1	0	2.8	0	1	0	8.9

次に、実際に解析を行うプログラムを示す。

```
proc phreg data=eye covsandwich(aggregate) outest=Estpl;
  model surv*censor(1,2)
    =RIBYOUR RIBYOUL BMIR BMIL SBPR SBPL GUNR GUNL SEXR SEXL
    HBA1CR HBA1CL
  /r1; *ハザード比の信頼区間を出力する指定;
  strata type; *層を表す変数で層別する;
  id ID;
  output out=Outpp1 dfbeta=dtR dtL ;
run;
```

対象内相関を考慮して分散にロバスト分散を用いるには、上記のようにPROC PHREG ステートメント内で“covsandwich(aggregate)”と指定し、ID ステートメントで ID を表す変数を指定する。

解析結果は次のアウトプットで得られる。

The PHREG Procedure

Model Information

Data Set WORK.EYE
 Dependent Variable surv
 Censoring Variable censor
 Censoring Value(s) 1
 Ties Handling EXACT

Summary of the Number of Event and Censored Values

Stratum	TYPE	Total	Event	Censored	Percent Censored
1	1	1294	207	1087	84.00
2	2	1300	193	1107	85.15
Total		2594	400	2194	84.58

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	5413.021	5189.714
AIC	5413.021	5213.714
SBC	5413.021	5261.612

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	223.3070	12	<.0001
Score	253.7374	12	<.0001
Modified Score	86.4660	12	<.0001
Wald	168.7898	12	<.0001

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
RIBYOUR	1	0.06000	0.00876	0.984	46.8864	<.0001	1.062	1.044	1.080
RIBYOUCL	1	0.06010	0.00871	0.949	47.6374	<.0001	1.062	1.044	1.080
BMIR	1	0.04702	0.02326	0.981	4.0858	0.0432	1.048	1.001	1.097
BMIL	1	0.02891	0.02365	0.961	1.4939	0.2216	1.029	0.983	1.078
SBPR	1	0.00878	0.00466	1.038	3.5468	0.0597	1.009	1.000	1.018
SBPL	1	0.00882	0.00473	1.024	3.4814	0.0621	1.009	1.000	1.018
GUNR	1	-0.32659	0.14133	0.995	5.3397	0.0208	0.721	0.547	0.952
GUNL	1	-0.22375	0.14797	1.010	2.2867	0.1305	0.800	0.598	1.069
SEXR	1	0.18687	0.14445	1.015	1.6737	0.1958	1.205	0.908	1.600
SEXL	1	0.27351	0.14950	1.013	3.3470	0.0673	1.315	0.981	1.762
HBA1CR	1	0.35458	0.03484	0.931	103.5561	<.0001	1.426	1.331	1.526
HBA1CL	1	0.33148	0.03750	0.935	78.1535	<.0001	1.393	1.294	1.499

“Parameter Estimate”の列にはパラメータ推定値が、“Hazard Ratio”の列にはハザード比が出力されている。いちばん上の行を例に取ると、右眼における罹病期間の効果を表すパラメータ推定値は 0.6000 であり、ハザード比は 1.062 であることがわかる。“Pr > Chisq”の列には帰無仮説「パラメータ推定値=0」に対する検定の p 値が出力されている。右眼における罹病期間の効果は、p 値が 0.0001 より小さく 5%水準で有意に 0 ではない。いちばん右の列の“95% Hazard Ratio Confidence Intervals”にはハザード比の 95%信頼区間が出力されている。

左右眼で平均した共変量の効果を知りたいときには、PROC IML で計算する。計算は以下のプログラムで行った。

```
*****ロバスト分散共分散行列の出力*****;
proc sort data=Outp1; by ID;
proc means data=Outp1 noprint;
  by ID;
  var dtR dtL;
  output out=Outp2 sum=dtR dtL;
proc iml;
  use Outp2;
  read all var{dtR dtL} into x;
  v=x` * x;
  reset noname;
  vname={"RIBYOUR", "RIBYOUL"};
  print, "ロバスト分散共分散行列", ,
    v[colname=vname rowname=vname format=10.5];
  create RCov from v[colname=vname rowname=vname];
  append from v[rowname=vname];
run;

proc iml;
  use Estp1;
  read all var{RIBYOUR RIBYOUL} into RIBYOU;
  b= RIBYOU`;
  use Outp2;
  read all var{dtR dtL} into x;
  v=x` * x;
  nparm= nrow(b);
```

```

se=sqrt(vecdiag(v));
reset noname;
stitle={"Estimate"," Std Error"};
vname={"RIBYOUR","RIBYOUL"};
tmpprt= b || se;
print,tmpprt[colname=stitle rowname=vname format=10.5];
print,"分散共分散行列",,
      v[colname=vname rowname=vname format=10.5];

```

* 左右眼で罹病期間の効果が等しいと仮定して

左右眼で重み付け平均した罹病期間の効果の出力:

```

c= {1 0, 0 1};
cb= c * b;
si= c * v * t(c);
e= j(2,1,1);
isi=inv(si);
h= inv(e` * isi * e) * isi * e;
bl= t(h) * cb;
se= sqrt(t(h) * si * h);
zscore= bl / se;
p= 1-probchi( zscore # zscore, 1);
print , "罹病期間の左右眼平均パラメータ",,
      "Optimal Weights = "h,
      "Estimate = " bl,
      "Standard Error = "se,
      "z-score = "zscore,
      "2-sided p-value = " p[format=5.4];
quit;

```

IML のプログラムによる出力結果は以下の通りである。

```

                Estimate Std Error
RIBYOUR      0.06000    0.00876
RIBYOUL      0.06010    0.00871

      ロバスト分散共分散行列
                RIBYOUR    RIBYOUL
RIBYOUR      0.00008    0.00006
RIBYOUL      0.00006    0.00008

      罹病期間の左右眼平均パラメータ
Optimal Weighs = 0.4877582
                  0.5122418
Estimate = 0.0600495
Standard Error = 0.0081517
z-score = 7.3665071
2-sided p-value = .0000
```

左右眼で罹病期間の効果が等しいと仮定して左右眼で重み付け平均した罹病期間の効果のパラメータ推定値は 0.6005 であった。

3.2 糖尿病合併症

糖尿病合併症についても、糖尿病網膜症、虚血性心疾患、脳梗塞にそれぞれ「type=1, 2 ,3」という変数を与えれば、同様の解析を行うことができる。IMLを用いれば各合併症間で平均した共変量の共通効果を推定することができる。

4. おわりに

今回バージョン8のマニュアルに記載されている方法で多重イベントに対する Cox 回帰モデルを適用した事例を紹介した。WLW モデルはこれらの事例以外にも再発を繰り返す疾患など、多くの応用適用例が考えられ、応用範囲が広い方法であると考えられる。

[参考文献]

1. Cox,D.R.(1972),“Regression Models and Life-Tables (with discussion),” Journal of the Royal Statistical Society, Series B, 34, 187-220.
2. Lin,D.Y.and Wei,L.J.(1989),“The Robust Inference for the Proportional Hazards Model,” Journal of the American Statistical Association, 84,1074-1078.
3. SAS/STAT User's Guide Version8.
4. Wei,L.J.,Lin,D.Y. and Weissfeld,L.(1989),“Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution,”Journal of the American Statistical Association, 84, 1065-1073.

日本SASユーザー会 (SUGI-J)

再発事象に対するモデルを用いた解析方法の検討

○中 牧子, 大橋 靖雄

東京大学大学院医学系研究科健康科学・看護学専攻生物統計学

Statistical models for recurrent events

Makiko Naka, Yasuo Ohashi

Department of Biostatistics, School of Health Sciences and Nursing,

The University of Tokyo

要 旨

再発事象の解析方法として、PHREG プロシジャで解析可能な Cox 回帰を拡張した様々なモデルや、GENMOD プロシジャで解析可能な GEE を用いたポアソン回帰が提案されている。これらの方法を紹介するとともに、多発性硬化症の臨床試験を想定したシミュレーションを試みた。

キーワード： PHREG プロシジャ, GENMOD プロシジャ, Cox 回帰, ポアソン回帰

1. はじめに

臨床試験には、観察期間中に再発が何度も起こる疾患を対象としたものがある。このとき、再発防止や再発遅延に対する薬剤の治療効果は長期的に評価する必要がある。治療効果を推定する際、再発を考慮することで情報量が増加し、関心のある治療効果の推定効率が上がることも期待される。再発を考慮したモデルとして、PHREG プロシジャで解析可能な Cox 回帰を拡張した種々のモデル (AG, PWP, WLW, LWA の各モデル) や、GENMOD プロシジャで解析可能なポアソン回帰モデルが提案されている。まず各モデルを紹介してプログラム例を示し、さらに再発と寛解を繰り返す神経疾患である多発性硬化症の臨床試験を例にとりてシミュレーションした結果を示す。

2. 再発を考慮したモデル

2-1. 記法

時点 t における対象者 i の j 回目のイベント(再発)について以下のように定義する。

$\lambda_{ij}(t)$: ハザード関数

$\lambda_{0j}(t)$: 基準ハザード関数

$Y_{ij}(t)$: 指示関数(リスク集合に含まれる場合には 1、それ以外には 0)

$X_{ij}(t)$: 共変量ベクトル

β_j : 推定すべきパラメータベクトル(治療効果を表すパラメータを含む)

これらが何回目かの再発によらず共通と仮定する場合には、添え字 j は省く。

また、本論文で例示する際のダミーデータ data=myelin を表 1 に示す。変数 ID は個人の識別、TRT は治療を表す変数(0,1 の 2 群)、obsday は観察期間、rec1 は 1 回目の再発の起きた日、rec2 は 2 回目の再発の起きた日、rec3 は 3 回目の再発の起きた日を表す。

表 1: データ例 data=myelin

ID	TRT	obsday	rec1	rec2	rec3
1	0	692	51	185	413
2	1	701	.	.	.
3	1	536	196	.	.

2-2. AG モデル

Andersen, Gill(1982)は、イベント発生過程に非定常ポアソン過程を仮定するモデルを提案した。同一対象者内の複数イベントを独立とみなし、一度イベントを起こした対象者も観察が続いている限りリスク集合に含まれるとするモデルである。対象者 i の時点 t におけるハザード関数は $\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(X_i(t)\beta)$ と表される。治療の全般的な効果に関心のあるときに特に有用なモデルである。

PHREG プロシジャでは、SAS6.10 から MODEL 文で再発事象型のデータを扱うことが可能である。さらに、対象内相関を考慮するために用いるロバスト分散は、SAS8.2 から covsandwich オプション(covs)で指定することで得られるようになった。ここで aggregate オプションを用いることで、各対象者(ID)を単位として集計することができる。AG モデルを適用するためにデータセット data=myelin を表 2 のように加工する。各対象者に対して、obsday を対象者がリスク集合に含まれている半閉区間(t_1, t_2]に分解し、イベントが起きたときは打ち切り変数 status は 1、打ち切りられたとき(観察が終了したとき)は 0 をとるとする。また、何回目の再発であるかは変数 type で表す。加工したデータセット data=ag について、プログラム例を表 3 に示す。

表 2: AG モデル解析用データセット data=ag

ID	TRT	t1	t2	status	type
1	0	0	51	1	1
1	0	51	185	1	2
1	0	185	413	1	3
1	0	413	692	0	4
2	1	0	701	0	1
3	1	0	196	1	1
3	1	196	536	0	2

表 3: AG モデルのプログラム例

```
proc phreg data=ag covs(aggregate);
  model (t1, t2)*status(0)=trt /rl ties=efron;
  id id;
run;
```

2-3. PWP モデル

Prentice, Williams, Peterson(1981)は、 j 回目のイベントに対するリスク集合は、 $(j-1)$ 回目のイベントを起こした対象者に限るとする条件付モデルを提案した。その際、生存時間の取り扱いにより 2 つのモデルが提案されている。イベント発生過程に非定常ポアソン過程を仮定し、時点を全て研究開始からとする total time モデルと、イベント発生過程にセミマルコフ過程を仮定し、前回のイベントから今回のイベントまでの経過時間で考える gap time モデルである。対象者 i の時点 t における j 回目の再発に対するハザード関数は、

$$\text{total time モデル: } \lambda_{ij}(t) = Y_{ij}(t) \lambda_{0j}(t) \exp(X_i(t) \beta_j)$$

$$\text{gap time モデル: } \lambda_{ij}(t) = Y_{ij}(t) \lambda_{0j}(t - t_{j-1}) \exp(X_i(t) \beta_j)$$

(t_{j-1} を $(j-1)$ 回目の再発発生時間とおく)

と表される。治療がどこから効果を発揮するかに関心のあるときに特に有用なモデルである。ただし、本研究ではパラメータを何回目かの再発によらず共通と仮定した。

解析プログラムは表 4 のようになる。データセットは AG モデルで用いたものと同じものが使えるが、gap time モデルに関しては前回のイベントからの時間 (gap time) の計算が必要である。また、複数のイベントについて異なったベースラインを仮定するために、STRATA 文でイベントの種類を指定する必要がある。ここでもロバスト分散を用いた。

表 4:PWP モデルのプログラム例

```

data pwp;
  set ag;
  gap=t2-t1;
run;
*total time modelについて;
proc phreg data=pwp covs(aggregate);
  model t2*status(0)=trt /rl ties=efron;
  strata type;
  id id;
run;
*gap time modelについて;
proc phreg data=pwp covs(aggregate);
  model gap*status(0)=trt /rl ties=efron;
  strata type;
  id id;
run;

```

2-4. WLW モデル

Wei, Lin, Weissfeld(1989)は、各再発を別々のイベントと捉える周辺モデルを提案した。再発発生に関する前提はおいていない。対象者はその集団で最大の再発数だけ設定された各リスク集合に入る。対象者 i の時点 t における j 回目の再発に対するハザード関数は、 $\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp(X_i(t)\beta_j)$ と表される。ここで対象内相関を考慮したロバスト分散を使うことを提案している。

解析データセットを表 5 に、プログラム例を表 6 に示す。ダミーデータセットである data=myelin の場合、最大の再発数が 3 回であったため、層の数は 3 である。まず、治療効果を層ごとに求める。そして、再発間で治療効果が共通とする共通パラメータは、ロバスト分散の逆数で重み付けして IML で求めることができる。

表 5:WLW モデルの解析用データセット data=wlw1

ID	TRT	t2	status	type
1	0	51	1	1
1	0	185	1	2
1	0	413	1	3
2	1	701	0	1
2	1	701	0	2
2	1	701	0	3
3	1	196	1	1
3	1	536	0	2
3	1	536	0	3

表 6:WLW モデルのプログラム例

```

data w1w2;
  set w1w1;
  if type<4;
  k1=trt*(type=1);k2=trt*(type=2);k3=trt*(type=3);
proc phreg data=w1w2 outest=Est1;
  model t2*status(0)=k1-k3 /r1 ties=efron;
  output out=out1 dfbeta=dt1-dt3 /order=data;
  strata type;
  id id;
run;
proc means data=out1 noprint;
  by id;
  var dt1-dt3;
  output out=out2 sum=dt1-dt3;
run;
proc iml;
  use Est1;
  read all var {k1 k2 k3} into trt;
  b= trt`;
  use out2;
  read all var {dt1 dt2 dt3} into x;
  v=x` * x;
  nparm= nrow(b);
  se=sqrt(vecdiag(v));
  reset noname;
  stitle={"Estimate", " Std Error"};
  vname={"k1", "k2", "k3"};
  tmpprt= b || se;
  print,tmpprt[colname=stitle rowname=vname format=10.5];
  print,"Estimated Covariance Matrix",,
      v[colname=vname rowname=vname format=10.5];
  c= {1 0 0 , 0 1 0 , 0 0 1};
  cb= c * b;
  si= c * v * t(c);
  e= j(3,1,1);
  isi=inv(si);
  h= inv(e` * isi * e) * isi * e;
  b1= t(h) * cb;
  se= sqrt(t(h) * si * h);
  zscore= b1 / se;
  p= 1- probchi( zscore # zscore, 1);
  print ,"Estimation of the Common Parameter for Treatment",,
      "Optimal Weights = "h, "Estimate = " b1, "Standard Error = " se,
      "z-score =" zscore, "2-sided p-value = " p[format=5.4];
quit;

```

2-5. LWA モデル

Lee, Wei, Amato(1992)は、各再発を別々のイベントと捉える周辺モデルのうち、基準ハザードが各再発で共通とするモデルを提案した。さらに、ここで対象内相関を考慮したロバスト分散を使うことを提案している。対象者 i の時点 t における j 回目の再発に対するハザード関数は、 $\lambda_{ij}(t) = Y_{ij}(t)\lambda_0(t)\exp(X_i(t)\beta_j)$ と表される。プログラム例を表 7 に示す。

表 7: LWA モデルのプログラム例

```
proc phreg data=wlw1 covs(aggregate);
  model t2*status(0)=trt /rl ties=efron;
  id id;
run;
```

2-6. ポアソン回帰モデル

再発事象の解析方法として、イベント(再発)の起きる時間までに注目した Cox 回帰やその拡張モデルの他に、ある期間内での再発数を数えるモデルを用いることもできる。

ポアソン回帰モデルは、単位時間内における対象者のイベント(再発)生起がポアソン分布に従うとするモデルであり、一般化線型モデルの枠組みでパラメータ推定される。 $\mu(y)$ を再発回数 y の期待値、 $N(y)$ を観察期間、 $r(y)$ を疾患の再発率、 X を共変量ベクトル、 β を治療効果を含む推定すべきパラメータとすると、リンク関数として \log をとり、

$$\log\{r(y)\} = \log\{\mu(y)/N(y)\} = X\beta$$

と表され、 $\log\{\mu(y)\} = X\beta + \log\{N(y)\}$ すなわち $\mu(y) = N(y)\exp(X\beta)$ と表すことができる。ポアソン回帰を用いた解析でも、Cox 回帰による解析と同様に、治療効果の違いをハザード比として推定することができる。

SAS では GENMOD プロシジャで解析でき、MODEL 文のオプションで、ポアソン分布とリンク関数を指定し、さらに観察期間の対数をとったものを OFFSET として指定する。プログラム例を表 8 に示す。

表 8: ポアソン回帰モデルのプログラム例

```
proc genmod data=poisson;
  class trt;
  model recnum=trt /dist=poisson link=log offset=logtime type3;
run;
```

ところが、再発率は全観察期間を通じて必ずしも一定とはいえないため、観察期間を再発率

が十分に一定とみなせる区間に区切り、区間ごとの再発回数を用いる方が妥当である。その際、対象者内の各区間の間に存在する再発率の相関は、対象者を 1 つのクラスターとみなして一般化推定方程式(Generalized Estimating Equations: GEE)を用いて考慮することができる。SAS では GENMOD プロシジャにおいて、REPEATED ステートメントでクラスターを指定することにより解析できる。表 9 に 180 日ごとに区切ったときの解析用データセットを示す。区間の中の再発回数を変数 recnum で表している。さらに、表 10 にプログラム例を示す。ここでは、相関構造として、区間の間に存在する相関は一定とする exchangeable 構造を仮定したため、オプションで type=exch と指定した。

表 9: GEE を用いたポアソン回帰の解析用データセット

ID	TRT	time	recnum
1	0	180	1
1	0	180	1
1	0	180	1
1	0	152	0
2	1	180	0
2	1	180	0
2	1	180	0
2	1	161	0
3	1	180	0
3	1	180	1
3	1	176	0

表 10: GEE を用いたポアソン回帰モデルのプログラム例

```
proc genmod data=geepoisson;
  class id trt;
  model recnum=trt /dist=poisson link=log offset=logtime type3;
  repeated subject=id /type=exch;
run;
```

GENMOD プロシジャ(ポアソン回帰と GEE ポアソン回帰)では、ハザード比やその信頼区間について ESTIMATE ステートメントの exp オプションによって、パラメータ推定値の指数をとることで、ハザード比(再発率比)とみなすことができる。

3. シミュレーション

以上のモデルのうち、多発性硬化症の臨床試験ではどのモデルを用いて解析するのが適当であるかを検討するため、当該疾患の時間的・空間的多発という病態から想定されるモデルで

シミュレーションを行った。それは、対象者が将来病巣になる潜在病巣をいくつか持っているとし、早く悪化して再発とみなされたものから1回目、2回目、・・・の再発が生じるとみなす病態モデルである。

想定される臨床試験の対象者は試験薬群 100 例、プラセボ群 100 例の計 200 例であり、試験期間は 3 年とした。また、プラセボ群に対する試験薬群のハザード比(再発率比)を 1/1.3 と設定した。打ち切りまでの時間(観察期間)や再発までの時間はそれぞれワイブル分布を仮定し、そのパラメータは、過去に行われた当該疾患の複数の臨床試験におけるプラセボ群のデータから得た。

ワイブル分布の生存関数は、 γ を形状パラメータ、 λ を尺度パラメータとしたとき $S(t) = \exp(-\lambda t^\gamma)$ と表される。本研究では SAS の rand 関数によってワイブル分布の乱数を発生させ、call streaminit ルーチンによってシードを指定し、再現性を保証した。rand 関数では、ワイブル分布の確率密度関数を $f(t) = a/b^a \cdot t^{a-1} \cdot \exp\left\{-\left(t/b\right)^a\right\}$ と定義し($a=\gamma$, $b=(1/\lambda)^{1/\gamma}$)、

$$x = \text{rand}('weibull', a, b)$$

によってワイブル分布に従う乱数を発生させることができる。

シミュレーションは 1000 回行い、各データセットに対して、1 回目再発までの時間に対する Cox 回帰、Cox 回帰を拡張した各モデル(AG, PWP, WLW, LWA モデル)、ポアソン回帰、半年ごとに期間を区切って GEE を用いたポアソン回帰を当てはめ、バイアスと MSE を算出した。WLW モデルでは、はじめから共通パラメータを設定して解析した。結果を表 11 に示す。バイアスはパラメータ推定値と真値($\log(1/1.3)=-0.26236$)の差である。1 回目再発までの時間に対する Cox 回帰の結果を基準とみて各モデルを相対評価すると、PWP gap time モデルと LWA モデルでバイアスが小さく、AG モデルや GEE を用いたポアソン回帰において MSE が小さいという結果が得られた。

表 11: シミュレーション結果

	パラメータ推定値	バイアス	MSE
1 回目再発までの時間に対する Cox 回帰	-0.2643	-0.0020	0.0489
AG モデル	-0.2188	0.0436	0.0138
PWP total time モデル	-0.3256	-0.0633	0.0376
PWP gap time モデル	-0.2553	0.0070	0.0167
WLW モデル	-0.4248	-0.1624	0.0760
LWA モデル	-0.2618	0.0005	0.0168
ポアソン回帰	-0.2186	0.0438	0.0162
GEE ポアソン回帰	-0.2167	0.0454	0.0138

4. おわりに

SAS ではプロシジャを用いて、再発事象に関する様々なモデルを用いた解析ができる。モデルの使い分けはケースバイケースだが、イベント発生過程にあまり多くの前提を必要としない GEE を用いたポアソン回帰は、シミュレーションの結果からも有用性が期待される。

5. 参考文献

- Andersen PK, Gill RD. Cox's regression model for counting processes : A large sample study. *Annals of Statistics* 1982;10(4): 1100-20.
- Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972;34: 187-202.
- Lee EW, Wei LJ, Amato DA. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, 237-47. Dordrecht: Kluwer Academic Publishers, 1992.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13-22.
- Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine* 1994;13: 2233-47.
- Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981;68(2): 373-9.
- Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using The SAS System* 2nd ed. Cary: SAS Institute Inc., 2000.
- Therneau TM, Grambsch PM. *Modeling Survival Data : Extending the Cox Model*. Springer-Verlag New York, Inc., 2000.
- Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989;84:1065-73.

日本SASユーザー会 (SUG I - J)

MIXED プロシジャを用いた 線形混合効果モデルの交互作用の指定方法

○寒水 孝司 *1 菅波秀規 *1*2

*1 東京理科大学大学院工学研究科 *2 興和株式会社臨床解析部

Consideration for Coding the Interaction of
Linear Mixed Effects Models Using MIXED Procedure

Takashi Sozu*1 Hideki Suganami*1*2

*1 Graduate School of Engineering, Tokyo University of Science

*2 Biostatistics and Data Management Dept., Kowa Co.Ltd

要 旨

MIXED プロシジャを用いて、交互作用を含む線形混合効果モデルを適用し、主効果の不均一性を評価することを考える。本稿では交互作用の定義の方法が異なる2つのモデルを取り上げ、得られる解析結果の違いについて考察する。特に、臨床試験において試験治療効果の施設における不均一性を評価する方法をいくつか想定し、どちらのモデルを用いるのが適切であるかを考察する。

キーワード：MIXED プロシジャ，多施設共同治験，線形混合効果モデル，交互作用

1 はじめに

臨床試験では複数の施設（病院，医院など）で試験治療の有用性を評価することが多い（以下，この試験を多施設共同治験と呼ぶ）。多施設共同治験では，施設によって試験治療と対照治療の効果の差が異なることがあり，得られた結果の一般化可能性を評価するために試験治療効果の施設における不均一性を評価する必要がある。このことに関して，1998年に発効された「臨床試験のための統計的原則」に，次のような記述がある [1]。

“施設当たりの被験者数が不均一性を評価しうる規模の試験で，試験治療の肯定的な効果が判明した場合，結論の一般化可能性に影響する可能性があるため，通常は施設間における試験治療効果の不均一性を探索すべきである。著しい不均一性は，個々の施設の結果を図示すること又は試験治療と施設間の交互作用の有意性検定などの解析手法によることでも確認される場合がある。～（中略）～これまで，多施設共同治験に関する議論は，固定効果モデルを用いることを前提としてきた。混合モデルも試験治療効果の不均一性を探索するために利用できる。混合モデルでは，施設及び試験治療と施設の交互作用を変量効果として扱っており，特に施設数が多い場合に用いることが適切である。”

日本の臨床試験は、多くの施設で実施されることが多い。そのため、MIXED プロシジャを用いて線形混合効果モデルによる推測を行う状況は少なくないと考えられる。

本稿では、臨床試験において試験治療効果の施設における不均一性を評価する際に、どのように MIXED プロシジャを指定すればよいかを考察する。具体的には、交互作用の定義の方法が異なる 2 つのモデルを取り上げ、試験治療効果の施設における不均一性を評価する方法に応じて、どちらのモデルを用いるのが適切であるかを考察する。ここで、2 つのモデルのうち 1 つは、一般的な教科書等に記載されていることが多く、単純な 2 元配置実験を想定したときに用いるモデルであり [2][4]、もう 1 つは、著者らが使用すべきと主張するモデルである。

2 比較する 2 つの方法

2.1 モデルと記号法

並行 2 群比較試験を想定する。すなわち、 J 施設の各々で $2K$ 人の被験者が K 人ずつの 2 群に分けられ、第 1 群では試験治療、第 2 群では対照治療が施されるとする。ここで、総被験者数は $N = 2JK$ とする。このような状況において、施設 j で試験治療 ($i = 1$) あるいは対照治療 ($i = 2$) を受けた k 番目の被験者の応答を y_{ijk} とし、 y_{ijk} に次式の治療と施設の交互作用を含む線形混合効果モデルを想定し、試験治療効果の施設における不均一性を評価することを考える。(今回取り上げた 2 つのモデルはいずれも次式で表現される。)

$$y_{ijk} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \epsilon_{ijk} \quad (1)$$

μ	...	総平均	
β_i	...	治療 i の主効果	$i = 1, 2$
γ_j	...	施設 j の変量効果	$j = 1, 2, \dots, J$
$(\beta\gamma)_{ij}$...	治療 i と施設 j の交互作用	$k = 1, 2, \dots, K$
ϵ_{ijk}	...	誤差	

説明のために、式 (1) の線形混合効果モデルを次のように行列表現する。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (2)$$

各記号の定義を次に示す。

\mathbf{y}	:	結果変数ベクトル, $\text{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$
\mathbf{X}	:	固定効果に対する計画行列
$\boldsymbol{\beta}$:	固定効果ベクトル
\mathbf{Z}	:	変量効果に対する計画行列
$\boldsymbol{\gamma}$:	変量効果ベクトル, $E(\boldsymbol{\gamma}) = \mathbf{0}$, $\text{Var}(\boldsymbol{\gamma}) = \mathbf{G}$
$\boldsymbol{\epsilon}$:	誤差ベクトル, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\text{Var}(\boldsymbol{\epsilon}) = \mathbf{R}$, $\text{cov}(\boldsymbol{\gamma}, \boldsymbol{\epsilon}') = \mathbf{0}$

2.2 2 つのモデルの違い

上記の式 (1) に対応する 2 つのモデルを考える。いずれも総平均、治療の主効果、および施設の変量効果の定義の方法は同じであるが、交互作用の定義の方法は異なる。モデル 1 では、2 つの治療に

対してそれぞれ交互作用を定義し、モデル2では、試験治療に対してのみ交互作用を定義する。すなわち、モデル2は、モデル1の交互作用のパラメータに $(\beta\gamma)_{2j} = 0, \forall j$ という制約条件を与えたモデルに相当する。

これらのモデルを MIXED プロシジャで指定するには、それぞれ次のように指定する。ただし、TREAT は治療を表す名義変数（試験治療を 1, 対照治療を 0 とする）、CENTRE は施設を表す名義変数、TR は TREAT とは別に治療を表す連続変数（TREAT と同様に、試験治療を 1, 対照治療を 0 とする）である。

モデル 1

```
proc mixed;
  class TREAT CENTRE;
  model y = TREAT;
  random CENTRE TREAT*CENTRE;
run;
```

モデル 2

```
proc mixed;
  class TREAT CENTRE;
  model y = TREAT;
  random CENTER TR*CENTRE;
run;
```

MIXED プロシジャでは、2 つモデルの違いが random ステートメントの治療を表す変数に現れる。モデル1では、random ステートメントの治療を表す変数を class ステートメントに指定している名義変数（TREAT）とし、モデル2では、これを別の連続変数（TR）とする。

3 試験治療効果の不均一性の評価方法とモデルの妥当性

3.1 試験治療効果の不均一性の評価方法

線形混合効果モデルを適用し、試験治療効果の施設における不均一性を評価するには、いくつかの方法がある。ここでは、次の3つの方法を取り上げ、それぞれ2つのモデルの妥当性を評価する。

- 分散の比較 … 変数効果の分散の推定値を比較する方法
- 試験治療効果の分布の図示 … 試験治療効果に関する分布を施設ごとに図示する方法
- 予測値の図示 … モデルの予測値を治療および施設ごとに図示する方法

3.2 モデルの妥当性

3.2.1 分散の比較

変数効果の分散の推定値をもとに、試験治療効果の施設における不均一性を評価するには、モデルの変数効果の分散のもつ意味を吟味する必要がある。そこで、各モデルの変数効果の分散の解釈の仕方とその根拠について説明し、2つのモデルの違いを明らかにする。具体的な評価方法は次節で例示する。

モデル1では、2つの治療 ($i = 1, 2$) に対してそれぞれ交互作用を定義するので、交互作用の変数効果の分散は、いずれかの治療を行うことの施設間の“ばらつき”を表し、施設の変数効果の分散は、施設のみの効果の“ばらつき”を表す。一方、モデル2では、試験治療 ($i = 1$) に対してのみ交互作用を定義するので（すなわち、 $(\beta\gamma)_{2j} = 0, \forall j$ ）、交互作用の変数効果の分散は、試験治療効果の施設間

の“ばらつき”を表し、施設の変量効果の分散は、対照治療（ベースライン効果）の施設間の“ばらつき”を表す。分散に対する解釈と記号の定義を表1にまとめた。

表 1: 変量効果の分散の解釈と記号の定義

	施設		交互作用	
モデル1	$\sigma_{\gamma 1}^2$	施設のみの効果のばらつき	$\sigma_{\beta\gamma 1}^2$	いずれかの治療を行うこと の施設間のばらつき
モデル2	$\sigma_{\gamma 2}^2$	対照治療（ベースライン効果）の施設間のばらつき	$\sigma_{\beta\gamma 2}^2$	試験治療効果の施設間のばらつき

このように整理すると、2つのモデルにおける変量効果の分散のもつ意味はまったく異なることがわかる。ここで、分散の比較に基づく試験治療効果の施設における不均一性の評価という目的と、2つのモデルの解釈を対比させると、モデル2を用いるのが適切であることがわかる。それは、試験治療効果の施設における不均一性を評価する際に、興味のあるばらつきは、試験治療効果（すなわち治療効果の差）の施設間のばらつきであり、これを対照治療（ベースライン効果）の施設間のばらつきと比較することに意味があるからである。逆に、モデル1では、このような目的に合った分散を評価することができない。（しかし、いずれかの治療を行うことの施設間のばらつきに興味がある場合は、必ずしも適切でないというわけではない。）

3.2.2 試験治療効果の分布の図示

試験治療効果の分布をもとに、試験治療効果の施設における不均一性を評価するには、治療の主効果 β_i の大きさを考慮して、各モデルの変量効果 $(\beta\gamma)_{ij}$ のもつ意味を吟味する必要がある。具体的な評価方法は次節で例示する。

ここで興味のあるパラメータは、試験治療効果の（施設における）違いを表すパラメータである。したがって、モデル1では、各治療の交互作用の変量効果の差を考えればよい。一方、モデル2では、交互作用の変量効果は試験治療効果の違いとして解釈できるので、交互作用をそのまま考えればよい。しかし、モデル1では、施設の変量効果を治療とは別に想定するため、交互作用の変量効果の差をもとにした試験治療効果の違いの検討に、施設の変量効果が考慮されない。これに対して、モデル2では、施設の変量効果は対照治療の交互作用として解釈され、試験治療効果の違いの検討に考慮される。すなわち、2つのモデルで想定する変量効果が異なることが、試験治療効果の違いを表すパラメータに影響する。モデル2による試験治療効果の分布の検討が、試験治療効果の違いを直接評価することは明らかであるが、一方で、モデル1による検討が不適切であるとするには、さらなる検討が必要と思われる。

3.2.3 予測値の図示

モデルの予測値をもとに、試験治療効果の施設における不均一性を評価するには、モデルの予測値の特性を吟味する必要がある。具体的な評価方法は次節で例示する。

モデルの予測値の特性として重要なのが、各モデルの予測値の分散である。ここで、各モデルにお

ける群ごとの応答変数の分散は次式で与えられる。ただし、 $\text{Var}(\epsilon_{ijk}) = \sigma^2$ とする。

$$\text{モデル 1} \quad \text{Var}(y_{1jk}) = \text{Var}(y_{2jk}) = \sigma_{\gamma_1}^2 + \sigma_{\beta_{\gamma_1}}^2 + \sigma^2$$

$$\text{モデル 2} \quad \text{Var}(y_{1jk}) = \sigma_{\gamma_2}^2 + \sigma_{\beta_{\gamma_2}}^2 + \sigma^2$$

$$\text{Var}(y_{2jk}) = \sigma_{\gamma_2}^2 + \sigma^2$$

このように、モデル1では、各群での応答変数の分散が対等に扱われるが、モデル2では、そうではない。したがって、モデルの予測値を図示して試験治療効果の施設における不均一性を評価する場合は、モデル1を用いるのが適切であることがわかる。

3.3 まとめ

これまでの結果をまとめると、次のようになる。

表 2: 試験治療効果の施設における不均一性を評価するモデル

評価方法	モデル 1	モデル 2
分散の比較	×	○
試験治療効果の分布の図示	△	○
予測値の図示	○	×

このように、試験治療効果の施設における不均一性を評価する場合には、目的（方法）に応じて2つのモデルを使い分ける必要がある。ただし、モデル2は一般的な教科書等に記載されている方法とは異なるため、さらなる吟味が必要であると考えている。

4 MIXED プロシジャの適用例

4.1 状況設定

ここでは、簡単な数値例をもとに、MIXED プロシジャを用いて2つのモデルを適用した場合の解析結果について考察し、いくつかの注意点を整理する。

次のような条件のもとで仮想データを発生させ、多施設共同治験のデータの例とした（ただし、このようなデータに対して、線形混合効果モデルを使うべきであると主張しているわけではない）。簡単のために、施設数が $J = 5$ 、1施設1群あたり被験者数が $K = 10$ 、すなわち総被験者数が $N = 2JK = 100$ の場合を想定する。施設の変量効果 γ_j 、交互作用の変量効果 $(\beta\gamma)_{ij}$ 、および誤差 ϵ_{ijk} はそれぞれ互いに独立に正規分布に従うものとする。

この条件はいずれの治療に対しても交互作用の変量効果を想定しているため、モデル1に準拠している。しかし、実際には真の状態を規定することは困難であり、この条件が本質的な問題になるとは

考えていない。

$$\begin{aligned}\mu &= 0 \\ \beta_1 &= 2, \beta_2 = 0 \\ \gamma_j &\sim N(0, \sqrt{2}) \\ (\beta\gamma)_{ij} &\sim N(0, \sqrt{2}) \\ \epsilon_{ijk} &\sim N(0, 2)\end{aligned}$$

4.2 MIXED プロシジャの出力

乱数を発生せさせて得られたデータに MIXED プロシジャを適用し、特記すべき結果についてまとめた。ただし、近似自由度の計算には Satterthwaite 法を用いた。MIXED プロシジャで、近似自由度の計算に Satterthwaite 法を指定するには、model ステートメントで “ddfm = satterth” と記述すればよい。

4.2.1 変量効果の計画行列

モデル 1 の変量効果ベクトル γ は、 $J + 2J = 3J = 15$ 行の列ベクトルになり、モデル 2 では、これが $J + J = 2J = 10$ 行の列ベクトルになる。これに対応して、モデル 1 の変量効果に対する計画行列 \mathbf{Z} は、 $[N \times 3J] = [100 \times 3]$ の行列になり、モデル 2 では、これが $[N \times 2J] = [100 \times 2]$ の行列になる。モデル 2 では、 $(\beta\gamma)_{2j} = 0, \forall j$ とするため、モデル 1 よりも変量効果に対する計画行列の列数が少なくなる。MIXED プロシジャでは、出力の “Dimension” の部分に計画行列の列数が表示される。

4.2.2 固定効果の推定値

固定効果の推定値に関する出力を表 3 に示した。MIXED プロシジャで固定効果に関する出力を表示させるには、model ステートメントで “solution (または s)” オプションをつければよい。

表 3: 固定効果の推定値に関する出力

モデル	要因	推定値	標準誤差	自由度	t 値	P 値
モデル 1	μ	0.93	1.854	5.24	0.50	0.638
	β_1	3.12	1.374	4	2.27	0.086
モデル 2	μ	0.93	1.616	4.19	0.57	0.596
	β_1	3.12	1.411	4.19	2.21	0.089

今回の数値例では 1 施設 1 群あたりの被験者数が等しい (すなわちバランスがとれた) データであるため、2 つのモデルで固定効果の推定値は一致する。これは次の数学的根拠に基づいている。

線形混合効果モデル (式 (2)) では、 $\mathbf{X}\beta$ の一般化最小二乗推定量 (Generalized Least Squares Estimator: GLSE) は次式で与えられ、最良線形不偏推定量 (Best Linear Unbiased Estimator: BLUE) となる。

$$\text{GLSE}(\mathbf{X}\beta) = \mathbf{X}\beta^0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \text{BLUE}(\mathbf{X}\beta) \quad (3)$$

ただし、一般に \mathbf{V} は未知であるので、 \mathbf{V} の推定量 $\hat{\mathbf{V}}$ を代入することで推定値を得る。ここで、バランスのとれたデータでは、次の関係式が成立し、いずれも最良線形不偏推定量となる [7]。ただし、 $\text{OLSE}(\mathbf{X}\beta)$ は通常の最小二乗推定量 (Ordinary Least Squares Estimator : OLSE)、 $\text{MLE}(\mathbf{X}\beta)$ は最尤推定量 (Maximum Likelihood Estimator : MLE) である。

$$\text{OLSE}(\mathbf{X}\beta) = \text{GLSE}(\mathbf{X}\beta) = \text{MLE}(\mathbf{X}\beta) = \text{BLUE}(\mathbf{X}\beta) \quad (4)$$

ここで、

$$\text{OLSE}(\mathbf{X}\beta) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K y_{ijk} = \bar{y}_{i..} \quad (5)$$

である。2つのモデルで \mathbf{V} の推定値 $\hat{\mathbf{V}}$ は異なり (後述)、 $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$ も異なるが、バランスがとれたデータであれば式 (4) の関係式から、 $\text{GLSE}(\mathbf{X}\beta) = \text{OLSE}(\mathbf{X}\beta)$ となる。したがって、 $\text{OLSE}(\mathbf{X}\beta)$ の式 (式 (5)) に \mathbf{V} が含まれないので、2つのモデルで固定効果の推定値は一致する。

一方、1施設1群あたりの被験者数が異なる場合、すなわちアンバランスなデータでは、例えば、 $\mathbf{VX} = \mathbf{XF}$ を満たす行列 \mathbf{F} が存在すれば、 $\text{GLSE}(\mathbf{X}\beta) = \text{OLSE}(\mathbf{X}\beta)$ となるが、この条件が成立するのは稀である [7]。

2つのモデルで、推定値の標準誤差およびその自由度は異なり、 t 値、 P 値も異なるが、試験治療効果の施設における不均一性を評価するには、本質的な問題ではない。それは、臨床試験のための統計的原則 [1]、および寒水 他 [3] において、治療の主効果は治療と施設の交互作用を含まないモデルを用いて調べることが推奨されており、その場合には2つモデルが同じになるからである。

4.2.3 変量効果の推定値

変量効果の推定値に関する出力を表 4 に示した。ただし、検定統計量に関する出力は省略した。MIXED プロシジャで変量効果に関する出力を表示させるには、random ステートメントで “solution (または s)” オプションをつければよい。モデル1とモデル2では、想定する変量効果ベクトル γ が異なるため、変量効果の推定値は一致せず、結果として $\hat{\mathbf{V}}$ も異なる。それに応じて標準誤差およびその自由度は異なる。

4.3 試験治療効果の不均一性の評価

前節では、試験治療効果の不均一性を評価するには、目的 (方法) に応じて、モデルを使い分ける必要があることを示した。ここでは数値例について、適切なモデルを用いて解析を行った結果を示す。

4.3.1 分散の比較

モデル2を用いて解析する。試験治療効果の施設間の分散は $\sigma_{\beta\gamma_2}^2 = 6.3$ 、対照治療 (ベースライン効果) の施設間の分散は $\sigma_{\gamma}^2 = 11.2$ 、誤差分散は $\sigma^2 = 18.2$ である。一般に、試験治療効果の施設における不均一性を評価するには、 $\sigma_{\beta\gamma_2}^2$ と $\sigma_{\beta_2}^2$ の大きさを相対的に評価すればよい。MIXED プロシジャでは、出力の “Covariance Parameter Estimates” の部分に変量効果の分散および誤差分散の推定値が表示される。

表 4: 変量効果の推定値に関する出力

モデル 1				モデル 2			
要因	推定値	標準誤差	自由度	要因	推定値	標準誤差	自由度
γ_1	4.44	2.020	5.88	γ_1	4.95	1.816	6.03
γ_2	-1.80	2.020	5.88	γ_2	-1.16	1.816	6.03
γ_3	-3.83	2.020	5.88	γ_3	-3.22	1.816	6.03
γ_4	-0.75	2.020	5.88	γ_4	-1.58	1.816	6.03
γ_5	1.95	2.020	5.88	γ_5	1.00	1.816	6.03
$(\beta\gamma)_{11}$	0.19	1.440	2.27	$(\beta\gamma)_{11}$	-0.15	1.740	4.39
$(\beta\gamma)_{12}$	-1.01	1.440	2.27	$(\beta\gamma)_{12}$	-1.78	1.740	4.39
$(\beta\gamma)_{13}$	-1.32	1.440	2.27	$(\beta\gamma)_{13}$	-2.15	1.740	4.39
$(\beta\gamma)_{14}$	0.77	1.440	2.27	$(\beta\gamma)_{14}$	1.61	1.740	4.39
$(\beta\gamma)_{15}$	1.37	1.440	2.27	$(\beta\gamma)_{15}$	2.47	1.740	4.39
$(\beta\gamma)_{21}$	0.84	1.440	2.27				
$(\beta\gamma)_{22}$	0.59	1.440	2.27				
$(\beta\gamma)_{23}$	0.43	1.440	2.27				
$(\beta\gamma)_{24}$	-0.95	1.440	2.27				
$(\beta\gamma)_{25}$	-0.92	1.440	2.27				

モデル 1 では、 $\sigma_{\beta\gamma_1}^2 = 2.9$ 、 $\sigma_{\beta_1}^2 = 12.5$ となるが、これらを比較しても試験治療効果の施設における不均一性を評価することにならない。

4.3.2 試験治療効果の分布の図示

モデル 1 およびモデル 2 を用いて解析する。モデル 1 では、 $\beta_1 + (\beta\gamma)_{1j} - (\beta\gamma)_{2j}$ の推定値を、モデル 2 では、 $\beta_1 + (\beta\gamma)_{1j}$ の推定値を、それぞれ 95% 信頼区間とともに図示すると図 1 が得られる。このように図示することで、試験治療効果の不均一性を視覚的に評価でき、いずれのモデルからも同様の結論が得られる。2 つのモデルによる結果は数値的に類似しているが、モデル 1 による検討が妥当であるかは議論の余地がある。MIXED プロシジャで、上記の推定値とその 95% 信頼区間を計算するには “Estimate” ステートメントを利用すればよい。(変量効果の推定値の 95% 信頼区間のみを出力するには、random ステートメントで “cl” オプションをつければよい。)

4.3.3 予測値の図示

モデル 1 を用いて解析する。モデルの予測値を 95% 信頼区間とともに図示すると図 2 が得られる。このように図示することで、試験治療効果の不均一性を視覚的に評価でき、試験治療効果の分布を図示した場合と同様の結論が得られる。MIXED プロシジャで、モデルの予測値とその 95% 信頼区間を出力するには、model ステートメントで “outpred (または outp) = dataset” オプションをつければよい。

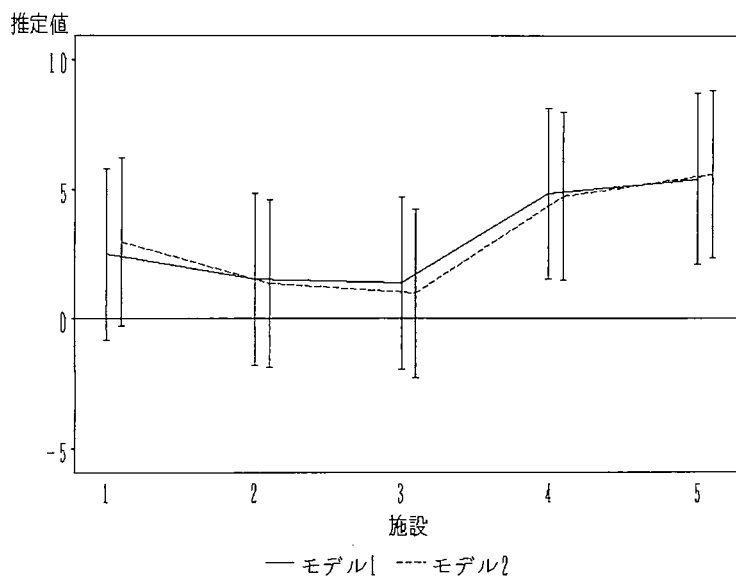


図 1: 試験治療効果の分布

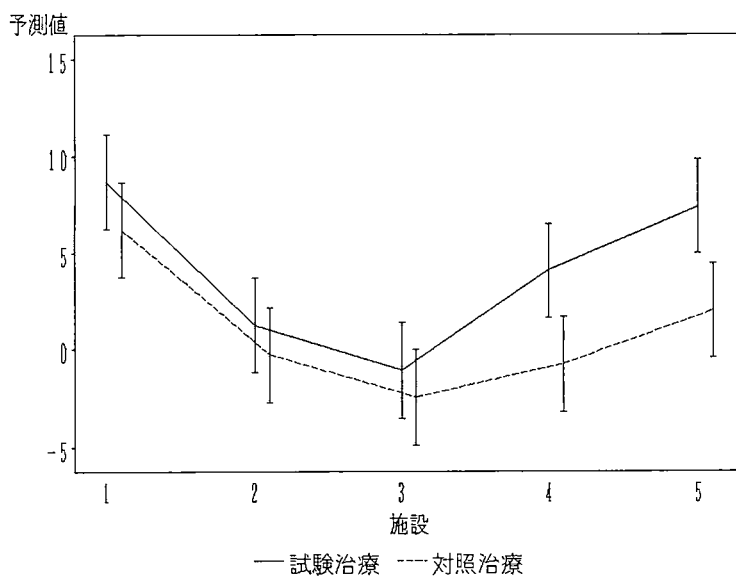


図 2: 各施設ごとのモデルの予測値

5 まとめ

本稿では、臨床試験において試験治療効果の施設における不均一性を評価する3つの方法を示し、どのように線形混合効果モデルを適用するのが適切であるかを考察した。取り上げた2つのモデルは、一見同じようなモデルに見えるが、パラメータのもつ意味が異なるので、目的に応じて使い分ける必要がある。特に、モデル1の交互作用が試験治療効果（治療効果の差）の不均一性を表していないことには注意が必要である。

今回取り上げたモデル2は、一般的な教科書等に記載されている方法とは異なるため、著者らは、このモデルに対する数理的な研究を行う必要があると考えている。

謝辞 本発表のきっかけは、東京理科大学工学研究科経営学専攻の医薬統計コースで東京大学の松山裕 助教授による「混合モデルとベイズ流解析法」の講義を受けたことである。本研究に対する動機付けと貴重なご助言をして頂いた松山裕 助教授に心より感謝致します。

参考文献

- [1] 厚生省医薬安全局審査管理課長 (1998) : 「臨床試験のための統計的原則」について (平成 10 年 11 月 30 日医薬審第 1047 号) . <http://www.nihs.go.jp/dig/ich/ichindex.htm>.
- [2] 菅波秀規, 吉村功 (2000) 混合効果モデルの実用化. 計量生物セミナー資料
- [3] 寒水孝司, 大森崇, 吉村功 (2001) 「臨床試験のための統計的原則」における交互作用の扱い方についての考察. 応用統計学 **30** No.1, 2001 1-18
- [4] Latour, D and Littell, R. (1996) Advanced General Linear Models with an Emphasis on Mixed Models Course Notes. SAS Institute Inc.
- [5] McCulloch, C. E., Searle, S. R. (2001) Generalized, Linear, and Mixed Models. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.
- [6] SAS Institute Inc. (2001) SAS/STAT User's Guide, Version 8.
- [7] Searle, S. R., Casella, G. and McCulloch, C. E. (1992) Variance Components, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.
- [8] Zyskind, G. (1967) On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Stat.* **38**, 1092-1109
- [9] Zyskind, G. and Martin, F. B. (1969) On best linear estimation and general Gauss-Markoff theorem in linear models with arbitrary non-negative structure. *SIAM J. Appl. Math.* **17**, 1190-1202

要因配置実験の効果成分の表示から生じる不定性*

○柴山忠雄
(前所属・名古屋市工業研究所)

Indeterminacy of effect components of factorial experiments
arising from the ways of expression*

Tadao Shibayama
(Retired: Nagoya Municipal Industrial Research Institute)

要 旨

効果成分の「加法性」のみを仮定しても効果要素は一通りには定まらず、実務的には、線形制約式の問題または一般逆行列の問題を無視できない。SAS/STATソフトウェア GLMプロシジャおよび付属文書が高度に完成された解決策を与えているが、変数、効果成分および効果要素が数多く、表示・定義もさまざま、個別の背景もあるから、活用するには広大な視野を展望し細部を熟視し周到な考慮をはらう必要がある。

単一要因系の直積として組み立てられる組み合わせ完全配置の上での応答関数の効果成分への展開は一般の複数要因系の解析の基礎となり、その不定性の吟味は各種の解析の手順の意味をわかりやすくする。

その結果、SAS/STATソフトウェア GLMプロシジャに組み込まれている推定可能性・平方和縮減・平方和分離・掃き出し法などの一般逆行列による数理も、また、通常の線形制約式による数理も見やすくなる。

キーワード： SAS/STATソフトウェア GLMプロシジャ 一般逆行列 (SAS)平方和 SSI…IV 線形制約式

●問題の背景

組み合わせ完全配置の上で処理ごとに確定する応答を応答方程式に基づいて効果成分の各種、すなわち、一般平均、主効果、および、2要因以上の複数要因交互作用に分解する方法は応答の性質を明らかにするために広く用いられている。しかし、効果成分の「加法性」のみを仮定しても、応答方程式に含まれる効果要素の総数が応答の値の総数よりも多いために、効果成分の要素（効果要素）は一通りには定まらない。

実務的には、応答の性質を明らかにするために効果要素を適当な方法で確定させることも必要な場合があり、そのときには、線形制約式の問題または一般逆行列の問題を無視できない。これを解決する試みの成果が高度に完成された形でSAS/STATソフトウェア GLMプロシジャおよび付属文書に集約されている。

ただし、その成果を活用するには、変数、効果成分および効果要素が数多く、表示・定義もさまざまであり、個別の背景もあるから、広大な視野を展望し細部を熟視し周到な考慮をはらう必要がある。

●解析的・代数的な立場 — 数理統計学の視点からの留意点

統計科学の役割は、実験事実・経験事実から、a) 集団、b) 集団内の変動、および、c) 事実の要約、この3つの視点で、人間活動に役立つ実用的に意味のある知識をとり出すことである (Fisher, 1節, 1925)。しかし、人間は今日も明日も安定している明白な確定の事実にはささえられて生きており、集団も変動も確定の事

*English handout available.

実に基づいて把握する。確定の事実の多くが定数・関数として要約され解析的・代数的な法則として表わされる。集団・変動を取り扱うためにも、定数・関数を取り扱う手段を整理しておくことが求められる。

最も重要なのは応答の値が複数個の要因の各々の設定値の組み合わせの関数（応答関数）となる場合であり、実験は要因の各々の設定値の組み合わせに対して応答関数の値を具体的に定めるために行なわれる。

●各種の構造模型・各種の効果成分 — 実質科学との接点

応答の効果成分はさまざまな形式で定義され、効果成分を加え合わせて応答とするさまざまな形式の構造模型が組み立てられ、さらに、応答から効果成分を定めるためのさまざまな実験配置が考案されている。

SAS/STATソフトウェア GLMプロシジャでは、加法性にしたがう構造模型に基づいて応答を組成する各種の効果成分として、0)定数、1)離散水準値を設定値とする要因（離散要因）の主効果成分、および、2)複数の離散要因の直積組み合わせ構造としての交互作用成分、のほか、3)複数の離散要因の階層枝分かれ構造としての階層分岐成分、などが用意されている（SAS/STAT® User's Guide, Vn 6, 4th edn, p.895-897）。

模型は実在の真実の姿を探索するための手段に過ぎないから、できるだけ単純な形式ですどくときずまされていることが好ましい。その意味では、組み合わせ完全配置の上の応答の効果成分の取り扱い、不定性の問題を中心として、今よりも、わかりやすく整理されてよい。また、枝分かれ配置の上の応答の効果成分の扱いは、組み合わせ完全配置の上のものとは全く別のものとして、やはり、整理されてよい。

●完全配置・省略配置・重複配置・変形配置・欠落配置・欠測配置

組み合わせ完全配置の上での効果成分のいくつかが恒等的に 0である場合には、各々の要因の設定水準値の組み合わせ（処理）の一部を省略した「省略配置」を用いて、ほかの効果成分を求めることができる。

処理のいくつかを重複して実施する「重複配置」を用いると、揺動の大きさを定量することができる。

処理の省略と重複とを不規則に施した「変形配置」も用いられる。計画した配置の処理の一つまたはいくつかについて、本来、測定が実行できない永久の原因があり、その処理の測定値が実現できない場合の配置は「欠落配置」となる。これは偶然の不都合で測定が実行できなかった「欠測配置」とは区別される。

SAS/STATソフトウェア解説文書の「4種類の推定可能関数」の章には欠落配置または欠測配置の実例が示されているが（SAS/STAT® User's Guide, Vn 6, 4th edn, p.120-124）、欠落と欠測との区別は触れられていない。応答から加法性に基づいて効果要素を求める Gauss-Jordan-Doolittle の前進消去・後退代入の計算（Goodnight 1978）はその区別をせず実行できる。組み合わせ効果要素と枝分かれ効果要素とに加法性を仮定して両方を同じ構造模型の中に取り入れる姿勢もこの事実によるところが大きいと思われる。

●加法性と分離性とに基づく（正準）効果成分の定義 — 正準展開・正準制約式

加法性のみに基づく一般線形模型の効果成分の定義を用いると、さまざまな問題を効果要素の連立一次方程式の系で表わして共通の解法で取り扱うことができる。しかし、通常は効果要素の数が応答の数より多いために解には不定性が生じる。また、組み合わせ完全配置の上で確定する応答関数については、応答関数平方和は効果成分平方和の総和と一致しない。加法性に基づく定義に応答関数平方和の効果成分平方和への分離性を補足すると、これを一致させることができる。また、各々の効果要素を確定させることができる。

この結果は実質科学の側からの効果要素の解釈および一般の効果要素の不定性の吟味に便利である。

組み合わせ完全配置の上の応答関数の（正準）効果成分を加法性と分離性とに基づいて定義する。

加法性： 応答関数 $yy(a, b, \dots, k)$ がつぎの式(1)の示すように効果成分の和に分解されること。

$$yy(a, b, \dots, k) = yy:M+yy:A(a)+\dots+yy:K(k)+yy:AB(a, b)+\dots+yy:HK(h, k) \\ +yy:ABC(a, b, c)+\dots+yy:UHK(u, h, k)+\dots+yy:ABC\dots UHK(a, b, c, \dots, u, h, k) \quad (1)$$

ここで、独立変数 a, b, \dots, k は要因 A, B, \dots, K の水準値の一組であり、従属変数 yy は応答の値である。右辺の項 $yy:M$ は定数（一般平均）であり、項 $yy:X(x)$ は単一の要因 X の水準値 x のみを独立変数とする関数（主効果）であり、項 $yy:X_1X_2(x_1, x_2)$ は要因 X_1 の水準値 x_1 と要因 X_2 の水準値 x_2 とを独立変数とする関数（2要因交互作用）であり、3要因以上の各項がつづき、最後は全部の要因を含む交互作用の項となる。

分離性： 式(1)の左辺の応答関数 $yy(a, b, \dots, k)$ のこの配置の上での平方和が、つぎの式の示すよう

に、式(1)の右辺の効果成分の各々のこの配置の上での平方和の和に分解されること。 「 (2)

$$\Sigma yy(a, b, \dots, k)^2 = \Sigma yy:M^2 + \Sigma yy:A(a)^2 + \dots + \Sigma yy:K(k)^2 + \Sigma yy:AB(a, b)^2 + \dots + \Sigma yy:HK(h, k)^2$$

$$+ \Sigma yy:ABC(a, b, c)^2 + \dots + \Sigma yy:UHK(u, h, k)^2 + \dots + \Sigma yy:ABC \dots UHK(a, b, c, \dots, u, h, k)^2$$

和記号 Σ は配置の処理(a, b, ..., k)の全部の上でとる。分離性はつぎの「正準」制約式から導かれる。

$$\text{制約式: } \Sigma @x\#yy:X(x)=0, \quad \Sigma @x1\#yy:X1X2(x1, x2)=0, \quad \Sigma @x2\#yy:X1X2(x1, x2)=0, \quad \text{「 (2a)}$$

$$\Sigma @x1\#yy:X1X2X3(x1, x2, x3)=0, \quad \Sigma @x2\#yy:X1X2X3(x1, x2, x3)=0, \quad \Sigma @x3\#yy:X1X2X3(x1, x2, x3)=0,$$

... 」 この式で、記号@は添え字の先頭を表わし、記号#は添え字の末尾を表わす。

整数x, 整数x1などは要因X, 要因X1などの水準値をそれぞれ表わす。和記号 $\Sigma @x\#$, 和記号 $\Sigma @x1\#$ などはそれぞれの添え字の表わす水準値x, 水準値x1などの可能な値の全部にわたって作用させる。

制約式を式(2)の「基礎」分離性のみから導くことはできないが、つぎの2つの条件を補足して「完全」分離性とすれば導くことができ、この「完全」分離性と制約式とは等価になる(柴山 2002a, b)。

A) 加法性の式(1)の右辺の効果成分の任意の一つか任意のいくつかを0としたときにその結果として左辺に得られる応答関数yy1(a, b, ..., k)についても分離性の式(2)がなりたつこと - 直交性。

B) 加法性の式(1)の右辺の効果成分の任意の一つか任意のいくつかを加法性の式(1)と分離性の式(2)とにしたがうまったく別の応答関数yy0(a, b, ..., k)の対応する効果成分でおきかえたときにその結果として左辺に得られる応答関数yy2(a, b, ..., k)についても分離性の式(2)がなりたつこと - 交換性。

このうち、交換性はつぎの「基本」交換性と等価である(柴山 2002b)。

B') 加法性の式(1)の右辺の効果成分の任意の一对にあてはまる直交性の式が、その式に含まれる効果成分のうちの一方に任意の微小変動を与えた場合にも、なりたつこと - 基本交換性(柴山 2002b)。

組み合わせ完全配置の上で確定する応答関数について、加法性(式(1))と分離性(式(2))または正準制約式(式(2a))とによって定まる「正準展開」は「正準」効果成分の基本的な定義を与える。また、正準展開ではない一般の多方展開、線形展開および線形直交展開の効果成分の性質を整理する基礎を与える。

●正準展開の演算子表示

要因(X)ごとに平均演算子と残差演算子とを定めると(正準)効果成分が応答関数の式で表わされる。

$$\text{平均演算子 } EX := (1/X) \Sigma @x\# \quad (3) \quad \text{残差演算子 } DX := 1 - EX \quad (3a) \quad (3) (3a)$$

ここで整数1Xは要因Xの水準数を表わす。平均演算子EXおよび残差演算子DXは任意の関数に左側から作用させる代数演算子であり、つぎの恒等式(3b)がなりたつ。右辺の各項を応答関数yy(a, b, ..., k)に作用させると、結果の各項が、全体として加法性および分離性を満足して、(正準)効果成分の一つ一つを表わす。

$$(1\equiv) (EA+DA)(EB+DB) \dots (EK+DK)$$

$$\equiv EA.EB \dots EK + DA.EB \dots EK + EA.DB \dots EK + \dots + EA.EB \dots DK + \dots + DA.DB \dots DK \quad (3b)$$

●応答ベクトル

組み合わせ完全配置の上で確定する任意の応答関数の値yy(a, b, ..., k)の全部を縦一列に紙面に書き、行列代数の縦ベクトルと見なし、応答縦ベクトルと名づけ、Dirac 右括弧を用いてつぎの式(4)のように表わす。また、応答関数の値yy(a, b, ..., k)の全部を同じ順序で左から右へ横一行に紙面に書き、行列代数の横ベクトルと見なし、応答横ベクトルと名づけ、Dirac 左括弧を用いてつぎの式(4a)のように表わす。

$$\text{応答縦ベクトル } \langle yy(a, b, \dots, k) \rangle (4) \quad \text{応答横ベクトル } \langle yy(a, b, \dots, k) \rangle (4a) \quad (4) (4a)$$

この記号法を用いると組み合わせ完全配置の上で確定する応答関数yy(a, b, ..., k)を縦ベクトルまたは横ベクトルとして書き、それぞれのベクトルの要素を応答関数yy(a, b, ..., k)として書き、たがいに書きなおし、さらに、ベクトルの間の、または、ベクトルに対する演算を式として表わすことができる。

●記号累乗による水準値の表示

組み合わせ完全配置をつくる要因の総数を整数nで表わし、その一つ一つを要因Xw (w=1, ..., n)として表わす。要因Xwの水準数を整数lwで表わし、水準値の一つ一つを整数xw (=1, ..., lw)で区別する。そして、要因A, B, ..., Kを要因X1, X2, ..., Xnとして表わせば、要因A, B, ..., Kの処理(a, b, ..., k)は処理(x1, x2, ..., xn)

として表わされる。また、応答関数 $yy(a, b, \dots, k)$ は応答関数 $yy(x_1, x_2, \dots, x_n)$ として表わされる。

このほかに、要因 X_w の水準値 $x_w (=1, \dots, l_w)$ を記号累乗 x_w^{tw} ($tw=0, \dots, g_w$)で書く表示がある。指数 tw の上限 $g_w (=l_w-1)$ は要因 X_w の自由度と名づけられる。この表示では、応答関数 $yy(x_1, x_2, \dots, x_n)$ は、記号累乗の積 (Finney-Kempthorne 処理記号積) を用いて、応答関数 $yy(x_1^{t_1} x_2^{t_2} \dots x_n^{t_n})$ とする。

●応答基本ベクトルまたは応答基本関数による応答の展開 (単一要因系の場合)

要因 X_w のみで定まる応答 $yyX_w(x_w^{tw})$ の応答空間では一次独立な応答基本縦ベクトル $m(X_w^{Tw}:x_w^{tw})$ とそれに双対の対比基本横ベクトル $\langle m^*(X_w^{Tw}:x_w^{tw})$ とが要因 X_w の水準数 l_w に等しい本数ずつ定まる。どちらも対比指数 $T_w (=0, \dots, g_w)$ で一本ずつ区別し、要素 $m(X_w^{Tw}:x_w^{tw})$ または要素 $m^*(X_w^{Tw}:x_w^{tw})$ を処理指数 $tw (=0, \dots, g_w)$ で一つずつ区別する。単位演算子 I_w をつぎの式(5)で定め、その式の両辺を応答縦ベクトル $yyX_w(x_w^{tw})$ に作用させて応答基本縦ベクトル $m(X_w^{Tw}:x_w^{tw})$ または応答基本関数 $m(X_w^{Tw}:x_w^{tw})$ の一次結合を得る。係数(対比) $\langle m^*(X_w^{Tw}:x_w^{tw}) \cdot yyX_w(x_w^{tw})$ は対比指数 T_w で一つずつ区別される。

$$I_w = \sum_{T_w=0, g_w} m(X_w^{Tw}:x_w^{tw}) \langle m^*(X_w^{Tw}:x_w^{tw}) \tag{5}$$

●応答基本ベクトルによる正準展開

単一要因系での応答基本縦ベクトルおよび双対の対比基本横ベクトルのうち対比指数 T_w の値が0のものに着目する。応答主方向基本縦ベクトル $m(X_w^0:x_w^{tw})$ を単位要素ベクトル($m(X_w^0:x_w^{tw})=1$)とし、対比主方向基本横ベクトル $\langle m^*(X_w^0:x_w^{tw})$ を均分要素ベクトル($m^*(X_w^0:x_w^{tw})=1/l_w$)とする。

そして、単位演算子 I_w (式(5))を平均演算子 E_w と残差演算子 Φ_w との和として、つぎのように書く。

$$\begin{aligned} I_w &= E_w + \Phi_w \\ E_w &= m(X_w^0:x_w^{tw}) \langle m^*(X_w^0:x_w^{tw}) \\ \Phi_w &= \sum_{T_w=1, g_w} m(X_w^{Tw}:x_w^{tw}) \langle m^*(X_w^{Tw}:x_w^{tw}) \end{aligned} \tag{5a}$$

単位演算子 I_w ($w=1, \dots, n$)の直積は完全配置の応答空間の単位演算子 I でありつぎの式で展開される。

$$\begin{aligned} I &= (I_1 \times \dots \times I_n = (E_1 + \Phi_1) \times \dots \times (E_n + \Phi_n)) \\ &= E_1 \times E_2 \times \dots \times E_n + \Phi_1 \times E_2 \times \dots \times E_n + E_1 \times \Phi_2 \times \dots \times E_n + \dots + E_1 \times E_2 \times \dots \times \Phi_n \\ &\quad + \dots + \Phi_1 \times \Phi_2 \times \dots \times \Phi_n \end{aligned} \tag{5b}$$

右辺の各々の項を応答縦ベクトル $yy(x_1^{t_1} \dots x_n^{t_n})$ に作用させると効果成分縦ベクトルが一つ一つ定まり、式(3b)の(正準)効果成分と一つ一つ対応する。この応答空間の応答基本縦ベクトルは単一要因応答空間の応答基本縦ベクトルの直積であり Finney-Kempthorne対比記号積 $X_1^{T_1} \dots X_n^{T_n}$ で指定される。

これに双対の対比基本横ベクトル $\langle m^*(X_1^{T_1} \dots X_n^{T_n}:)$ も定まり、応答縦ベクトル $yy()$ との内積

$$\langle m^*(X_1^{T_1} \dots X_n^{T_n}:x_1^{t_1} \dots x_n^{t_n}) \cdot yy(x_1^{t_1} \dots x_n^{t_n}) \tag{5c}$$

は対比と名づけられ、その一つ一つは Finney-Kempthorne対比記号積 $X_1^{T_1} \dots X_n^{T_n}$ で区別される。

●一般の多方展開、線形展開および直交線形展開の定義

複数要因系を組み立てている単一要因系の応答主方向基本縦ベクトル $m(X_w^0:x_w^{tw})$ (式(5a))を単位要素ベクトル($m(X_w^0:x_w^{tw})=1$)とし、対比主方向基本横ベクトル $\langle m^*(X_w^0:x_w^{tw})$ を均分要素ベクトルではない一般のベクトル($m^*(X_w^0:x_w^{tw}) := rX_w(tw)$)とすると、正準展開の場合と同様に、単位演算子 I_w 、平均演算子 E_w および残差演算子 Φ_w が定まり、(多方)効果成分の各々が確定する(式(5b))。応答関数に対する平均演算子 EX (式(3))は係数 $1/IX$ を関数 $rX_x (=rX_w(tw))$ でおきかえてつぎの式で定義される。

$$EX := \sum_{x\#} rX_x \tag{6}$$

その結果、加法性の式が与える展開(式(1))は、正準展開ではなくなり、一般の「多方展開」となる。

効果成分の加法性(式(1))は成り立つが、応答平方和の効果成分平方和への分離性(式(2))は成り立たない。重みつき平方和により重みつき分離性を確保してもその意味を自然に解釈することは容易でない。

なお、組み合わせ完全配置の上の応答関数から、加法性の式(式(1))に基づいて、(多方)効果成分を定めるためには、正準制約式(式(2a))のかわりに、その中に含まれる効果要素の各々に対比主方向基本横ベクトル $\langle m^*(X_w^0:x_w^{tw})$ の要素 $rX_w(tw)$ を係数として付与した(多方)制約式を用いる必要がある。

複数要因系を組み立てている単一要因系の応答主方向基本縦ベクトル $m(Xw^0: xw^t w)$ (式(5a)) を単位要素ベクトルではない一般のベクトル ($m(Xw^0: xw^t w) := mXw(tw)$) とすると、同様の手順で、単位演算子 I_w 、平均演算子 E_w および残差演算子 D_w を定めて、一般の「線形展開」を組み立てることができる。なお、応答関数に対する平均演算子 E_X (式(3)) は関数 $mX_x (=mXw(tw))$ を付け加えてつぎの式で定義される。

$$E_X := mX_x \cdot \sum_{x\#} rX_x \quad (6')$$

加法性の式は、正準展開または多方展開の場合(式(1)) よりも複雑になり、つぎの式となる。

$$\begin{aligned} yy(Aa, Bb, \dots, Kk) &= yy(Aa, Bb, Cc, Dd, \dots, Qq, Uu, Hh, Kk) \\ &= mAa \cdot mBb \cdot mCc \cdot \dots \cdot mUu \cdot mHh \cdot mKk \cdot yy:M \\ &\quad + mBb \cdot mCc \cdot \dots \cdot mUu \cdot mHh \cdot mKk \cdot yy:A(Aa) \\ &\quad + mAa \cdot mCc \cdot \dots \cdot mUu \cdot mHh \cdot mKk \cdot yy:B(Bb) \\ &\quad + \dots + mAa \cdot mBb \cdot \dots \cdot mQq \cdot mUu \cdot mHh \cdot yy:K(Kk) \\ &\quad \quad + mCc \cdot \dots \cdot mUu \cdot mHh \cdot mKk \cdot yy:AB(AaBb) \\ &\quad + \dots + mAa \cdot mBb \cdot \dots \cdot mQq \cdot mUu \cdot yy:HK(HhKk) \\ &\quad \quad + \dots + \dots + yy:AB \cdot \dots \cdot K(AaBb \cdot \dots \cdot Kk) \end{aligned} \quad (6a)$$

なお、要因 X の単一要因系では $yy(Xx) = mX_x \cdot yyX:M + yyX:X(Xx)$ (6b)

加法性の式(式(6a))にあわせて、正準制約式(式(2a))の中に含まれる効果要素の各々に対比主方向基本横ベクトル $m^*(Xw^0: xw^t w)$ の要素 $rXw(tw)$ を係数として付与した線形制約式を用いると、組み合わせ完全配置の上の応答関数から、線形効果成分を定めることができるが、分離性(式(2)) は成り立たない。

ただし、線形制約式の効果要素の係数となる関数 $rX_x (=rXw(tw))$ として関数 $mX_x (=m(Xw^0: xw^t w))$ の定数倍 ($1/\sum_{x\#} mX_x^2$ 倍) を用いると、組み合わせ完全配置の上の応答関数平方和は式(6a)または(6b)の右辺の各項(線形効果成分)の平方和の総和に等しくなり、その意味では、分離性が成り立つ。

その線形効果成分はたがいに直交し、加法性の式(式(6a))は「直交線形展開」を与える(柴山 2003b)。

●単一要因系の応答方程式の表示から発生する不定性

要因 X のみを含む応答関数 $yy(Xx)$ について、要因 X の水準値 $X_x (x=1, \dots, 1X)$ ごとに、線形展開の加法性の式(式(6b))を書き、その全部(1X個)をその応答関数の応答方程式としてつぎの形で表わす。

$$\begin{aligned} y1 &= m1 \cdot c0 + 1 \cdot c1 + 0 \cdot c2 + \dots + 0 \cdot cx + \dots + 0 \cdot c1 \\ y2 &= m2 \cdot c0 + 0 \cdot c1 + 1 \cdot c2 + \dots + 0 \cdot cx + \dots + 0 \cdot c1 \\ \dots &\dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ yx &= mx \cdot c0 + 0 \cdot c1 + 0 \cdot c2 + \dots + 1 \cdot cx + \dots + 0 \cdot c1 \\ \dots &\dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ y1 &= m1 \cdot c0 + 0 \cdot c1 + 0 \cdot c2 + \dots + 0 \cdot cx + \dots + 1 \cdot c1 \end{aligned} \quad (6c)$$

ただし、最後の式の左辺の値 $y1$ の添え字 1 、最後の式の右辺の第1項の係数 $m1$ の添え字 1 、および、各々の式の右辺の最後の項の値 $c1$ の添え字 1 は、どれも、要因 X の水準数 $1X$ を表わしている。

各々の式の右辺の第1項の係数 $mx (x=1, \dots, 1)$ は応答主方向基本関数 $mX_x (=mXw(tw))$ (式(6')) の値である。各々の式の左辺の応答測定値 $y1, y2, \dots, yx, \dots, y1$ の各々は要因 X の各水準での応答関数 $yy(Xx)$ の値の各々 $yy(X1), yy(X2), \dots, yy(Xx), \dots, yy(X1)$ を表わしている。各々の式の右辺の第1項の値 $c0$ は一般平均の効果要素 $yy:M$ (定数) を表わしており、そのあとの各項の値 $c1, c2, \dots, c1$ はそれぞれ要因 X の主効果の効果要素の各々 $yy:X(X1), yy:X(X2), \dots$ または $yy:X(X1)$ を順に表わしている。

各々の式の右辺の効果要素 cx の係数の各々はこの単一要因系の要因配置実験の「計画行列」をつくる：

$$\begin{bmatrix} m1 & 1 & 0 & \dots & 0 & \dots & 0 \\ m2 & 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ mx & 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ m1 & 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \quad (6d)$$

応答方程式(式(6c))の左辺の応答測定値 $y1, y2, \dots, y1$ を縦1列に書いて応答縦ベクトル yy' (式(6e))とし、右辺の効果要素 $c0, c1, c2, \dots, c1$ を縦1列に書いて効果要素縦ベクトル cc' (式(6f))とすると、この効果要素縦ベクトル cc' に左から計画行列(式(6d))を掛けたものは応答縦ベクトル yy' になる。

$$yy' = (y1, y2, \dots, y1)' \quad (6e) \quad cc' = (c0, c1, c2, \dots, c1)' \quad (6f) \quad (6e)(6f)$$

応答測定値(式(6e))の全部(1個)が定まっても、応答方程式の解となる効果要素(式(6g)) (1+1 個) は一通りには定まらない。解となる効果要素の一通りがつくる効果要素縦ベクトル cc_o' をつぎの形に書くと、

$$cc_o' = (c0_o, c1_o, c2_o, \dots, cx_o, \dots, cl_o)' \quad (6g)$$

この縦ベクトル cc_o' につぎの縦ベクトル cc_m' の任意定数 (仮に r とする) 倍を加えたものも解となる。

$$cc_m' = (-1, m1, m2, \dots, m(x-1), mx, m(x+1), \dots, ml)' \quad (6h)$$

この縦ベクトル cc_m' は計画行列(式(6d))に対して、不定倍数 r を除き、一通りただ一通りに定まる。効果成分の表示から発生する不定性が、この縦ベクトルを用いて、具体的に表わされる (Shibayama 2003a)。

この縦ベクトル cc_m' を、仮に、「効果‘零’方向基本縦ベクトル」と名づける。応答方程式(式(6e))の解の一つとして効果要素縦ベクトル一本 cc_o' が定まると、これに効果零方向基本縦ベクトル cc_m' の任意定数倍を加えたものは同じ応答方程式の解の一つとなる。また、同じ応答方程式のどの解ももとの効果要素縦ベクトル cc_o' に効果零方向基本縦ベクトルの定数倍 $r \cdot cc_m'$ を加えたものとして表わされる。

したがって、線形制約式の差異または一般逆行列の差異は任意倍数 r の差異となる。また、同じ応答測定値に対する異なる効果 (要素) の差は効果零方向基本縦ベクトルの任意定数倍 (の要素) に等しい。

●複数要因系の応答方程式の表示から発生する不定性

複数要因のつくる組み合わせ完全配置の上の応答関数の線形展開で、加法性のみによって定まる線形効果成分の不定性については、要因ごとに単一要因系の応答方程式、計画行列および効果零方向基本縦ベクトルをつくり、全部の要因にわたって、それぞれ、直積をつくると、基本的に同じ数理があてはまる。

ただし、こうして得られる複数要因系では、単一要因の効果零方向基本縦ベクトルのつくる「純粹」直積ベクトル一本のほかにも、「複合」直積ベクトル多数本も、それぞれ、効果零方向基本縦ベクトルとなる。

複合直積ベクトルをつくるには、1)各々の単一要因系の計画基本横ベクトルの直積として複数要因系の計画基本横ベクトルをつくり、2)これを転置して、複数要因系の「計画基本縦ベクトル」とし、3)その直積因数の一つまたはいくつかを対応する単一要因系の効果零方向基本縦ベクトルでおきかえる。

この複合直積ベクトルは純粹直積ベクトルと直交する。また、同様にして得られる複合直積ベクトル多数本のうちのどの一本とも直交する。また、複数要因系の計画基本横ベクトルのどの一本とも直交する。

複数要因系の応答方程式の解となる効果要素縦ベクトルの不定部分は複数要因系の効果零方向基本縦ベクトル (すなわち、純粹直積ベクトルおよび複合直積ベクトル) の任意の一次結合で表わされる。

●一般線形模型の解析の手順の代数的な吟味

単一要因系の直積として組み立てられる組み合わせ完全配置の上での応答関数の効果成分への展開は一般の複数要因系の解析の基礎となり、その不定性の吟味は各種の解析の手順の意味をわかりやすくする。

SAS/STATソフトウェア GLMプロシジャに組み込まれている推定可能性・平方和縮減・平方和分離・掃き出し法などの一般逆行列による数理が見やすくなり、線形制約式による数理との関係も明瞭になる。

推定可能性は、代数的には、複数要因系の応答方程式の解となる効果要素縦ベクトルに不定部分がつけ加わっているにも関わらず、その不定性の影響を受けない結論をとり出すことが可能なことを意味する。

応答の測定値 y を真値 yy と揺動 vv との和と仮定する。真値 yy と効果要素の真値 cc については応答方程式があてはまるとする。その一方で、応答の測定値 y をその推定値 yv と残差 vy との和と仮定し、推定値 yv と効果要素の推定値 cv については、応答の真値 yy と効果要素の真値 cc についてあてはまると同じ形の応答方程式を仮定し、実験配置の上の残差平方和を最小にするよう各々の推定値を定める演算を組み立てる。

ここで、応答の推定値 yv はその真値 yy と揺動 vv の一部分を含む。効果要素の推定値 cv はその真値 cc と揺動 vv の一部分を含む。残差 vy は揺動 vv の一部分のみを含むから揺動 vv の大きさの推定に用いる。

揺動 vv が 0 の場合には、応答の測定値 y はその真値 yy に等しく、応答方程式の解は効果要素の真値 cc となるが、不定部分を含む。残差平方和を最小にする応答の推定値 yv と効果要素の推定値 cv についても同じ形の応答方程式があてはまるが、その解となる効果要素の推定値 cv は、やはり、不定部分を含む。

ただし、応答の推定値 y_v （および、応答の推定値 y_v の定数係数一次結合）、残差 v_y 、および、実験配置の上での残差平方和は不定部分の影響を受けない — 応答方程式の最小2乗解の推定可能性の定理。

応答の測定値 y が(活動)効果要素を含まず揺動 v_v のみを含むとすれば実験配置の上の応答平方和が残差平方和 SSE_0 となる。効果要素のいくつかを無効要素(0)と見なし、残りのいくつかのみを活動要素 cc として含む応答方程式を組み立て、その推定値 c_v 、応答の推定値 y_v 、および、残差 v_y を求め、実験配置の上での残差平方和 SSE_1 を求めると、残差平方和 SSE_0 より小さい。その差 R ($:=SSE_0-SSE_1$) (平方和縮減)は、活動要素の影響を表わしているが、その真値 cc の不定部分または推定値 c_v の不定部分の影響を受けない。

SAS/STATソフトウェアでは、応答方程式に組み入れる活動効果要素を変化させ、さまざまな平方和縮減を定量して、平方和 SSI または SSII (SAS/STAT® User's Guide, Vn 6, 4th edn, p.115-118) を定める。なお、平方和 SSIII または SSIV も定めるが、この4種類の平方和の全部に共通な定義としては、応答方程式または規準方程式のGauss-Jordan-Doolittle解法から得られる推定可能関数に基づく定義が用いられる。ただし、どの平方和の値も効果要素の真値 cc の不定部分または推定値 c_v の不定部分の影響を受けない。

● 応答方程式または規準方程式のGauss-Jordan-Doolittle解法

応答の測定値 y が応答の真値 y として確定する場合には応答方程式を解いて効果要素の真値 cc を求め、応答の測定値 y が揺動 v_v を含む場合には応答方程式を規準方程式に変化させて解き効果要素の推定値 c_v を求める。Gauss-Jordan-Doolittle解法 (Goodnight 1978) を用いると、不定部分の実例値も算出できる。

たとえば、単一要因系の連立一次方程式(式(6c))で、第2式以下の各々の式から、第1式の(m_x/m_1)倍を引き、効果要素 c_0 を消去する。つぎに、第3式以下の各々から、類似の手順で、効果要素 c_1 を消去する。特別の不具合が発生しなければ、この前進消去を重ねて、最後に、第1式に効果要素 $c_{(l-1)}$ と効果要素 c_l とだけが残る。ここで、効果要素 c_l を0とすると、効果要素 $c_{(l-1)}$ が定まる。その後、後退代入により、効果要素 $c_{(l-2)}, \dots, c_1, c_0$ を順に定める。効果要素 c_l の値を0のかわりに任意の値として後退代入を行なうと、効果要素 c_0, c_1, \dots, c_l として、ほかの一通りが定まる。もとの一通りととの差は効果要素の不定部分である。

SAS/STATソフトウェア解説文書・参考文献で水準番号の最大の効果要素の値が0になっている例はこの解法の結果と思われる(たとえば、SAS/STAT® User's Guide, Vn 6, 4th edn, p.967, Output 24.9 (17))。

● 推定可能関数 - SAS/STAT® User's Guide, Vn 6, 4th edn, Ch. 9, p.109-124

一般の単一または複数要因系の応答方程式または規準方程式について前進消去を行なうと、最初の連立一次方程式は変化する。その結果、各々の式の右辺には効果要素のさまざまな一次結合が現われて、それぞれが「推定可能関数」となる。たがいに一次独立な推定可能関数の一揃いの個数(q とする)の上限(q_u とする)は最初の応答方程式の右辺の効果要素の真値 cc または推定値 c_v の係数のつくる「計画行列」(X とする)の階数に等しい。一揃いの推定可能関数の各々の一次結合係数がつくる「推定可能行列」(L とする)は計画行列 X (行数 x)に左から「縮約行列」(K とする)を掛けて得られる($L = KX$)。

推定可能行列 L の行をそれぞれ「推定可能行ベクトル」 $L^{\sim}1, \dots$ 、または、 $L^{\sim}q$ とする。推定可能行列 L を効果要素縦ベクトル cc または c_v に左から掛けて「推定可能要素縦ベクトル」($L^{\sim}1.cc, \dots, L^{\sim}q.cc$)' または ($L^{\sim}1.c_v, \dots, L^{\sim}q.c_v$)' を得る。一次独立な推定可能関数の「完全な」一揃い(q_u 個)がつくる推定可能行列(Lu とする) (行数 q_u) が含む推定可能行ベクトル $Lu^{\sim}1, \dots, Lu^{\sim}q_u$ の任意の一次結合 $Lu^{\sim}o$

$$Lu^{\sim}o = Lu_{.1}.Lu^{\sim}1 + \dots + Lu_{.q_u}.Lu^{\sim}q_u \quad (7)$$

は「推定可能横ベクトル」の一般形であり、これと効果要素縦ベクトル cc または c_v との内積が推定可能関数の一般形を与える。結合係数 $Lu_{.1}, \dots, Lu_{.q_u}$ を変化させて任意の推定可能関数が生成される。

● 推定可能関数による平方和の計算

前進消去の結果として得られる連立一次方程式から消去の各々の段階を逆にたどって「回復行列」(J とする)が得られる場合には、これを推定可能行列に左から掛けて計画行列 X が得られる($X = JL$)。

計画行列 X を効果要素推定値縦ベクトル cv に左から掛けると応答推定値縦ベクトル $yv (=X \cdot cv)$ が得られるから、実験配置の上での応答推定値平方和 SS_{yv} が推定可能関数 $L \cdot cv$ の2次形式で表わされる。

$$SS_{yv} = yv' \cdot yv = cv' \cdot X'X \cdot cv = (L \cdot cv)' \cdot J'J \cdot (L \cdot cv) \quad (7a)$$

行列積 $J'J$ をつくる回復行列 J は完全な(「完全系の」ともいう)推定可能行列 Lu (行数 qu)に作用して計画行列 X の行の全部を生成させる。SAS/STATソフトウェアでは行列積 $J'J$ の表示に一般逆行列 $(X'X)^{-}$ を含む逆行列 $[L \cdot (X'X)^{-} \cdot L']^{-1}$ を用いている(SAS/STAT® User's Guide, Vn 6, 4th edn, p.110)。

完全系の推定可能行列 Lu (行数 qu)の行のいくつか $L^{\sim}1, \dots, L^{\sim}q$ をえらんで「有効行」とし、ほかの行を「無効行」と見なして要素の各々を0でおきかえ、部分系の推定可能行列 L を形式的に定める。

この推定可能行列 L に左から完全系の推定可能行列 Lu の回復行列 J を掛けて計画行列 X を形式的に定めると、規準方程式から、有効行に対応する推定可能関数 $L^{\sim}1 \cdot cv, \dots, L^{\sim}q \cdot cv$ が定まり、部分系の推定可能成分 $yyP (=J \cdot (L \cdot cv))$ と部分系の推定可能成分平方和 $SS_{yyP} (=yyP' \cdot yyP)$ (式(7a))とが定まる。

計画行列 X の一次独立な行から生成される完全系の推定可能行列 Lu (および、その部分行列 L)はいくつもあり(SAS/STAT® User's Guide, Vn 6, 4th edn, Ch.9 The Four Types of Estimable Functions), 各種の回復行列 J , 各種の推定可能成分 yyP および各種の推定可能成分平方和 SS_{yyP} が生成される。

●推定可能成分平方和による仮説検定

応答測定値縦ベクトル y は、規準方程式 $(X'X \cdot cv = X' \cdot y)$ の解を用いると $(cv = (X'X)^{-} \cdot X' \cdot y)$, 応答推定値縦ベクトル yv とそれに直交する残差縦ベクトル vy との和に分解でき、つぎの式が成り立つ。

$$yv = X \cdot cv = X \cdot (X'X)^{-} \cdot X' \cdot y \quad SS_{yv} = yv' \cdot yv = (L \cdot cv)' \cdot J'J \cdot (L \cdot cv) = y' \cdot X \cdot (X'X)^{-} \cdot X' \cdot y \quad (7b)$$

$$vy = y - yv = (Ix - X \cdot (X'X)^{-} \cdot X') \cdot y \quad (\text{ただし, } Ix \text{ は計画行列と同じ行数の単位行列とする})$$

$$SS_{vy} = y' \cdot (Ix - X \cdot (X'X)^{-} \cdot X') \cdot y = SS_y - SS_{yv} \quad SS_y = y' \cdot y \quad (7c)$$

応答測定値平方和 SS_y は推定可能成分平方和 SS_{yv} と残差平方和 SS_{vy} との和に等しい。

零仮定として、応答測定値 y の含む真値 yy が恒等的に0であり、応答測定値 y が揺動の標本値 vv のみを含む場合を想定する。応答測定値縦ベクトル y は x 次元線形空間のベクトルであり、その要素は確率的に独立な揺動の標本値 vv の x 個となる。その母集団を正規分布母集団と仮定する。この空間の(直交)座標系に直交変換を作用させると、応答測定値縦ベクトル y の要素(x 個)はその一次結合(x 個)にそれぞれ変換され、結果として得られる一次結合(x 個)も同じ正規分布母集団の確率的に独立な揺動の標本(x 個)となる。

適切な直交変換を用いると、有効行数 q の推定可能行列 L から規準方程式によって定まる応答推定値縦ベクトルすなわち部分的な推定可能成分 yyP を変換後の座標軸(x 本)のうちの q 本の一次結合として表わすことができ、残差縦ベクトル vyP を残り $(x-q)$ 本の座標軸の一次結合として表わすことができる。

各々の一次結合係数はもとの応答測定値縦ベクトル y の要素 vv_i に直交変換を作用させて得られる要素 vv の一次結合であり、それぞれ、もとの要素 vv と同じ正規分布母集団の独立な標本となる。そこで、推定可能成分 yyP については、尤度比 $W (:= (SS_{yv}/q)/(SS_{vy}/(x-q)))$ を用いて、零仮定を検定する。

一連の計算の結果は、どれも、効果成分の表示にともなう不定性の影響を受けず、一通りに確定する。これは計算手順の一段階ごとに確認される。なお、一般逆行列解と線形制約式解との関係も追跡できる。

参考文献

Fisher, R.A. (1925, 1948): Statistical methods for research workers., Oliver & Boyd.

Goodnight, J.H. (1978): Sweep operators: Its importance ..., SAS Technical Report R-106., SAS Institute, Inc.

SAS Institute, Inc. (1990): SAS/STAT® User's Guide, Vn 6, 4th edn, vol.1 and 2., SAS Institute, Inc.

Shibayama, T. (2003a): Effect components being defined ... in the indeterminate expressions., The 54th ISI Session.

柴山忠雄(2002a): 要因配置実験の結果整理のための定理., 日本行動計量学会大会, 抄録集p. 164-167.

柴山忠雄(2002b): 効果成分の直交性に伴う交換性., 日本品質管理学会年次大会4-3, 要旨集p. 115-118.

柴山忠雄(2003b): 任意の ... 関数に基づく直交応答分解., 日本品質管理学会第71回研究発表会6-8.

口頭論文発表
統計教育

日本SASユーザー会 (SUGI-J)

CROにおけるSASプログラマの育成教育

○竹田 眞* 佐藤 智美**

株式会社 CRC ソリューションズ / CRO 業務部 統計解析チーム

*関西支社, **東京本社

The Education for SAS Programmer at CRO

Makoto Takeda Tomoyoshi Sato

CRC Solutions Corp.

CRO Department Data Management & Biostatistics Section

要 旨

人件費が費用の大半を占めるCROにおいては人材の育成が急務である。またCROでは出力成果物を商品として提供していることから単に正確な出力結果を出すだけでなく、レイアウトの見やすさなど出力結果の美麗さについても求められることがある。よってCROのSASプログラマには、出力結果を自由自在に加工するSASの技術も必要となる。そこでより効果的に新入社員を一人前のSASプログラマに育成するため、体系だった社内教育カリキュラムの整備を行った。目標は少なくとも1年以内に基本的なプロシジャ、SAS関数、データステップによるデータ加工の知識を習得し、上級プログラマが作成したプログラム仕様書に基づきSASプログラミングが出来るレベルまでの育成である。今回はこの教育カリキュラムの概要について紹介する。

キーワード： 模擬解析演習

1. はじめに

当社は情報処理企業であるが、臨床試験に関わるデータマネージメントや統計解析処理、モニタリングといったCRO(開発業務受託機関)業務のサービスも提供している。特にデータマネージメント・統計解析部門では入力データや集計解析結果といったものを納品することにより対価を得ることを生業としているが、そのコストの大半は人件費である。またサービスを商品として提供している以上、単に結果を求めるだけでなくレイアウトの見やすさなど付加価値のある出力成果物を作成する必要がある。そこでより効果的に新入社員を一人前のSASプログラマに育成するため従来はOJT[On the Job Training]中心だったSASプログラミングの社内教育方針を変更し、体系だった社内教育カリキュラムとして整備することにした。

2. 社内教育制度について

当社での新入社員に対する社内研修は大きく3つ分かれる。まずは全新入社員に対して行われる研修で、これは入社後2ヶ月間にわたって開催される。内容的には会社自身がCRO業務だけでなく金融や流通、建築といった多岐の分野でのITサービスを提供する情報処理企業であるため、一般的な社会人研修に加えて、情報処理に関する研修が大部分を占める。新入社員はプログラムの経験がない場合でもこの研修期間を通じて基本的なプログラミング技術やアルゴリズムを習得することが出来る。次にCRO部門の研修としては「医薬品開発」「CRO業務」「医薬品の基礎」「法令・規則」「医学基礎講座」といった講座が配属後1ヶ月間開催され、医薬系学部出身者以外でも最低限必要な業務知識の習得が可能である。更にデータマネージメント・統計解析チームに配属される新入社員に対してはSAS社のトレーニングコース、OJTによる指導、月例チーム勉強会などが用意されている。

図1:2002年度入社新入社員に対する教育カリキュラム

1. 全社研修(4月~5月)
 - ・ コンピュータの基礎知識
 - ・ PMLとリテラシー
 - ・ ネットワーク/データベース入門
 - ・ UNIX基礎技術入門
 - ・ システム設計入門
 - ・ アルゴリズム/フローチャート入門
 - ・ C言語入門
 - ・ 企業人研修
 - ・ マナー研修
2. CRO部門研修(6月)
 - ・ 医薬品開発に対する理解
 - ・ CRO業務部の業務に対する理解
 - ・ 医薬品の基礎
 - ・ 法令・規則に対する理解
 - ・ 医学の基礎
3. DM・統計解析チーム研修(6月~)
 - ・ SAS社トレーニングコース
 - ・ OJTによる指導
 - ・ 月例チーム勉強会(DM・統計)
4. その他(随時)
 - ・ ステップアップ研修
 - ・ 新入社員IT研修(VB,ACCESS, JAVA)

3. 現チーム体制と育成方針

現在弊社ではデータマネジメント業務と統計解析業務を同一チーム内で行っている。これはデータマネジメントを行うスタッフであっても、自らが作り出すデータがより正しく、より効率よく解析されるためにはどのような構造であるべきかを理解しておく必要があり、また統計解析を行うスタッフであってもCRFの内容がどのようにコンピュータデータとして表現されているかを把握しておくことが必要と考えるからである。従って入社3年目くらいまでは両方の業務を経験し、その後本人の適性・希望などを考慮し、どちらかの専門性を高めるというキャリアパスを通例としている。

4. 新規教育カリキュラム

以上のように新入社員には上記に述べた全社研修・CRO 部門研修が用意されているため、PC の操作方法はもちろんのこと基本的なプログラミングやアルゴリズム、GCP や SOP などを含む業務知識習得はクリアされることになる。

次に現場で必要となるのは SAS のプログラミング技術であるが、これまでに入社した新入社員は SAS プログラミング経験がない場合が多く、よって SAS 社で実施されるトレーニングコースへの参加を実施している。これはDATAステップ、プロシジャといった概念や操作方法については既に SAS 社で洗練されたトレーニングコースが用意されているため、社内で研修を実施するより効率的で、利便性も高いと考えている。

また実務で使用するプログラムについては、従来はOJT制度のもと新入社員はOJTトレーナーの指示に従い、簡単なプログラミングから作業を始めていた。しかし、この方法ではその時々業務によって体験する内容が異なることがあり、ある者は症例一覧表のみ、ある者は集計表のみと、状況によっては大きな偏りが生じ、実際に入社後2年で初めて SAS GRAPH を使用したという例もあった。

そこで今回はこのような偏りをなくすべく、ダミーデータを用いた模擬解析演習を教育カリキュラムに取り入れ、その中で実践的なプログラミング技術を習得できるように考えた。

5. 模擬解析演習

模擬解析演習ではあらかじめ用意されたデータ、仕様書を元に解析データの作成から解析結果を作成する。模擬解析演習を通じてさまざまな SAS プログラミング技術やプロシジャ、関数の使い方を体験し、その使い方を習得することを目標としている。

具体的には以下のレベルの内容を作成することを想定しており、実施開始時期は CRO 部門研修及び SAS 社トレーニングコース終了後からを予定している。

<模擬解析演習演題>

- ① 入力データから解析用データへの変換
- ② 症例一覧表
- ③ 頻度集計表(例数、%)
- ④ 基礎統計量表(例数、平均、標準偏差、最小値、最大値)
- ⑤ グラフ(散布図・経時的推移)
- ⑥ 検定(χ^2 検定、t検定、Wilcoxon 検定)

6. 演習手順

演習において事前に与えられる教材は以下のものである。

- ① 入力データ(SAS データセット)
CRF からの入力をイメージしたダミーデータ。データ間の不整合は存在しない。
- ② 入力データ変数定義書
「①入力データ」のデータベース定義書
- ③ 解析用データ定義書
入力データから変換される解析用データベースの定義書
- ④ 統計解析計画書(図表レイアウトを含む)
- ⑤ 解析プログラム仕様書
各図表に出力される項目に使う変数名やプロシジャ名、パラメータ或いは合成変数の生成定義を規定したプログラムの仕様書
- ⑥ プログラムサンプル
プログラム作成にあたって参考にするプログラムサンプル。各機能の部分的な箇所が記載されている。

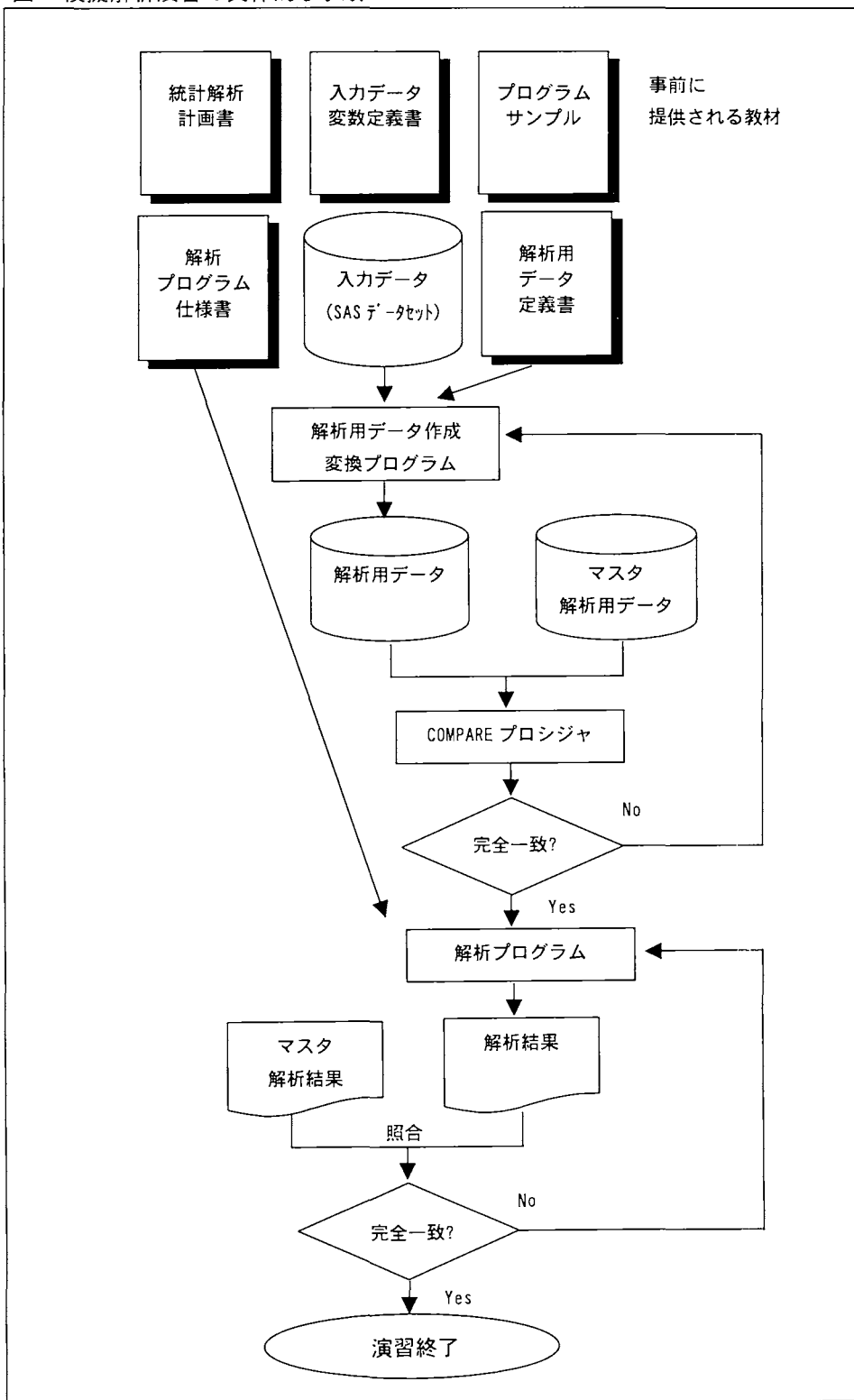
また各作業完了時に提供される教材として以下のものがある。

- ⑦ マスタ解析用データ(SAS データセット)
バリデーション済みの解析用データ
- ⑧ マスタ解析結果
バリデーション済みの解析結果

次に演習は以下の手順で進められる。(図2参照)

1. 解析用データ定義書に従い、入力データから解析用データを作成する変換プログラムを作成する。
2. 解析用データ作成後、マスタ解析用データと SAS COMPARE プロシジャを使って比較し、相違点を確認する。不一致箇所があれば変換プログラムにバグがあることを意味するので完全一致するまで訂正を繰り返す。
3. 統計解析計画書と解析プログラム仕様書に従い、解析プログラムを作成し、各図表に解析結果を出力する。
4. 解析結果を出力後、マスタ解析結果と照合し、正しく出力が行われたかと確認する。不一致箇所があれば解析プログラムにバグがあることを意味するので完全一致するまで訂正を繰り返す。

図2: 模擬解析演習の具体的な手順



7. 期待されるメリット

本模擬解析演習の実施による期待される一番のメリットはSASプログラマの早期育成である。模擬解析演習を通じて、業務で作成頻度の高い図表の作成技術を短期に且つ体系的に会得することを期待している。また OJT トレーナの負担軽減も期待できる。OJT トレーナは新入社員ひとりに1名ずつ任命されるが、新人教育以外に実務も担当している訳で、繁忙時には作業指示のための十分な時間の確保が困難な場合もある。しかしながら模擬解析演習ではあらかじめ仕様書及びバリデーションされた結果が用意されているため、新入社員が自ら演習を進め、結果を確認することが出来、OJT トレーナは本人が自力で解決できない場合のみヘルプすればよい。また演習終了後の実務では OJT トレーナが一から十までの指示せずともある一定以上の作業量が期待できるものと考えている。

8. おわりに

本模擬解析演習の実施は 2003 年度入社の新入社員からを対象に現在鋭意作成中である。内容的には 200 例程度のシンプルな 2 群比較を予定している。初心者への演習ということで解析手法もシンプルなものに留める予定であるが、データ上では不完全な記載の日付データを混ぜ、その取り扱い処理を必須にしたり、タイムウインドウによる採用時期の選定等の処理も含める予定である。また将来的にはクロスオーバーや薬物動態といった様々なデザインの問題を作ることや、入力データベースを作るところでデータマネージメントの演習までを視野に入れることが可能と考えている。

口頭論文発表
システム

日本SASユーザー会 (SUGI-J)

CALL EXECUTEを用いたマクロの再帰呼び出しと統計計算への応用

伊藤 要二

アストラゼネカ株式会社
臨床統計・プログラミング部

A recursive SAS macro technique using CALL EXECUTE and its application to statistics

Yohji Itoh

Statistics & Programming Department, AstraZeneca K.K.

要 旨

再帰呼び出しは反復計算を行う上では重要な機能であり、他のいくつかのプログラミング言語では利用可能であるが、SAS言語ではそのような機能は提供されていない。しかし、ここで紹介するテクニックを用いれば、マクロの再帰呼び出しを簡単に行うことができる。この方法はCALL EXECUTEを用いるものであるため、まずCALL EXECUTEについて説明し、次にCALL EXECUTEによるSASマクロの再帰呼び出しの解説をおこなう。最後に統計処理への応用事例として、MIXEDプロシジャを用いたPower-of-the-mean modelの反復計算を紹介する。

キーワード： 再帰呼び出し、マクロ、CALL EXECUTE

1. はじめに

いくつかのプログラミング言語 (例えば、PascalやPL/Iなど) においては、サブルーチンの再帰呼び出しの機能が利用できる。この再帰呼び出しのテクニックを用いれば、自分自身を呼び出すようなサブルーチンプログラムを書くことができ、それにより反復処理が非常に容易になることがある。しかしながら、SASシステムではそのような機能は提供されておらず、よって、他の方法を用いなければならない。Benjamin (1999)は他言語における再帰処理に用いられているrun-time stackを模倣した擬似的再帰SASマクロ(pseudo-recursive SAS macro)のテクニックを提案したが、この方法は再帰処理についての特殊な知識を必要とする非常に煩雑なものであり、一般のSASユーザが容易に利用できるものではない。

本発表では、マクロの再帰呼び出しの新しい方法を提案する。この方法は非常に簡単であり、再帰呼び出しについての特殊な知識を必要としないものである。

まずSASマクロの一般的な問題点を示し、その欠点を補う方法としてCALL EXECUTEの利用に

ついて解説する。次に、それを発展させたものとして、CALL EXECUTEによるSASマクロの再帰呼び出しの解説をおこなう。そして最後に、統計処理への応用事例として、MIXEDプロシジャを用いたPower-of-the-mean modelの反復計算などを紹介する。

2. CALL EXECUTE

CALL EXECUTEはDATAステップで用いられるCALLルーチンで、Riba (1997)によって詳細に解説されている。CALL EXECUTEはプログラム1に示すように用いる。CALL EXECUTEの引数はSASステートメントからなる文字列で、文字定数でも文字変数でも構わない。

CALL EXECUTEの処理の流れについての知識はマクロの再帰呼び出しにとって重要である。しかし、それを理解するには通常のSASプログラムの処理の流れについて理解しておく必要がある。まず通常のSASプログラムの処理の流れについて簡単に説明する。

図1は通常のSASプログラムの処理の流れを示したものである。SASプログラムがサブミットされても、それが直ぐにSASシステムによってコンパイルされるわけではなく、先ずはプログラム・スタックという場所に記憶される。RUNステートメントや次のステップが見つかったら、プログラム・スタックに蓄えられたSASプログラムがコンパイルされ、実行される。

図2はCALL EXECUTE ステートメントを含むDATAステップのプログラムの処理の流れを描いたものである。プログラムの実行開始のところまでは通常のプログラムの処理の流れと同じである。DATAステップの中のCALL EXECUTEが実行されると、その引数(この例では「abc」)がプログラム・スタックに蓄えられる。そのDATAステップの実行が終了すると、制御はプログラム・スタックに蓄えられていたプログラムに移され、それがコンパイル・実行される。

ここで注意すべきことは、CALL EXECUTEによって生成されたプログラム・ステートメントは、それを生成したDATAステップの実行が完了するまではコンパイルされないということである。この点については後に再び述べる。

プログラム1

```
data ...;
...
call execute(' SASステートメント ');
...
run;
```

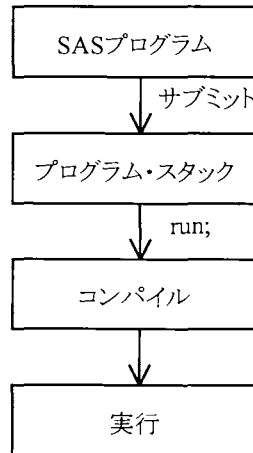


図1. 通常のSASプログラムの処理の流れ

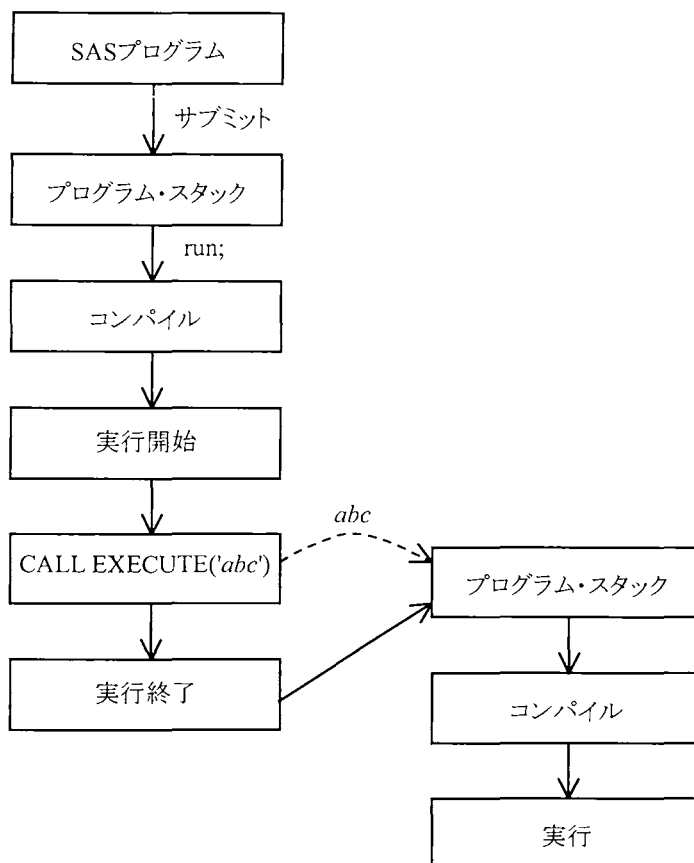


図2. CALL EXECUTEを含むSASプログラムの処理の流れ

3. マクロ処理の問題点

マクロ処理の問題点を説明するために、プログラム2のようなプログラムを考える。プログラム2が実行された時、IFステートメントの条件判定に従って%macroxが展開されるものと期待するかもしれない。すなわち、条件が「真」なら「a」を引数として展開し、「偽」であるなら「b」を引数として展開されるものとも考えるかもしれない。

プログラム2

```

data ...;
...
if (条件) then %macrox(a);
           else %macrox(b);
...
run;
  
```

しかし、実際にはそのようにはならない。図3に示すように、SASのマクロ・プロセッサはそのDATAステップの翻訳・実行に先立って、マクロを展開してしまうからである。よって、そのDATAステップが実行される時には、既にマクロの展開は終了してしまっているのである。Riba (1997)が詳しく述べているように、一般的に、プログラムの実行結果に従ってマクロの展開を変更することはできない。

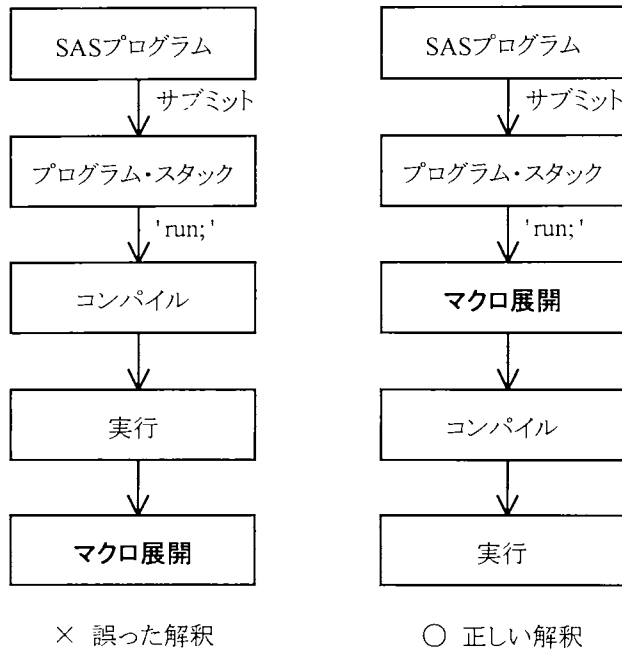


図3. プログラム2の処理の流れの解釈

4. CALL EXECUTEによるマクロの呼び出し

CALL EXECUTEを用いると、プログラムの実行結果に従ってマクロ展開を変更することが可能となる。プログラム3においては、IFステートメントの条件が「真」の場合には「%macrox(a)」がプログラム・スタックに蓄えられ、条件が「偽」の場合には「%macrox(b)」が蓄えられる。

プログラム3

```

data ...;
...
if (条件) then call execute('%macrox(a);');
                    else call execute('%macrox(b);');
...
run;
```

そして、このDATAステップが完了した後に、プログラム・スタックに蓄えられたステートメントが展開・コンパイルされることになる。よって、このプログラム3のようにすれば、IFステートメントの条件判定に従ってマクロ展開をコントロールすることが可能となる。

5. マクロの再帰呼び出し

この考えを拡張すれば、自分自身を呼び出すマクロ・プログラムを構築することが可能となる。プログラム4はこの考えを示したものである。このプログラムにおいては、マクロ%mcxはCALL EXECUTEによって自分自身を呼び出している。このマクロ・プログラムの実行が終了すると、CALL EXECUTEによって呼び出された自分自身であるマクロ・プログラムの展開、コンパイル、

実行が開始される。よって、マクロの再帰的な実行がなされる。もしIFステートメントの条件が「偽」であるなら、マクロ・プログラムは呼び出されず、処理は終了する。

類似の処理はPL/IやPASCALのような他の言語では利用可能であり、「サブルーチンの再帰呼び出し」と呼ばれている。そのようなサービスはオリジナルのSASシステムでは提供されていないが、CALL EXECUTEを用い

れば、そのような再帰処理が可能となる。この方法をここでは他の言語にならって、「マクロの再帰呼び出し」と呼ぶことにする。

プログラム4

```
%macro mcrx;
  ...
  data ...;
  ...
  if (条件) then call execute('%mcrx;');
  ...
run;
...
%mend;

%mcrx;
```

6. マクロの再帰呼び出しの統計への応用 — Power-of-the-mean model

6.1 Power-of-the-mean modelとは

マクロの再帰呼び出しの統計への応用例として、power-of-the-mean model (Carroll & Ruppert, 1988, Littell *et al.*, 1996参照)を考える。このモデルにおいては、各観測値の誤差分散はその期待値のべき乗に比例すると仮定する。すなわち、 i 番目の個体の誤差分散は次のように表される:

$$\sigma_{e_i}^2 = \sigma^2 | \mathbf{x}'_i \boldsymbol{\beta} |^\theta$$

ただし、 σ^2 は未知の分散パラメタ、

\mathbf{x}'_i はデザインマトリックス \mathbf{X} の*i*番目の行、

$\boldsymbol{\beta}$ は未知の固定効果のベクトル、

θ は未知のべきパラメタ。

SASにおいては、power-of-the-mean modelは、MIXEDプロシジャのREPEATEDステートメントにおいてLOCAL=POMオプションによって指定することができる(Littell *et al.*, 1996):

```
REPEATED /LOCAL=POM(SASデータセット);
```

POMの後の括弧の中には、固定効果の値が収められたSASデータセットを指定する。すなわち、MIXEDプロシジャのpower-of-the-mean modelにおいては固定効果の値は既知であることが前提とされていて、それをSASデータセットとして与えてやらなくてはならない。実際にはそれが既知であることはほとんどなく、通常はデータから推定しなければならない。何らかの固定効果推定値(例えば $\theta = 0$ とした場合の推定値)がデータから得られれば、それをを用いてpower-of-the-mean modelに基づいて新たな固定効果およびべきパラメタの推定値を得ることができる。その結果を再び用いて新たな推定値を得ることができ、この過程を反復すれば、より正確な推定値を得ることができる。

6.2 数値例

話を具体的にするため、ここでプログラム5で与えられる数値例を用いてpower-of-the-mean modelを例示する。

プログラム5

```
data doseres;
  input dose @;
  do i=1 to 10;
    input res @;
    output;
  end;
  keep dose res;
cards;
  1  9.2  6.8 10.0 12.4  9.2 11.6 10.9  7.1 12.9  5.9
  2 23.9 24.8 23.9 19.6 18.3 12.7 10.7 18.4 17.1 21.1
  3 26.0 22.5 36.9 27.8 29.0 30.8 23.3 39.8 29.9 16.1
  4 44.6 47.7 30.2 55.4 18.8 40.0 39.4 55.5 28.4 38.0
;
run;
```

このデータはDOSEとRESという2つの変数からなる。我々はRESがDOSEにどのように依存しているかを知りたいとする。図3はこのデータをプロットしたものであり、RESはDOSEと共に直線的に増加しているが、RESの分散もDOSEと共に増加していることを示している。このような分散の特徴からpower-of-the-mean modelがデータに当てはまることが示唆される。

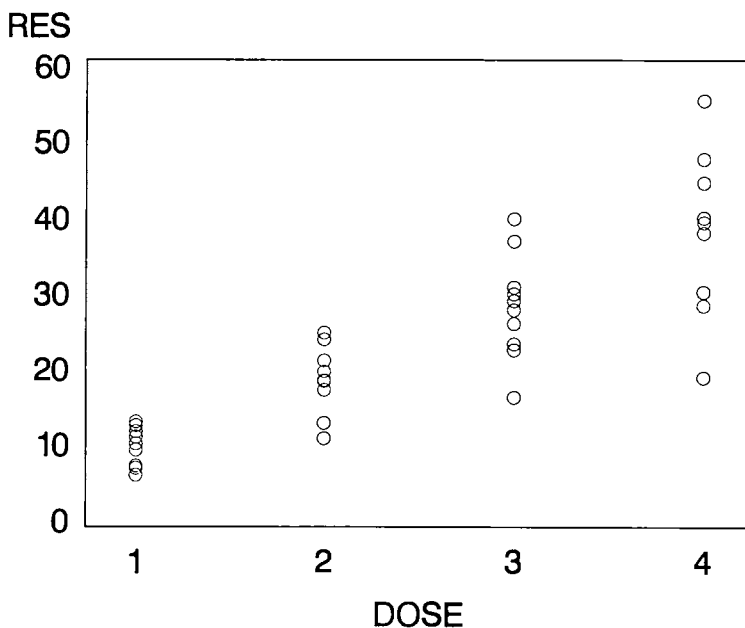


図4. Power-of-the-mean modelの数値例のプロット

6.3 反復計算のためのSASプログラム

プログラム6はMIXEDプロシジャを用いてpower-of-the-mean modelに基づいてこのデータを解析する際の基本的な考え方を示している。このプログラムは2つのステップからなっている。

最初のステップは予備的なステップであり、power-of-the-mean modelに必要な回帰パラメタの値を提供するため、MIXEDプロシジャを用いて、等分散を仮定した回帰分析を行うものである。そして得られた回帰パラメタ推定値はデータセットsolflに収められる。

次のステップのMIXEDプロシジャにおいては「repeated/ local=pom(solfl);」が指定されており、よってpower-of-the-mean modelに基づく推定がなされる。その際、前のステップで計算された予備的な回帰パラメタの値を用いて新たに回帰パラメタを求め、それをデータセットsolfl2に収めている。されにこれを用いて次のステップでもう一度MIXEDプロシジャを実行することができる。このようなステップを反復し、回帰パラメタの値が収束すれば、それが我々が得たい推定値である。

しかし、このような反復を手作業でするのは面倒であり、自動的に行えることが望ましい。そのような自動的な反復計算は前節に述べた方法を用いれば容易に行うことが可能となる。そのようなプログラム示したのがプログラム7である。

プログラム7の最初の%LETステートメントでは、回帰係数推定値を記憶しておくマクロ変数®に初期値1を与えている。このマクロ変数は、後のステップで、回帰係数推定値の収束の判定に用いられることになる。

マクロ・プログラム pom は MIXEDプロシジャおよびDATAステップの2つのステップからできている。

マクロ・プログラム pom はマクロ・パラメタ first を持っており、その値が1である場合は最初のサイクルであることを表している。その場合には、%IFステートメントの条件が「真」とならず、よってMIXEDプロシジャにおいて

プログラム6

```
ods output solutionf=solf1;
proc mixed data=doseres;
  model res=dose / s;
run;

ods output solutionf=solf2;
proc mixed data=doseres;
  model res=dose / s;
  repeated / local=pom(solfl);
run;
...
```

プログラム7

```
%let reg=1;

%macro pom(first);
  ods output solutionf=solf2;
  proc mixed data=doseres;
    model res=dose / s;
    %if &first=1 %then repeated / local=pom(solfl);;
  run;
  data solfl;
    set solfl;
    if effect='dose'
      and abs(estimate - &reg)>1e-8 then do;
      call symput('reg', left(put(estimate,e17.10)));
      call execute('%pom();');
    end;
  run;
%mend;

%pom(1);
```

REPEATEDステートメントは指定ず、等分散モデルが仮定されることになる。&firstの値が1でない場合には「repeated/ local=pom(solfl);」が指定されることになり、power-of-the-mean modelによる計算が行われる。MIXEDプロシジャによって得られた回帰パラメタ推定値はデータセットsolfl2に収められる。

次のデータ・ステップでは、回帰係数の収束判定がなされる。今計算した回帰係数の値をデータセットsolfl2から読み取り、マクロ変数®として記憶されている前サイクルの値との比較を行い、もしその差の絶対値が 10^{-8} よりも大きければ、まだ収束に達していないものとみなし、次のサイクルの準備を行う。すなわち、CALL SYMPUTを用いてマクロ変数を新しい回帰係数の値で置き換える。そして、「call execute('%pom(');)」によりマクロの再帰呼び出しを行い、次のサイクルに入る。ただし、その際にはマクロに引数は指定されておらず、よって次のステップではpower-of-the-mean modelが指定されることになる。一方、もし差の絶対値が 10^{-8} 未満であった場合には収束したものとみなされ、新たにマクロは呼び出されず、反復は終了する。

最後の行はマクロ・プログラムpomを最初に呼び出すためのものである。

表1はこの数値例における反復計算の過程を示したものである。2番目の列は各反復における回帰係数の値を示し、3番目の列は反復間の回帰係数の値の差を示している。回帰係数の値は6回目のサイクルで 10^{-8} 未満となっている。最後の列はべきパラメタ推定値を示しており、これも収束しているようである。

表1. Power-of-the-mean modelの数値例の計算結果

反復	回帰係数	変化	べきパラメタ(θ)
1	9.9760000000	-	-
2	9.6919601675	-0.2840398325	2.0816386164
3	9.6898237032	-0.0021364643	2.1587574892
4	9.6898171411	-0.000065621	2.1590014405
5	9.6898171209	-0.000000202	2.1590021894
6	9.6898171209	-0.000000000	2.1590021917

7. 一般的なアルゴリズム

前節で紹介した反復計算のためのアルゴリズムをより一般的な形で模式的に表せばプログラム9のようになる。

ここに示したように、CALL EXECUTEを用いた反復計算では、マクロ変数を用いて反復を制御する必要があり、その初期値の設定には%LETステートメント、その値の更新にはCALL SYMPUTが必要となる。

プログラム8

```
%let マクロ変数 = ...;          ← マクロ変数に対する初期値の設定

%macro マクロ名;
  ...
  data ...;
  ...
  if (さらに反復すべきか?) then do; ← 前サイクルの結果を記憶しているマクロ変数と新
    しいサイクルの結果との比較による収束判定
    call symput('マクロ変数', 変数); ← 新しいサイクルの結果によってマクロ変数を更新
    call execute('マクロ名');       ← マクロの再帰呼び出し
  end;
  run;
%mend;

% マクロ名;                     ← マクロの最初の呼び出し
```

8. まとめ

本発表ではCALL EXECUTEを用いたマクロの再帰呼び出しについて提案した。このテクニックは特に統計学における反復計算にとって非常に強力な道具である。

反復計算はDATAステップやSAS/IMLのDOループによっても可能であるが、その場合には計算アルゴリズムの全てのプログラムコードを書かなければならない。よって、DOループによる反復はアルゴリズムが単純な場合に限定される。

一方、マクロの再帰呼び出しでは強力なSASの種々のプロシジャを反復して利用することができる。この点がこのテクニックの最も有用な点である。例えば、先述の数値例ではMIXEDプロシジャを反復実行させた。統計では種々の問題に対して反復計算がなされる。例えば、非線形モデルに対する推定、不完全データに対するEMアルゴリズム(Dempster et al., 1977)などがその代表的な例である。その中には、既存のSASのプロシジャを組み合わせることで反復することにより、その計算が遂行可能なものがあると考えられる。そのような場合には、マクロの再帰呼び出しによる反復のテクニックが非常に有用であると考えられる。

引用文献

- Benjamin, W. E. Jr. (1999), A Pseudo-Recursive SAS Macro, Observation, 07MAY1999, obswww18. ([http:// support.sas.com/documentation/periodicals/obs/obswww18/index.html](http://support.sas.com/documentation/periodicals/obs/obswww18/index.html))
- Carroll, R. J., Ruppert, D. (1988), Transformation and weighting in regression, Chapman & Hall

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1-38.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. (1996), *SAS System for Mixed Models*, SAS Institute Inc., Cary, NC

Riba, S. D., (1997), Self-modifying SAS programs: a DATA step interface. *Obsevation*, 07SEP1997, obswww03
(<http://support.sas.com/documentation/periodicals/obs/obswww03/toc.html>)

日本SASユーザー会（SUGI-J）

SAS未経験者をSAS内部構造を理解したDATA ステップSASプログラマに 短期間で育成するカリキュラムの紹介

○山田 大志、小澤 康彦、宮浦 千香子
アストラゼネカ株式会社
臨床統計・プログラミング部

Curriculum for SAS beginner's DATA step programming course " *Understanding SAS's inner structure* "

Tajji Yamada, Yasuhiko Ozawa, Chikako Miyaura
Statistics & Programming Department, AstraZeneca K.K.

要 旨

我々は、SAS未経験者を対象とした教育カリキュラムを作成する上で、プログラムデータベクトルのコンセプトなどSAS内部構造を理解してDATA ステップの仕組みを習得することが、ミスの少ない作業効率が良いSASプログラマになる近道であり、かつ学習意欲の向上が期待できるであろうと考えた。本稿では、SAS内部構造を理解したDATA ステップSASプログラマを短期間で育成することを目的として作成した教育カリキュラム、及びその実施例を紹介する。

キーワード： DATA ステップ、プログラムデータベクトル

1. はじめに

弊社では業務体系の変更に伴い、それまでプログラミンググループが実施していた臨床データ論理チェック用のSASプログラム作成業務を、別の部署が実施することとなった。しかし、そのスタッフの多くはSAS未経験者であり、かつ膨大な日常業務のため十分な教育時間を確保することが困難な状況であった。

SAS未経験者に対するSASプログラミングの基礎教育を、どのようにして効率的に実施するかは各社共通の悩みであろう。弊社でも教育時間が充分取れないため、とりあえず業務に必要な幾つかのSAS言語を、「呪文」のように扱い教育しようとしていた。しかし、教育カリキュラムを遂行するうちにSAS言語を「呪文」として教える事に限界を感じていた。

本稿では、最初に我々がなぜSAS内部構造を教えようと思ったか、経緯を述べる。次に、実際にセミナーで使用したプログラムデータベクトルの説明資料を紹介する。更に、セミナー終了後に実施した卒業試験の問題例とその結果を紹介し、最後にSAS内部構造の教育が実際の業務にどのような影響を与えたかを考察する。

2. SASセミナーの開始

「臨床データ論理チェック用のSASプログラム作成業務が実施できるSASプログラマを、短期間で育成して欲しい」との依頼に基づいて、SAS未経験者を対象とした「SAS DATA ステップセミナー」を計画し始めたのは2003年2月であった。我々は当初、依頼に基づき既存のマニュアルを元にSASの簡単な使い方を教える程度のセミナーを計画した。しかし、業務に必要と思われる内容を考えるうちに、既存のマニュアルでは紹介している範囲が広く、初心者が短期間でその全てを理解し実際にプログラムを作成するのは困難であり、また本来の目的である「短期間で業務に必要な情報を習得する」を達成するには適切ではないと考えた。

よって、セミナー内容を原点から見直し、多くのSASの機能から今回の業務に必要なものを選択しながら、約30ページのオリジナルテキストを独自に作成した。このオリジナルテキストを基準に全体的なセミナースケジュールを作成し、1回1時間、全7回のセミナーを講義形式で計画した。セミナースケジュールを表1に示す。

表1 セミナースケジュール

第1回	SAS概要 SASシステムとは・・・ マニュアルの紹介 プログラムの表記ルール プログラムの構成 ライブラリの割り当て システムオプション SASログ/エラー
第2回	DATA ステップ ① DATA ステップ処理の流れ SAS データセット グローバルステートメント (title/footnote/options) 変数 (型/長さ/フォーマット/インフォーマット/ラベル)
第3回	DATA ステップ ② setを使ったデータ処理 基本的なステートメント (keep/drop/rename/output) 変数の追加 (割り当て/合計) SAS演算子
第4回	DATA ステップ ③ if, selectを使用した条件分岐 doループ処理 オブザベーションの削除 (where/delete/サブセット化IF)
第5回	プログラミング基礎 ① 基本的なプロシジャ (sort/print/format) mergeを使ったデータ処理
第6回	プログラミング基礎 ② グループ処理 (first.by/last.by) 複数データセット結合後処理 (in)
第7回	プログラミング基礎 ③ 基本的な関数の紹介及び使用方法 (put/input/sum/substr/round) データセットオプション

3. 呪文の限界

セミナー受講者は16名であった。セミナーにはSAS 8.2を用いた。参加者のほとんどはSAS未経験者であり、最初はSASの起動方法、用語の説明から始めた。第1回、第2回はSASの一般的な話で、それほど問題もなくスムーズにセミナーは進んだが、第3回の「基本的なステートメント」から、多種多様な質問が頻繁に出てくるようになった。その質問の多くは「テキストに書いてあることは理解できるが、プログラムの一部を変更した時に、どのような結果になるかがわからない」といった類のものであった。我々はプロジェクターを使用して、サンプルプログラムを受講者に見せながら講義を行っていたが、1つ1つの質問に対して実際にプログラムの一部を変更し、その実行結果を表示しながら「プログラムをこう変更すれば、このような結果になる」というように、SAS言語を「呪文」のように教えた。すると、皆一応は納得した様子であったが、逆に我々から「では、この場合はどうなるか」と応用問題を投げ掛けると、答えに詰まる状況であった。我々は、SAS言語を「呪文」のように教える限界を感じていた。

4. セミナー方針の見直し

当初の目的をスムーズに達成できない原因として、我々はSAS言語を「呪文」のように教えるだけでは応用力が身につかないからであると考えた。「応用力のあるプログラマ」、それこそ我々が考える真のプログラマの姿である。

セミナー方針を見直していた時、我々自身が2002年12月にSAS認定プロフェッショナルプログラムを受験する際に自己学習した、「プログラムデータベクトル」のことを思い出した。プログラムデータベクトルとは、SASシステムがオブザベーションのデータ値を処理するために使う一時的なメモリ領域である(図1 DATA ステップ処理の流れ 参照)。我々自身、それまでSASの内部構造を特に意識することなくSASプログラムを作成していたが、実際に内部構造を知ることによって応用力が付き、業務においてもミスが少なく、効率良くプログラムが作成できるようになった。

我々はこのような経験から、「呪文」のようにSAS言語を覚えるのではなく、プログラムデータベクトルのコンセプトなど、SAS内部構造を理解した上でDATA ステップの仕組みを習得することが、応用力のあるSASプログラマになる近道であろうと考えた。

では、実際にSAS内部構造を説明する場合、まず考えたのは何から説明するかということであった。ある程度の経験をもったSASプログラマに対する講義とは異なり、初心者に対して講義を行う場合、セミナー中に用いる用語の一つ一つにも十分な注意が必要となる。セミナー方針を見直していた時も、「これを説明するためにはその前にこちらを・・・」というような堂々巡りの議論が、我々セミナー講師陣の間でしばしば交された。そこで我々が出した結論は、セミナー中盤でそれまでの内容を反映した練習問題を受講者に配布し、実際にプログラムをさせることで、各自が漠然と抱いていた不明点をまず明らかにする。次に、通常セミナー以外にその問題解説の時間を設け、そこで「プログラムデータベクトル」という本来外側からは見ることができない、SASの内部構造を交えたDATA ステップの説明を行うというものであった。

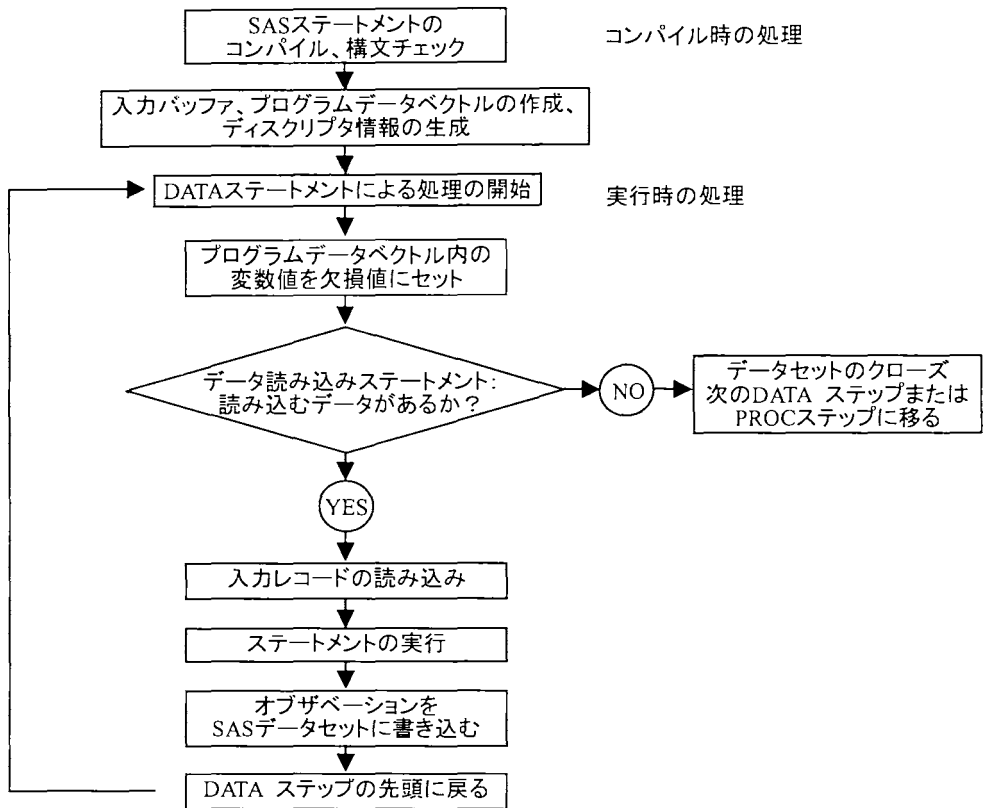


図1 DATA ステップ処理の流れ (SASランゲージ リファレンスより引用)

SAS内部構造を説明するために作成した演習問題の一部を表2に示す。

表2 演習問題

《 LIBRARY 》ライブラリにあるデータセット [DEM] を読み込み、以下の処理を実行して下さい。
 ・以下の 2 変数を新規変数として設定して下さい。

変数名	型/長さ	ラベル
DAY	数値/8 バイト	Day number
COUNT	数値/8 バイト	OBS No.

・現在の各症例 1 オブザベーションの [DEM] データセットから、新規変数 [DAY] の値が 「1」「8」「15」「22」となるような各症例 4 オブザベーションの [DEM_NEW] データセットを WORK ライブラリに作成して下さい。
 出力変数(順不同): [SUBJECT] [PATIENT] [SEX] [BIRTHDAT] [DAY] [COUNT]
 ・新規変数 [COUNT] に全オブザベーションを通して連番をつけてください。

[DEM] データセット

	FNAME	LNAME	SUBJECT	PATIENT	BIRTHDAT	SEX
1	YASUHIKO	OZAWA	E0001004	101	1980-01-31	1
2	TAIJI	YAMADA	E0001010	102	1979-05-05	1

↓

[DEM_NEW] データセット

	SUBJECT	PATIENT	BIRTHDAT	SEX	DAY	COUNT
1	E0001004	101	1980-01-31	1	1	1
2	E0001004	101	1980-01-31	1	8	2
3	E0001004	101	1980-01-31	1	15	3
4	E0001004	101	1980-01-31	1	22	4
5	E0001010	102	1979-05-05	1	1	5
6	E0001010	102	1979-05-05	1	8	6
7	E0001010	102	1979-05-05	1	15	7
8	E0001010	102	1979-05-05	1	22	8

この演習問題はdoステートメント、outputステートメントの働きを正確に理解させるために作成した問題であり、2オブザベーションのSAS データセットを、DATA ステップを使用して8オブザベーションにするというプログラムが要求されている。この演習問題は熟練したSASプログラマの場合、たとえSAS内部構造を知らなくても、過去の多くの経験から正しいプログラムを作成することは可能であろう。しかし、SASの経験が乏しい場合、作成は非常に困難である。なぜなら、SAS初心者は自分の作成したプログラムがどのような結果を出力するのかということが、プログラム記述段階では自分の中で明確ではないからである。我々は、作成したプログラムがSAS内部でどのように処理され、どのように目的とする結果が生み出されているのか、というプログラムの基礎となるべき部分を教える必要があると考えた。

次に我々が考えたことは、どのように説明すればSAS内部構造を受講者に正確に理解させることができるか、ということであった。ただ、マニュアルに記載されていることをそのまま読んで説明するだけでは、SAS未経験の受講者が充分内容を理解することは困難であろう。そこで我々は、プログラムがサブミットされてから、コンパイル、プログラムデータベクトルの作成、DATA ステップの開始、プログラムデータベクトルへの初期値セットなど、SAS内部の動きを順に追ったPowerPointファイルを作成した。

問題解説時にスライドショーを実行しながら説明を行ったところ、それまでセミナー中はほとんど反応がなかった受講者からも質問がくるといった、明らかにこれまでとは違った雰囲気が受講者の中に流れていた。受講者全員が一度の解説で全てを理解できた訳ではないが、スライドショーを見ながら説明を聞くことで、DATA ステップの流れ及びSAS内部構造に対する理解はそれまでよりは確実に深まったと考えられる。実際、ある受講者からは「セミナー後に自分でスライドショーを見ながら、何度もプログラムを実行していくうちに、なぜこんなデータセットが出力されたのかが理解できた」という意見もあった。

PowerPointファイルでの説明手順を表3に示す。

表3 PowerPointファイルでの説明手順

【DEM】データセット

STUDY	CENTRE	SUBJECT	PATIENT
AAA		1E0001004	101
AAA		1E0001010	102



Program 例

```

data DEM_NEW;
  set DEM;
  do I = 1 to 4;
    if I = 1 then DY = 1;
    else DY = DY + 7;
    COUNT + 1;
    output;
  end;
  rename DY = DAY COUNT = CNO;
  keep PATIENT DY COUNT;
run;
    
```

□ の部分はコンパイル時に実行済みのため、DATA ステップループでは実行されない

プログラムデータベクトル ②

STUDY	CENTRE	SUBJECT	PATIENT	I	DY	COUNT	_N_	_ERROR_
		0	1	0
D	D	D	K	D	K	K	D	D

【DEM_NEW】データセット

PATIENT	DAY	CNO
101	1	1
101	8	2
101	15	3
101	22	4
102	1	5
102	8	6
102	15	7
102	22	8

- ① 構文チェック、コンパイル (SASステートメントを実行可能なマシンコードに変換する)
- ② プログラムデータベクトル (DATA ステップで使用する全ての変数について、SASシステムがオブザベーションのデータ値を処理するための一時的なメモリ領域) の作成
→ [DEM] データセット内の変数/自動変数/DATA ステップで使用する変数の割り当て
- ③ 情報ステートメント (rename/keep) の内容を取得
→ 出力データセット [DEM_NEW] に影響する
- ④ dataステートメントによるDATA ステップ処理の開始
→ プログラムデータベクトル内の変数に、初期値がセットされる
データセット変数 or 割り当て変数欠損値
合計変数 or 自動変数 [_ERROR_] ...0
自動変数 [_N_]1
- ⑤ DATA ステップ内のステートメントの実行
→ setステートメントにより1オブザベーション分のデータをプログラムデータベクトルへ読み込む
→ doループや合計ステートメントによりプログラムデータベクトル上の各変数値が変化する
→ outputステートメントによりプログラムデータベクトル上のデータを [DEM_NEW] データセットへ書き込む
→ [DEM] データセットのオブザベーション数だけ、DATA ステップループが行われる

5. 卒業試験

セミナー終了後の2003年3月に卒業試験を実施し、セミナーの理解度を調査した。卒業試験は60分間、100点満点とした。卒業試験問題は予め複数のSAS経験者に模擬的に試験を受講してもらい、試験の難易度を調整した。卒業試験にはセミナー受講者だけでなく、SAS経験者も含めた20名が参加した。試験内容の要約を表4に示す。

なお、テスト終了後、数名に対しては更なる教育が必要であると判断し、新たな演習問題による自己学習の後、追試を実施した。追試問題は先の卒業試験と同じSAS経験者に模擬的に試験を受講してもらい、前回の試験と同程度の難易度とした。その結果、ほとんどの受講者が合格ラインを超えた。

表4 卒業試験内容の要約

問題番号	要約	キーワード
1	SASの基礎知識を問う文章問題（選択肢あり）	エラーの種類、変数等の命名規
2	SASの基礎知識を問う文章問題（選択肢なし）	則、SAS日付値、算術・比較・論理
3	SASの基礎知識を問う文章問題（3択）	演算子、グローバルステートメント
4	SASの基礎知識を問う文章問題（○×問題）	等
5	SASプログラムを与え、正しい実行結果を問う問題（一部選択肢あり）	合計ステートメント、サブセット化 IFステートメント、selectステートメント等
6	SASデータセットと実行結果（Output画面）を与え、正しいSASプログラムを問う穴埋め問題	データセットオプション、outputステートメント等
7	SASデータセットとSASプログラムを与え、正しい実行結果（Output画面）を問う穴埋め問題	フォーマット
8	SASデータセットと実行結果（SASデータセット）を与え、正しいSASプログラムを問う穴埋め問題	first.by変数、last.by変数
9	SASデータセットとSASプログラムの目的（年齢計算）を与え、正しいSASプログラムを問う穴埋め問題	データセットオプション、substr関数、input関数
10	SASデータセットと実行結果（SASデータセット）を与え、正しいSASプログラムを問うプログラム記述問題	lengthステートメント、反復doステートメント、文字結合等

6. 卒業試験問題例

卒業試験問題として出題した問題の一部を表5、6に示す。

表5 卒業試験問題 1

次のプログラムを実行し、作成されるSASデータセット【TEST5】の変数[X]の値を記述して下さい。

```
data TEST5 ;  
  S = 1 ;  
  do X = 1 to 7 by 2 ;  
    S + 1 ;  
    X + S ;  
  end ;  
run ;
```

表5の問題は、インデックス変数 X が反復doループ内で再計算されるプログラムであり、インデックス変数の値がカウントアップされる場所、及びカウントアップ後に反復doループの条件が判定されるというSAS内部処理を正確に把握していない場合、正答は困難である。

SASの反復doループ処理は、endステートメントが実行された時に、byステートメントで指定した数だけインデックス変数のカウントアップを行い、その上で反復doループの条件判定を行うため、インデックス変数の最終的な値は必ず反復doループ条件の最終値を超えることになる。この問題では、2度目のループ時に X が 8 となり、endステートメントで X を 10 にカウントアップしてから条件式の判定が行われ、その結果FALSEとなり反復doループを抜ける。そして、runステートメントの直前の「暗黙のoutputステートメント」によりSAS データセット[TEST5]の変数 X に 10 が出力される。よって、正解は "10" となる。

不正解の中で最も多かった解答は、"9" であった。この原因としては、インデックス変数の値が反復doループ条件の最終値を超えた場合は、byステートメントの引数に関係なく 1 が加算されると考えたのではないかと推測する。しかしながら、"9" と答えた受講者たちは、少なくともカウントアップしてから条件判定を行うというSAS内部処理は理解できているであろうと考えられる。また、解答を "8" と答えた受講者もいた。これはおそらくカウントアップ後に反復doループの条件が判定されることを理解できていなかったためだと考えられる。

表6 卒業試験問題 2

次のプログラムを実行した結果はどうなるでしょう？選択肢の中から選んでください。
但し、データセット【VIT】は変数[CTEMPORA][WEIGHT]のみが存在することとします。

```
data TEST6 ;
  set VIT ;
  rename CTEMPORA = CTMP ;
  format WEIGHT 8.1 ;
  label CTMP = 'TEMP (C)' ;
run ;
```

A：フォーマットは設定されない B：ラベルは設定されない C：全て問題なく設定される

表6の問題は、プログラムデータベクトルの存在、及びDATA ステップループでは実行されない情報ステートメントがいつどのように実行されるか、というようなSAS内部構造を問うた問題であり、その内部構造が正確に理解できていない場合、正答が困難であると思われる。

SASには、プログラムデータベクトルやデータセット内の変数の属性に関する情報をSASシステムに与えるための「情報ステートメント」と呼ばれるステートメントが存在する。代表的なものとして、keep / drop / label / format / rename / retainなどがある。この情報ステートメントはDATA ステップループ内では実行されず、コンパイル時にプログラムデータベクトルに情報として記憶される。

この問題では、データセット変数の変数名を変更するためのrenameステートメント、変数ラベルを設定するためのlabelステートメント、フォーマットを設定するためのformatステートメントという3種類の情報ステートメントが含まれている。SAS内部構造を認識せずにこのプログラムを考えた場合、変数名を変更した後、その変更された変数名に対してラベルを設定すると考え、“C”を正答としてしまうかもしれない。しかし、renameステートメントはプログラムデータベクトル内に存在する変数の変数名を実際に変更する訳ではなく、変更の情報だけをプログラムデータベクトル内に保持し、結果として出力データセットの変数名を変更する。このため、labelステートメントで変更後の変数名 [CTMP] を指定した場合、プログラムデータベクトル内には [CTMP] は存在しないことから、LOG画面に「NOTE: 変数 CTMP は初期化されていません。」というメッセージが表示され、labelステートメントは正しく設定されない。よって、正解は”B”となる。

この問題は、試験対象者の約9割が正答であった。このことから、我々が必要であると判断しカリキュラムに組み込んだ、SAS内部構造の説明について成果が見られたと考えられる。

7. まとめ

我々は、セミナー実施を計画してから約1ヶ月間で、ほとんどのSAS未経験者をSAS内部構造を理解した初級DATA ステップSASプログラマに育成することができた。

本セミナー後、受講者の何人かが実際の臨床データ論理チェック用のSASプログラム作成業務を行った。ある受講者から「実際に業務でSASを使用して、予想とはちがう結果が出たときにSAS内部構造を思い出すとエラーの原因が容易に分かった」との声が聞かれた。実際、SASの内部構造だけを学んでも、それを実践に生かせなければ意味はない。我々は、ケアレスミスを減少させ、効率的なプログラムの作成を促進し、かつ自分の力で更なるSASプログラミング技術を身につける上で、SASの内部構造を理解することは非常に重要であり、また、結果的に実際の業務において非常に有用であると確信している。

他の受講者からは「最初からこの作業に必要な方法のみ教えてくれればすぐに作業ができるようになったのに」とのコメントを頂いた。我々はSASを学ぶ上で、最初にプログラムデータベクトルを学ぶことが必須であるとは思っていない。しかし、最初に「呪文」のみ教えると、後でプログラムデータベクトルについて学ぶ気になったのか、また将来的に応用力のあるプログラマに早く成長できるのか、どうかについては疑問である。

確かにSASの内部構造がブラックボックスでも、ある程度の業務に耐えうるプログラムを作成する能力は取得可能であろう。しかし、それだけではSASプログラマの実力を上げるために多くの「呪文」を覚える必要があり、また、経験という名の恐ろしく時間がかかる方法でしか成長できない。プログラムデータベクトルなどのSAS内部構造を学べば、SAS言語を幾つかのパターンに分類でき、その後の自主学习もスムーズにいくと考える。

また、受講者の多くがSASに対して興味を持ったのも事実である。ただ単に業務をこなすためにSASの使い方を訓練するだけでなく、学問としてSAS内部構造を学ぶ喜びを知ることが、高いレベルのSASプログラマに成長する近道であると考えます。

今回は、SAS未経験者を対象としてSAS内部構造を教えた例を紹介したが、もちろんSAS経験者にもSAS内部構造の理解は有用である。もし、プログラムデータベクトルをご存知ない方がいらしたら、是非マニュアルを紐解いて欲しい。筆者自身、SASを使用して10年目に初めてプログラムデータベクトルについて勉強したことで、飛躍的にSASプログラマの実力が上がったと実感している。

最後に、今回作成したセミナーのテキスト、演習、卒業問題及びその解答例等の教育カリキュラムは、まとめて広く皆様にご紹介できればと思っている。

参考文献

1. SASランゲージ リファレンス Version 6 First Edition (1995), SAS出版局
2. Step-by-Step Programming with Base SAS® Software (2001), SAS Institute Inc.
3. Base SASソフトウェア使用法ガイド Version 6 First Edition (1993), SAS出版局
4. Robert Virgile, An Array of Challenges - Test your SAS® Skills (1996), SAS Institute Inc.

日本SASユーザー会 (SUGI-J)

Microsoft Access と SAS によるデータマネジメントシステム

○中村 竜児 松沢 享
メディカル統計株式会社

The data management system by Microsoft Access cooperating with SAS

Ryoji Nakamura / Akira Matsuzawa
MEDICAL TOUKEI CO,LTD.

要 旨

臨床試験のデータマネジメント業務を Microsoft Access で管理することにより、入力画面や出力書式など比較的 SAS が苦手とする部分を補完するとともに、SAS プログラムを自動作成する方法を検討したので報告する。

キーワード： COMPARE プロシジャ、SQL プロシジャ、ODBC、Microsoft Access

1 はじめに

当社では、ダブルエントリーによるデータ入力業務を行っているがコンペアリストをより見やすくすることと SAS プログラミングの自動化と標準化をすることがかねてより課題となっていた。そこで Microsoft Access のテーブル上に SAS データセットのコンテンツ情報を入力することでこれらの問題を解決する簡単なシステムを作成したのでその機能の概要を紹介する。

2 データベース定義

2.1 データセット情報の定義

作成する症例 SAS データセットについて、SAS データセット名・日本語名・ID 変数名を設定し(図 1)、定義されたデータセット毎に変数を定義する(図 2)。定義作成後、変数名の重複や SAS フォーマットの型に矛盾がないかどうかチェックを行い、問題がなければ DB 定義書を出力する(図 3)。

解析用データセットについてもここで同様に定義を行うが、これは主に解析用 DB 定義書を作成することが目的であり、解析用データセット作成のためのプログラミング機能については今後の検討課題である。

2.2 入力画面の作成

入力は転記シートに転記したものをスキャナで読み取りテキストデータ化してから SAS データセットを作成する方法と、Microsoft Access のテーブルに対してパンチ入力を行い、入力完了後 CSV または ODBC 経由で SAS データセット化する方法をとっている。転記入力の場合は変数定義の LENGTH をもとに SAS DATASTEP 文の INPUT ステートメントで入力カラム位置を指定するプログラムを作成する。パンチ入力の場合には変数定義に基づき入力用 Microsoft Access mde と入力画面を作成する(図 4)。

2.3 SAS プログラムの作成

SAS データセット化された以降の編集については Microsoft Access のクエリー機能等を使うのではなく、あくまでも SAS プログラムを記述して行うことになる。予め指定しておいたフォルダパスに SAS プログラムが書き出される(図 5)。

2.4 フォーマット情報

変数定義で指定された SAS フォーマット名をもとに、フォーマット情報入力用テーブルを作成しフォーマット情報を入力する(図 6)。入力されたフォーマット情報は ODBC 経由で SAS データセット化し、FORMAT プロシジャの CTRL IN オプションを利用して SAS フォーマットカタログ化する。

2.5 ロジカルチェック定義の作成

ロジカルチェック定義は日本語のエラー内容とプログラム上の論理文を変数に対等させて入力をする(図 7)。入力された内容をもとにロジカルチェック基準書とロジカルチェックプログラムを出力する。ロジカルチェックプログラムの修正もこの画面上で行えるので、プログラムエディタで修正するよりも確認が容易である。

No	テーブル名	入力種類	マルチ	SAS データセット名	備考
0	全件として	全件仕様	☐	ID ID変数名1 ID変数名2	
1	背景	通常転記	☐	HAI NUM	
2	合併症	通常転記	☑	GAP NUM ID	
3	臨床検査値	通常転記	☑	RIN NUM RNUM	
4	使用状況	通常転記	☐	HON NUM	
5	コメント	コメント	☑	CMT NUM CNUM	
6	併用薬投与状況報告	計算加工	☑	DRUG NUM RNUM	
*			☐		

図 1

背景

入力種類 通常転記 マルチレコード

SAS変数名1 ID変数名2 NUM

備考:

順番	Page	項目名	型	SAS変数名	備考	シート	開始	終了
1	1	整理番号	数値	num	番号	1	1	5
2	3	群別	数値	gun	1:A群 2:B群	1	6	6
3	2	施設名医師記載内容①	文字	size	施設名コード参照	1	7	16
4	1	医師名1	文字	dr1	医師名コード参照	1	17	19
5	1	医師名2	文字	dr2	医師名コード参照	1	20	22
6	2	症状	文字	hoyo	症状コード参照	1	23	25
7	2	症状程度	文字	hstei	症状程度コード参照	1	26	26
8	2	褥瘡安全度	数値	anzen	褥瘡安全度コード参照	1	27	27
9	2	重篤	数値	jutok	重篤度コード参照	1	28	28

レコード: 1 / 13

定義機構チェック コメント入力用データベース作成

シートカラム割 閉じる

図 2

1: 背景

入力種類 通常転記 SASデータセット HRI

No	Page	項目名	SAS変数名	属性	長さ	SASformat	備考
1	1	整理番号	num	数値	8		番号
2	3	群別	gun	数値	8	gun.	1:A群 2:B群
3	2	施設名医師記載内容①	size	文字	3	size.	施設名コード参照
4	1	医師名1	dr1	文字	3	dr.	医師名コード参照
5	1	医師名2	dr2	文字	3	dr.	医師名コード参照
6	2	症状	hoyo	文字	2	hoyo.	症状コード参照
7	2	症状程度	hstei	文字	2	hstei.	症状程度コード参照
8	2	褥瘡安全度	anzen	数値	8	anzen.	褥瘡安全度コード参照
9	2	重篤	jutok	数値	8	jutok.	重篤度コード参照
10	2	性別内容	sex	数値	8	sex.	性別コード参照
11	2	入院外来	goin	数値	8	goin.	
12	2	身長	shin	数値	8		
13	2	体重	wei	数値	8		

図 3

背景 0: フォーム

背景

整理番号

群別

施設名医師記載内容①

医師名1

医師名2

症状

症状程度

褥瘡安全度

重篤

性別内容

入院外来

身長

体重

レコード: 1 / 1

図 4

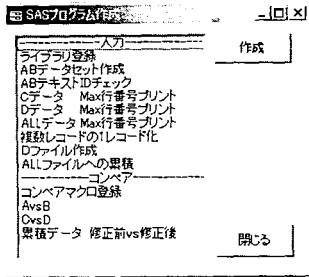


図 5

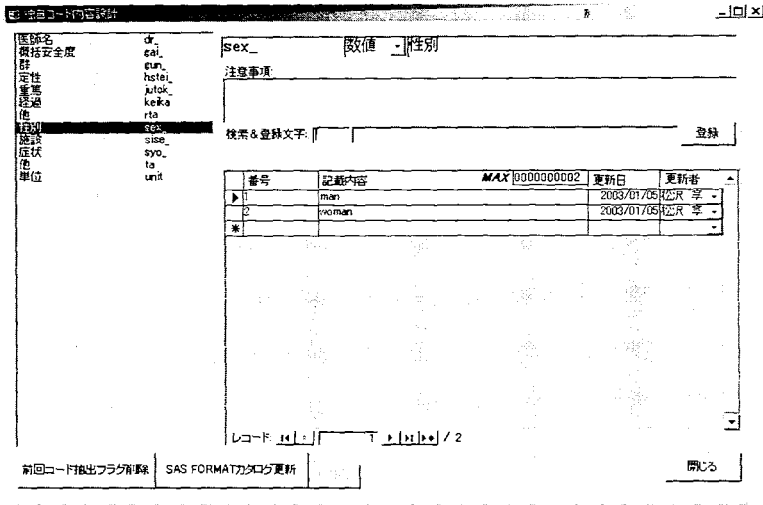


図 6

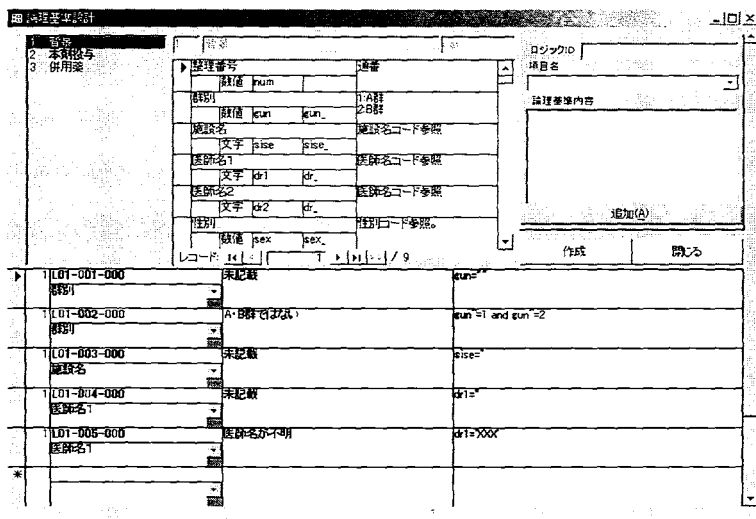


図 7

3 データマネジメント業務

3.1 業務の流れ

症例報告書コピーを受領したら入力ルーチンを登録し(図 8)、症例報告書コピーの記載内容から独自コード拾い出し等を行った後、ダブルエントリーにより入力を行い、出来上がった 2 群の SAS データセットをコンペアする。アンマッチの結果と症例報告書を付け合せ、正しい方の値を採用して新たなデータセットを作成し読合せ帳票を出力する。読合せ帳票と症例報告書を付け合せ校正を行い、入力ミスが認められれば修正を行う。修正確認は修正前後のデータセットのコンペアにより行う。

3.2 ダブルエントリーコンペア

ダブルで入力した SAS データセット化した後で両群データを比較するのだが、COMPARE プロシジャの出力は以下の点で見難いものであった。

- ① LENGTH が長い変数は値が途中で切られて出力される。
- ② アンマッチの内容は症例単位ではなく変数単位で区切って出力される。
- ③ SAS フォーマットをはりつけた数値変数についてはフォーマットの内容を反映して出力されるが、文字変数の場合は反映されない。

①、②については COMPARE プロシジャの OUT=オプションにより作成されるデータセットを編集することで解決できるが、③の問題は残ってしまう。文字変数についてもフォーマットの値を出力するだけでなく、元の値とフォーマットの内容を両方出力できることが望ましい。

そこで COMPARE プロシジャに比べれば多少実行速度が遅くなるが、DATASTEP によるコンペアを行っている。まず、CONTENTS プロシジャによりコンテンツ情報を SAS データセット化し、ID 変数以外の変数名とラベルを全て通番をふったマクロ変数に格納する。両群のデータセットを片方のデータセットの変数名に“_”を加えてマージし、マクロ変数に落とした変数名毎に両群で値が一致しないオブザベーションを抽出するループをまわす。サンプルプログラムを最後に載せておく。

プログラム実行の結果出来上がったデータセットを Microsoft Access に載せ、SAS データセットの日本語名等をひっぱりつけてきてレポート出力する(図 9)。

3.3 読合せ帳票の出力

本システムでは読合わせ帳票出力機能は持っておらず、症例一覧表作成ツール CATS(有限会社電助システムズ)を使用している。現状では CATS 用のレイアウトシートを作成するために CATS で指定している“@SAS 変数名(フォーマット名)”の文字列を Microsoft Excel に吐き出す機能を備えるのみであり、現在 DDE を利用した読合せ帳票出力プログラムの自動作成を検討中である。

3.4 データ修正

以前は SAS データセットに対して直接キーパンチで修正を行っていたが、修正処理を再現できるようにしておくため、PROC SQL を利用して修正用 SAS プログラムを記述して修正作業を行う。値の変更は PROC SQL の UPDATE 文節、オブザベーション挿入は INSERT 文節、削除は DELETE 文節を記述する。SAS プログラムや SQL が分からない人間でも修正プログラムを作成でき、修正内容の確認が日本語でできるようにするため図 10 のような修正画面を設け、この情報をもとに修正プログラムを作成する。

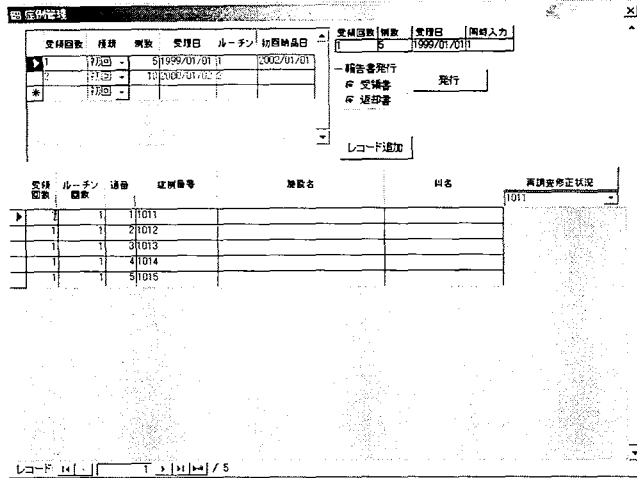


図 8

コンパリスト A-B

背景		【不一致回数:34】	
合併症		【不一致回数:23】	
臨床検査値		【不一致回数:0】	
使用状況		【不一致回数:0】	

症例番号	呼番号	A	B	既用数	既用不使用時
ファイル名	項目名	呼番号	呼番号	Record Table	修正内容
384					
合併症	終了時刻	2	16:05	16:00	
		3	16:05	16:00	
		4	07:05	07:00	
385					
背景	推込形態制御番号の	1		2	あり
	有無				
	カルテNo. (入院)	1		320887-9	
	科名	1	10	内科	
	検査番号	1	1	特別調査	

図 9

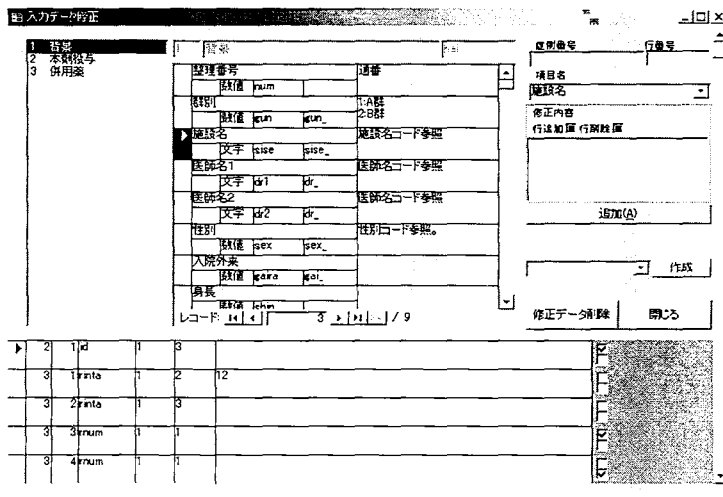


図 10

4 今後の検討課題

本システムは今後もバージョンアップを重ねていくがその際に課題となる点についてまとめておく。

- ① ODBC により SAS と Microsoft Access でデータをインポート／エクスポートすると、やりとりが終了しても SAS は実行状態のままである。そこからさらに SAS 上で何か作業をしようとする場合はプロセスを強制終了させなくてはならないが、その後さらに ODBC インポート／エクスポートすることができず、Microsoft Access を一旦終了させなくてはならない。これについては SAS/ACCESS ソフトウェアを導入するなどして解決したい。
- ② 読合せ帳票出力機能を追加する。
- ③ 解析用データセット作成プログラム作成機能と解析プログラム作成機能を追加する。
- ④ 現在ログオン画面からシステムに入るようにしているが、これは作業記録を残すためと、使える機能に制限を与えるためであるが、セキュリティについてはほとんど考慮されていない。機能追加の前にセキュリティ強化が先決である。

5 プログラムサンプル

5.1 コンペアプログラム

```
/*-----コンペアマクロ-----*/
data comp; delete;run;

%macro comp(tid,m,dt,id1,id2);/*データセット1、データセット2、id1、id2*/
proc sort data=&lib1..&dt.&D1 out=&dt._1; by &id1 &id2; run;
proc sort data=&lib2..&dt.&D2 out=&dt._2; by &id1 &id2; run;

proc contents data=&dt._1 out=_out noprint; run;
proc sort data=_out; by varnum; run;

data _out; set _out;                               /** 変数名の大文字に揃え、フォーマット
名を追加 **/
  if upcase(name) ^= upcase("&id1");
  if upcase(name) ^= upcase("&id2");
  if upcase(format) ^= 'YYMMDD' & upcase(format) ^= 'NENGO' & upcase(format) ^= 'TIME' & format ^= ''
then fflg=1;
  if type=1 & format='' then format='BEST8';
  if type=2 & format='' then format='$200';
  if upcase(format)='YYMMDD' then format='YYMMDD10';
  if upcase(format)='NENGO' then format='NENGO9';
  if upcase(format)='TIME' then format='TIME5';
run;

data _null_; set _out;
call symput(' flg' || left(_n_), compress(fflg));
call symput(' a' , left(_n_));                               /** 変数の総数を抽出 **/
call symput(' name' || left(_n_), compress(name));          /** 変数名を抽出 **/
call symput(' lab' || left(_n_), compress(label));          /** 変数名を抽出 **/
call symput(' fmt' || left(_n_), compress(format));         /** フォーマット抽出 **/
run;

data c2;
set &dt._2;
rename
```

```

%do i=1 %to &a;
  &&name&i = &&name&i...
%end;
;
run;
proc sort data=&dt._1 out=c1; by &id1 &id2; run;
proc sort data=c2; by &id1 &id2; run;

data c3;
keep tid nobsf compno comprno vname vlabel base comp based compd;
length tid 8. nobsf $1. compno comprno $20. vname vlabel base comp based compd $200.;
merge c1(in=in1) c2(in=in2);
by &id1 &id2;
%do i=1 %to &a;
  if &&name&i^=&&name&i... then do;
    if in1=0 and in2=1 then nobsf=1;
    if in1=1 and in2=0 then nobsf=2;
    compno=left(input(&id1, $20.));
    %if &m=1 %then comprno=left(input(&id2, $20.));;
    vname="&&name&i";
    vlabel="&&lab&i";
    base =left(input(&&name&i, $100.));
    comp =left(input(&&name&i..., $100.));
    based=left(put(&&name&i, &&fmt&i...));
    compd=left(put(&&name&i..., &&fmt&i...));
    if &&flg&i^=1 then do;
      base=""; comp="";
    end;
    output;
  end;
%end;
run;

data comp;
set comp c3(in=d);
if d then tid=&tid;
cat=&cat;
run;
%mend;

/* ACCESS搭載用データ作成 */
%macro comp2;
data comp;
set comp;
rename compno=num comprno=rnum vname=name base=oldd comp=newd based=old compd=new;
run;

data &lib1..comp&cat;
set comp;
run;

proc export dbms=csv data=&lib1..comp&cat outfile=&outfl replace;run;
%mend;

```

5.2 修正プログラム

```
/******SASデータセット修正プログラム*****/  
%macro up;  
proc sql;  
  update &lib..&dt.&d  
  set &v = &eq  
  where &w ;  
quit;  
%mend;  
  
%macro in;  
proc sql;  
  insert into &lib..&dt.&d.(&num.)  
  values(&nvar.);  
quit;  
%mend;  
  
%macro de;  
proc sql;  
  delete from &lib..&dt.&d  
  where &wdel ;  
quit;  
%mend;  
  
/******修正マクロ実行*****/  
%let lib=eril;  
%let d=_D;  
/***合併症***/  
%let dt=GAP;  
/*行番号*/  
%let num=num, id;  
%let nvar=1, "3";  
%in;  
/***臨床検査値***/  
%let dt=RIN;  
/*臨検検査項目*/  
%let v=rinta;  
%let eq=12;  
%let w=num="1" AND rnum=2;  
%up;  
/*臨検検査項目*/  
%let v=rinta;  
%let eq. ;  
%let w=num="1" AND rnum=3;  
%up;  
/*行番号*/  
%let num=num, rnum;  
%let nvar="1", 1;  
%in;  
/*行番号*/  
%let wdel=num="1" AND rnum=1;  
%de;
```

日本SASユーザー会 (SUGI-J)

SASによるメタデータマネジメント

○鹿渡 圭二郎 李 錦実 江口 英男 (訳)
カスタマーサービス本部プロフェッショナルサービス第1部
SAS Institute Japan 株式会社

SAS® Metadata, Authorization and Management Services — Working Together for You
Michelle Ryals
SAS Institute Inc., Cary, North Carolina

要 旨

企業内のリソースを十分に活用するために必要となるメタデータについて述べるとともに、SAS におけるメタデータ管理方法を Case Study を交えて説明する。
なお本書は、Michell Ryals 氏の SUGI28 での論文を翻訳したものである。

キーワード： メタデータ SAS 9.1 SAS Metadata Architecture

1 はじめに

あなたの会社ではどのようにリソースを管理していますか。機密情報へアクセスするユーザーを制限できていますか。データが必要な時、そのデータがどこにあるのか、またそれが使用できるのかをご存知ですか。あなたの会社ではメタデータを利用していますか。メタデータが企業に利益をもたらすということをご存知ですか。データを管理するために必要な情報を提供するものがメタデータであり、それはデータからインテリジェンスを生み出す際の鍵となります。ビジネスが成長するにつれてデータ量も増えます。データを正しく利用し競合企業に差をつけたいのであれば、メタデータはこれまで以上に重要になってきます。

この論文では、企業内のリソースを十分に活用するために必要となるメタデータについて述べるとともに、SAS におけるメタデータ管理方法を Case Study を交えて説明します。

2 メタデータ

メタデータは、よく「データのデータ」と説明されます。具体的には、データリソースについての情報のことであり、データの構造やその中身、データを利用するアプリケーションについての情報もその定義の中に含まれます。SAS が「知る力(The Power to Know)」を提供するのであれば、メタデータは

「理解する力(The Power to Understand)」を提供するといえます。

またメタデータを利用することで、必要な情報に素早くアクセスすることができます。データを管理することに労力を割くのではなく、本来のビジネスに集中するために、メタデータは必要不可欠です。

Break it down !

メタデータには、テクニカルメタデータとビジネスメタデータの次の2つの種類があります。

テクニカルメタデータは、IT 環境の構築、保守、管理をサポートします。物理的なストレージ構造、サーバーシステムやデータの加工プロセスといった情報がテクニカルメタデータの一例となります。テクニカルメタデータを参照することで、以下のような情報を取得できます。

- サーバーはどこにあるのか
- サーバーはどのように設定されているのか
- サーバーはいくつ利用できるのか
- データライブラリはどのように定義されているのか
- そのデータはどのように加工されているか

ビジネスメタデータは、データやサービスをより簡単に利用しやすくします。ビジネスメタデータは、ビジネスアナリストが根拠をもって正しい判断を行なうための情報を提供します。データ分類や表示形式の定義、ビジネス上の定義、実業務での使われ方といった情報がビジネスメタデータの一例となります。ビジネスメタデータを参照することで、以下のような情報を取得できます。

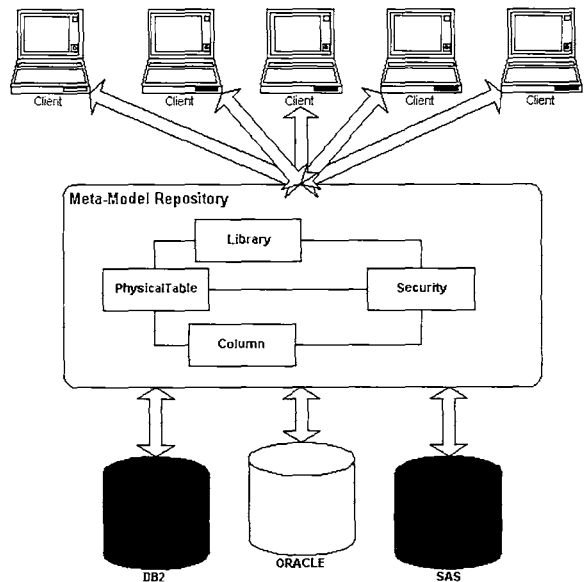
- 「売上」はどのように定義されているのか
- なぜこのデータが必要なのか
- 誰がこのデータを使用しているのか

なぜメタデータに注意を払うべきなのか

企業が情報を活用する目的は、現状維持でも衰退でもなく成長のためです。企業の成長と共にデータ量が増え、その環境は複雑になってきています。ユーザーが理解できないデータに、どのような意味があるのでしょうか？データの理解を助ける豊富な情報がなければ、たとえデータが大量にあろうと何の利益にもつながりません。

企業内のデータリソースは、最も価値のある資産の一つです。しかし、このリソースを管理するメタデータがなければ、データの保存場所も分からずアクセスできません。メタデータはこの問題を解決し、これらすべてのリソースを利用可能にします。

例えば、様々なデータが異なるアプリケーション上で管理されているとします。それらのデータを活用



図A

したいとき、通常は「どこに、どのようなエンジンで、どのような形でデータが格納されているのか」といった情報を知る必要があります。図Aのようなシステムでは、先に挙げた情報をメタデータとして管理し、クライアントアプリケーションはそのメタデータを利用して、データにアクセスします。これにより、ユーザーはそのようなことを意識せずにデータを適切な形で活用できます。

3 SAS によるメタデータ管理

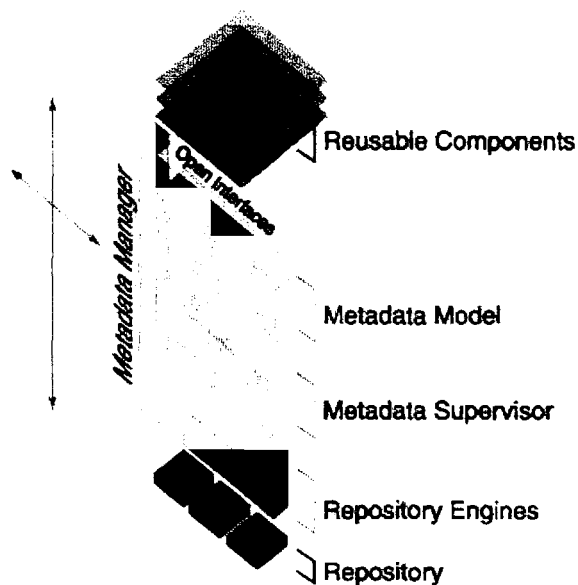
メタデータを利用した IT 環境を作り始めることは大変難しく、また非常に時間がかかるものだと言われています。企業はどのようにメタデータを組み込んだ環境を構築するのでしょうか。SAS は、その問題に対するソリューションとして、「SAS メタデータアーキテクチャ」を提供します。

SAS メタデータアーキテクチャ

SASメタデータアーキテクチャは、SASやその他のアプリケーションに共通のメタデータサービスを提供するための仕組みです。この仕組みによって、メタデータを統合します。統合されたメタデータを使用することによって、SASアプリケーションは、データに共通性、一貫性、信頼性を与えます。

このアーキテクチャは、メタデータインターフェース、メタデータモデル、メタデータスーパーバイザー、リポジトリエンジンの4層で構成されています。

- メタデータインターフェースは、メタデータを管理するための API(アプリケーション・プログラミング・インターフェース)です。クライアントとサーバー間での情報のやりとりは、XML を使用しています。業界標準のメタデータモデルと XML をサポートすることによって、SAS アプリケーションとその他のアプリケーション間の互換性を高めました。
- メタデータモデルは、メタデータのタイプやその属性、また個々のメタデータ間の関係などを定義したものです。SAS は、様々なメタデータを最適な形で管理するための土台として、このメタデータモデルを提供しています。メタデータモデルに基づいてメタデータを登録することにより、メタデータの関連性等を適切な形で保存することができます。
- メタデータスーパーバイザーは、ランタイムサービスを提供し、メタデータへのアクセスの承認を行います。メタデータスーパーバイザーは、マルチユーザーやマルチスレッド環境で、よりその効果を発揮します。
- リポジトリエンジンは、リポジトリへのインターフェースを提供します。メタデータはリポジトリエンジンを介してリポジトリ内に保存されます。リポジトリは、SAS、Oracle、DB2 といったデータ形式で保存が可能であり、それらの相違をリポジトリエンジンは吸収しているため、ユーザーはその相違を認識する必要はありません。



図B

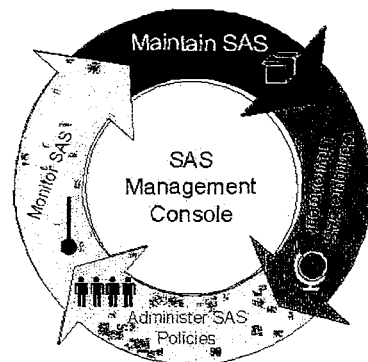
図 B は、SAS メタデータアーキテクチャを示したものです。この図におけるメタデータマネージャは、リソースやメタデータ、その他の重要な情報を管理しています。このメタデータマネージャは、「SAS Management Console(図D)」から利用できます。

SAS のメタデータ管理機能

SAS Management Console とは、様々なメタデータを一元で管理するための標準インターフェースです。図 C は SAS Management Console の作業サイクルを表したものです。

メタデータを一元管理することは、多くの利益につながります。例えば、以下のような利点が挙げられます。

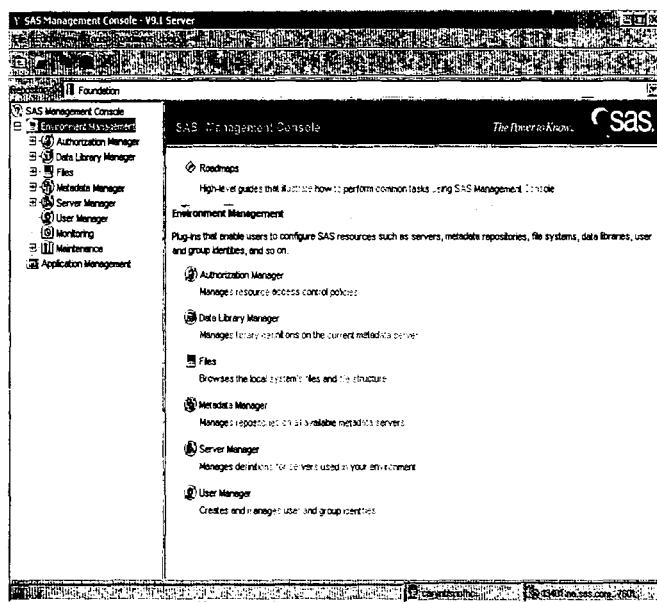
- プロセスの共通化・標準化が行える
- データを探すとき、探す場所は1箇所が良い
- 変更が発生したとき、修正する場所は1箇所が良い
- データの重複をなくせる



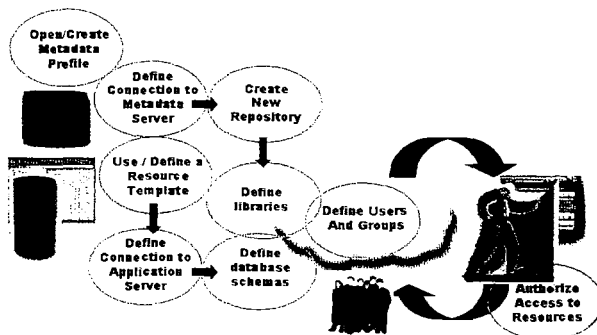
図C

図 D は、SAS Management Console の様々な機能を表したものです。これらの機能について詳細に説明します。

- **メタデータマネージャ**
メタデータマネージャでは、以下の作業を行うことができます。
 - メタデータサーバーの管理 (停止、一時停止など)
 - メタデータの変更履歴管理
 - メタデータのインポート・エクスポート
 - メタデータのリプリケーション・プロモーション
- **サーバーマネージャ**
サーバー設定情報を登録・管理します。複数のサーバーが存在する環境であれば、よりその機能を活かすことができます。
- **ユーザーマネージャ**
ユーザーとユーザーグループを作成します。ログイン情報等の定義も行います。



図D



図E

- 権限マネージャ
ユーザーやユーザーグループにアクセス権を設定します。アクセス権やアクセス・コントロール・テンプレートも作成します。
- ライブラリマネージャ
SAS ライブラリの定義を行います。データベーススキーマの管理も行います。
- ライセンスマネージャ
SAS プロダクト情報を登録できます。プロダクトのインストールやセットイニット更新時などに利用します。

なお、これ以外の管理機能が必要な際には、プラグインとして作成することが出来ます。図Eは、管理コンソールが企業のすべての管理タスクに役立つ様々な方法を表しています。

4 Case Study

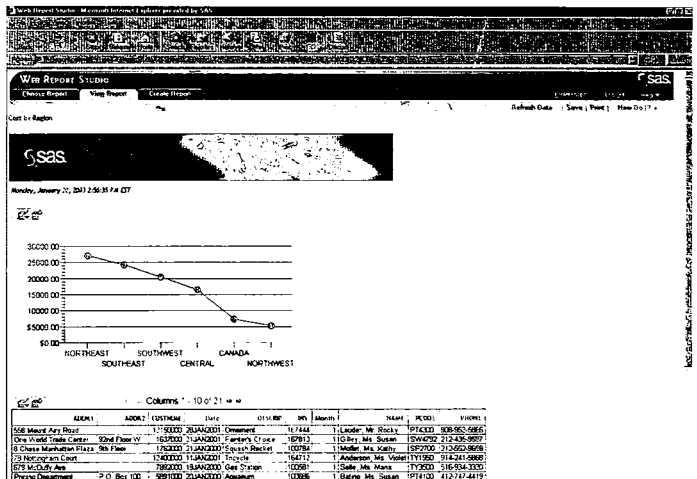
それではここまでに説明した内容を実際に試してみたいと思います。

メタデータは、単に情報技術の枠組みの中でのみ有用なわけではありません。メタデータは、ビジネス上の決定が必要な場面においても有益な情報を提供します。それでは始めましょう。

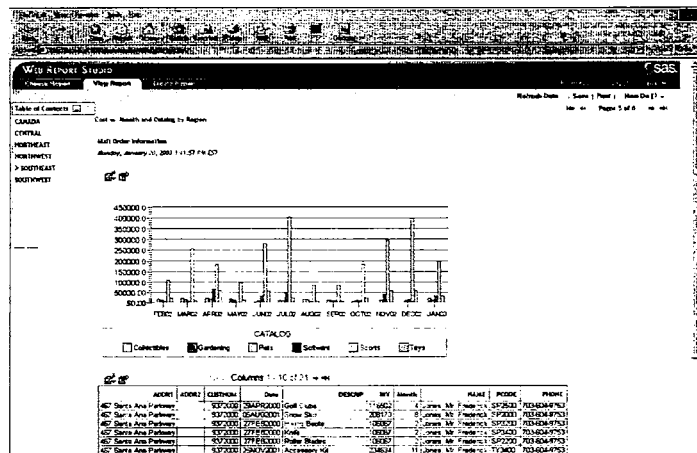
サンプルビジネスシナリオ

米国のとある会社を例に話を始めましょう。その会社は国内に支店をいくつか抱えています。経営幹部である Bob の1日は最新の販売報告書を確認することから始まります。ある日 Bob は、南東地区のオフィスの販売報告書に、売上の大きな落ち込みを発見します。あまりにも大きな落ち込みなので、なんらかの手違いがあるように Bob は感じました。

図Fと図Gは、SAS Web Report Studio の画面です。SAS Web Report Studio は、SAS メタデータアーキテクチャ(SAS アプリケーション間でメタデータを共有するための



図F



図G

仕組み)を利用した SAS 9.1 のプロダクトです。図Fは地区別の売上を示しています。図Gはその中でも南東地区の売上にフォーカスし、ここ1年間の売上推移を月別に表示しています。

図Gでは、売上額が2002年12月の40万ドルから2003年1月の20万ドルまで減少しています。この劇的な変化に Bob は「これは本当なのか?」という疑問を抱きました。その1ヶ月の間に、一体何が起こったのでしょうか。メタデータはこの質問に答えることができます。

メタデータの設定について

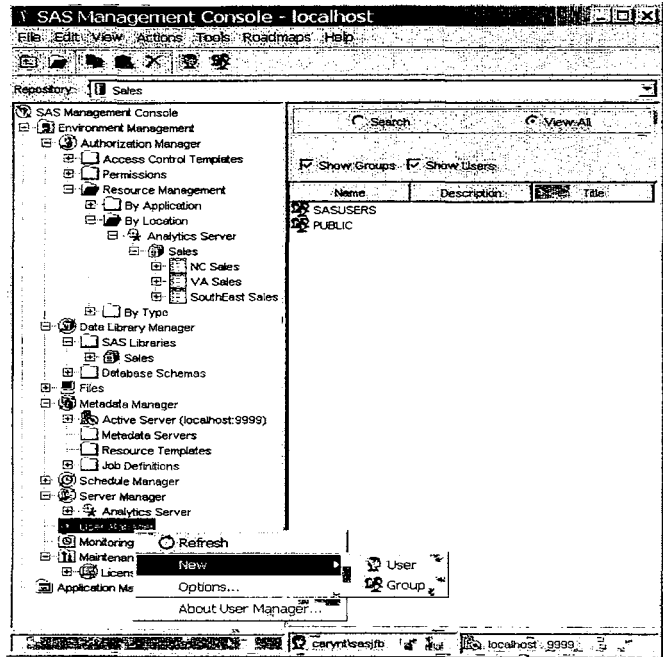
いくつかのメタデータオブジェクトの中に、この疑問に関連する答えが含まれています。図Hは SAS Management Console の画面です。この SAS Management Console には、「Sales」と呼ばれるライブラリが定義されています。その定義はあくまでもメタデータとしての定義です。Sales ライブラリ自体は、実際は分析処理を行う別のサーバーに割り当てられています。

まずは、管理者がサーバーとリポジトリを定義します。サーバーは複数のリポジトリを保持することができます。リポジトリは通常、明確な目的別に定義されます。例えば、「人材」と「在庫」というように異なる業務領域のものは、別のリポジトリとして定義することができます。この例では、「Sales」というリポジトリが用意されています。SAS Management Console においてメタデータの設定を行う際は、リポジトリの設置場所、名称、メタデータにアクセスする際に使用する SAS エンジンといった定義をウィザード形式で設定していきます(図I)。これによりリポジトリを簡単に作成することができます。

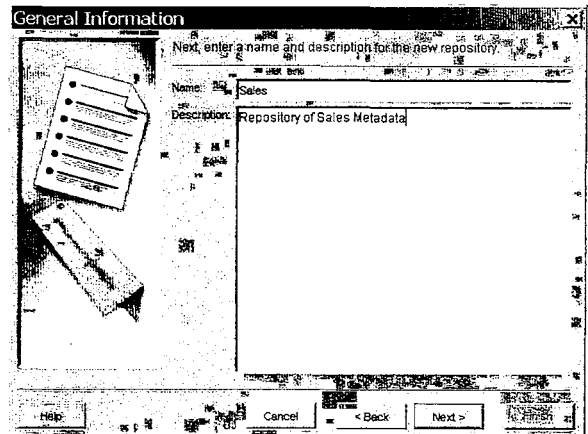
アクセス権の設定について

次に、メタデータへのアクセス制御について説明します。

メタデータへのアクセス制御は、ユーザーまたはユーザーグループごとに任意に設定することができます。図Hは、ユーザーまたはユーザーグループを設定するためのユーザーマネージャの画面です。



図H



図I

まずはユーザーとユーザーグループの作成・管理方法について説明します。

● ユーザーとグループ

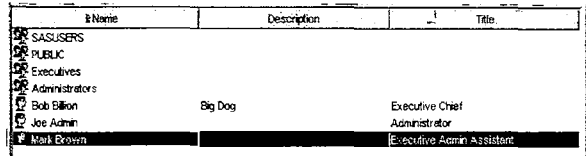
このサンプルシナリオの登場人物は、経営幹部、システム管理者、管理部門スタッフの3名です。アクセス権は、SAS Management Console において設定することができます。図Jは、このシナリオでのユーザー定義状況を示しています。

この例では、Bob Billion、Joe Admin、Mark Brown という3名のユーザーがいます。また SASUSER、PUBLIC、Executives、Administrators という4つのユーザーグループが登録されています。Executives グループは経営幹部が属するグループで、そこには Bob Billion と Administrators グループが登録されています。Administrators グループには Joe Admin が登録されています。Mark Brown はどのグループにも属していません。Mark には現在、読み取り権限のみが与えられています。

SAS Web Report Studio のようなクライアントアプリケーションは、ここで定義されたアクセス権によってレポートへのアクセスを制御することができます。例えば、図Gの売上レポートは、Executives グループのみ閲覧が可能です。

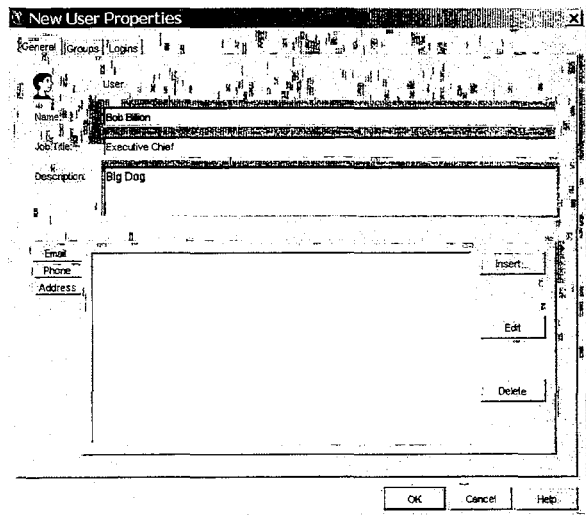
なおメタデータテーブル自体の使用・更新は、Administrators グループのみが行えるように設定されています。もしメタデータテーブルを更新したいのであれば、ユーザーやグループを作成する際に、その権限を与えなければなりません。

図Kは、実際にユーザー Bob Billion を作成する際の画面です。図Lでは Bob のログイン ID とパスワードの設定を行っています。管理者がユーザーやグループを作成する際、それぞれのユーザーごとにログイン情報を定義します。

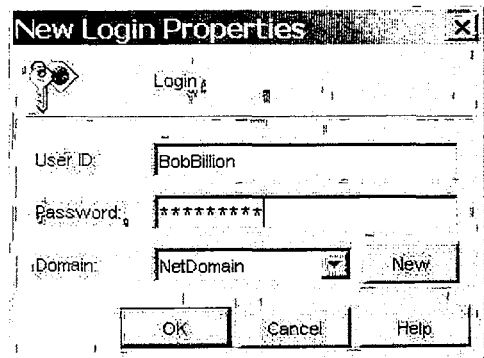


Name	Description	Title
SASUSERS		
PUBLIC		
Executives		
Administrators		
Bob Billion	Big Dog	Executive Chief
Joe Admin		Administrator
Mark Brown		Executive Admin Assistant

図J



図K



図L

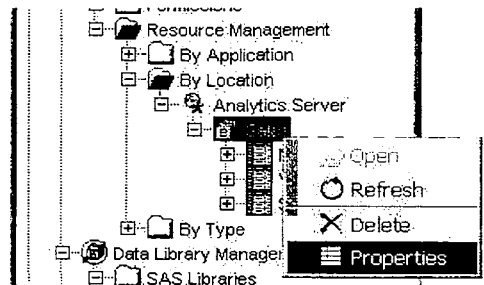
● ログインプロセス

ところで、そのログイン情報はどのように使用されるのでしょうか。SAS 9.1 では、メタデータは SAS Metadata Server によって管理されます。SAS Metadata Server 上のメタデータを使用するには、SAS Metadata Server にログインしなければなりません。

SAS Metadata Server におけるログインプロセスは、次のような流れになります。

- ① ホストOSによってユーザーの認証を行う。
- ② ホスト認証済みのドメイン名とユーザーIDを受け取る。
- ③ SAS Metadata Server 上の全てのログインオブジェクト(ユーザーに関するメタデータのうちのひとつ)の中から、そのユーザーIDと合致しているユーザーを探し出す(厳密に言うと、ログインオブジェクトが持つユーザーID属性の値とホスト認証済みのユーザーIDを比較することになります)。
- ④ 合致するログインオブジェクトが見つかったら、そのログインオブジェクトを保持したユーザー識別オブジェクトを取得します。
- ⑤ 以降、承認プロセスは、このユーザー識別オブジェクトによって行われます。

ログインオブジェクトに設定するユーザーIDは、OSのユーザーIDと同一のものを使用します。例えば、Windows 上でユーザーの認証が行われるのであれば、認証に使用されるユーザーIDをログインオブジェクトのユーザーID属性に、"domain¥userid" または"userid@domain"といった形式で入力します。同様に、ログインオブジェクトのドメイン属性に、認証で使用されるドメイン名を入力します。そのユーザーが複数のドメインで定義されているのであれば、ログインオブジェクトはドメインごとに作成します。

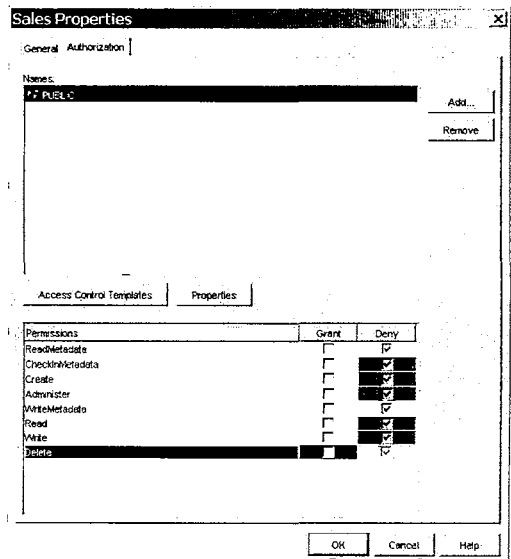


図M

• アクセス権

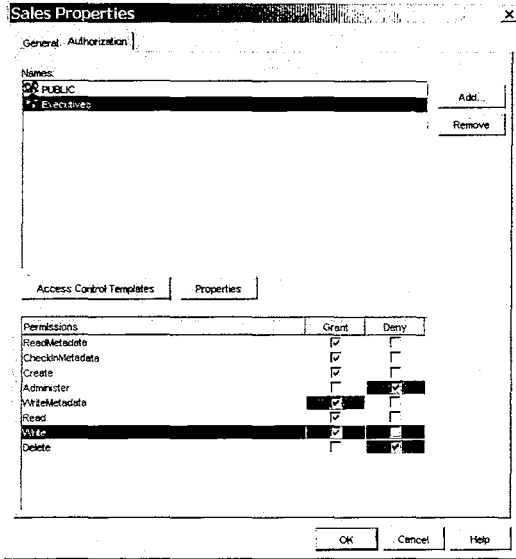
次に、「誰が何をすることが出来るのか」といったアクセス権の設定を行います。アクセス権も SAS Management Console 上で設定します。Sales ライブラリのプロパティから、それに設定されたアクセス権を確認できます(図M)。

SAS Management Console では、様々なレベルのアクセス権を設定することができます。デフォルトの設定では、PUBLIC グループにメタデータへの読み書き権限が与えられています。ユーザーやグループとして定義されていない全てのユーザーは PUBLIC グループのユーザーとしてみなされます。なおサンプルシナリオでは、PUBLIC グループに対して全てのアクセス権を拒否しています(図N)。

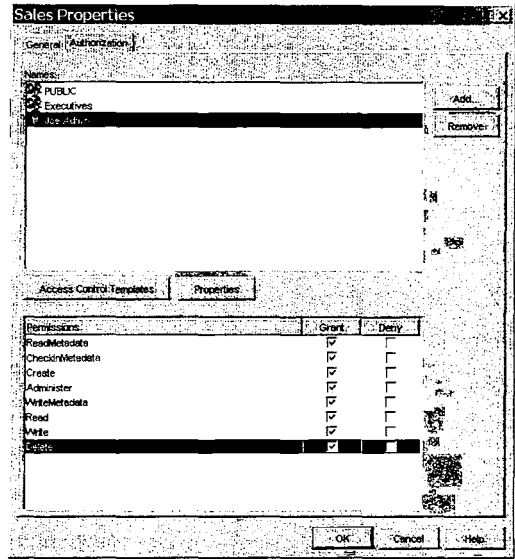


図N

一方、Executives グループのユーザーには様々なアクセス権が与えられています(図O)。しかし、彼らにメタデータを管理したり削除する権利はありません。経営幹部の人々が誤ってメタデータを消去しないようにするためです。図Pは、Joe Admin に設定されたアクセス権を示しています。Joe はこのシステムの統括管理者であるため、全ての情報に対して十分なアクセス権を持つ必要があります。



図O



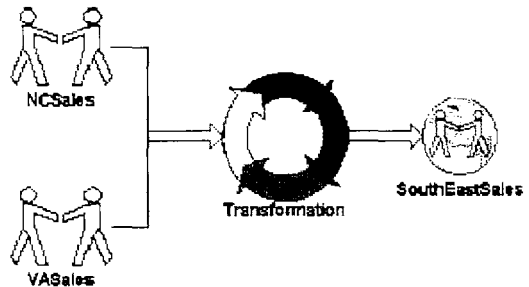
図P

意思決定の裏側

それでは、サンプルシナリオに戻ります。Bob は、問題のレポートの信頼性を判断するために2つの疑問を問いかけてました。

- 「このレポートはいつ作成されたのか？」
- 「このレポート作成の責任者は誰なのか？」

もしかしたら、このレポートは何週間も前に作成されたものかも知れません。誤ったデータテーブルを用いて作成されているかも知れません。またはレポートを生成する過程でエラーが生じた可能性もあります。



図Q

各々のメタデータには「作成日付」、「更新日付」という属性があり、いつレポートが定義され生成されたのかを知ることができます。データの出所もメタデータとして管理することができます。図 Q は、南東地区の売上データの出所を表したものです。図における「Transformation」は、南東地区の売上データを作成する際の加工ステップを表しています。

メタデータを詳細に調査した結果、南東地区の売上メタデータテーブルの「作成日付」と「更新日付」という属性から、問題のレポートは2、3日前に作成されたものであるということが判明しました。レポートの作成時期に関しては問題がなさそうなので、2つ目の疑問「このレポート作成の責任者は誰なのか？」に取りかかります。メタデータにはそれぞれ所有者・責任者が定義されていたため、Joe Admin が問い合わせるべき相手だということが明らかになりました。Bob はJoeに電話で問い合わせることにしました。

調査は続く

Bob から調査依頼を受けたJoeは、次のような疑問を投げかけます。

「このレポートは正しいのか？」

これまでの調査で、レポートの数字が最近のものであるということは判明していますが、その数字が正確かどうかまでは判明していません。

Joe は、その数字の正確さを判断するために2つの疑問を問いかけてみました。

「南東地区の売上レポートの『売上』項目はどのように計算されているのか」

「レポートに何かしらの変更が加えられていないか」

『売上』項目の算出ロジックを知ることは重要です。このレポートはどのデータを利用しているのでしょうか。そのテーブルのうちの1つに間違いがあるのでしょうか。レポートの内容が変更され、誤った計算結果を招いているのかもしれませんが。何らかの理由により単位が変更されたのかもしれませんが。

メタデータによって、Joe はこの数値がどのように計算されたのかという情報に素早くたどり着くことができました。Joe はデータ加工に関するメタデータを参照して、データの出所を突き止めました。このレポートは、ノースキャロライナ州とバージニア州の売上データをもとに作成されているようです。しかし、それらのデータと、『売上』項目の算出ロジックについて、何ら問題は見受けられませんでした。

次に、2つ目の問いかけ「レポートに何かしらの変更が加えられたのか」についても調査しました。しかし、変更が加えられている様子はありませんでした。

ということは、このレポートは「正しい」ということなのでしょう。レポートとしては正しい値を示しているようですが、この劇的な変化には何らかの理由があるはずです。

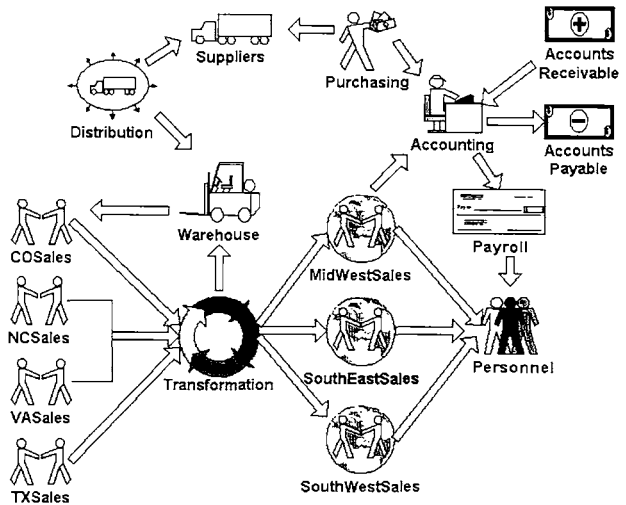
この問題を解決するために、もう少し広い範囲で調査を行って見ます。Joeは、このレポートを作成するためのデータフローだけでなく、ビジネス全体のデータフローを見てみることにしました(図R)。

この図を見ると、売上データは倉庫からの商品の仕入れデータもインプットしているようです。そこでJoeは仕入れに関する全てのデータの調査を、部下のMarkに依頼しました。Markはそれらのデータについて、「南東地区」というキーワードでメタデータ探索を行いました。

明らかになった原因

メタデータを探索することによって、Mark は南東地区の売上レポートにおける売上減少の原因を突き止めることができました。

どうやら、南東地区の店舗において、商品の在庫切れが多数発生していたようです。この会社は、最近コロラドに倉庫を購入したのですが、なぜかその新しい倉庫が、南東地区の店舗の仕入先としてシステムに登録されていました。これらの店舗は商品の仕入先として、本来、フロリダの倉庫を使用すべきなのですが、地理的に遠いコロラドから商品を仕入れていたことで商品の輸送に時間がかかり、商品の供給量が劇的に落ち、在庫切れを招いていたようです。



図R

Bob、Joe、Mark は、メタデータを利用することで物流の問題を発見することができました。南東地区

の店舗の仕入先をフロリダの倉庫に変更し、その結果、売上とコストを改善することができました。また Bob は、商品の搬出ポリシーに関して新たな意思決定を行い、再発の防止に努めました。

5 最後に

ここまで、メタデータについて論じるとともに、SAS によるメタデータ管理方法を簡単に説明しました。

現在、多くの企業が巨大な先行投資を抱えています。ひとたびシステムが構築されたとしても、今度はそれを維持するためのランニングコストと莫大な手間がかかります。

今日の情報システムには、便利であるとは言い切れない面が数多く存在することだと思います。メタデータには、「情報システムを便利なものにする」という可能性が秘められています。

Bob、Joe、Mark の 3 人の奮闘は、多少、現実離れしている感もありますが、メタデータを管理することで様々な利益を享受することができるという事実が変わりはありません。あなたにしか行うことのできない創造的な作業に、より多くの時間を割いていただくために、SAS はこれからもソリューションを提供し続けます。

日本SASユーザー会 (SUGI-J)

Enterprise Guide2.0 による add-in 機能について

木下 貴文

SAS Institute Japan 株式会社

カスタマーサービス本部 プロフェッショナルサービス第一部

Creating Custom Task for Enterprise Guide2.0

Takafumi Kinoshita

Professional Service Department 1/SAS Institute Japan Ltd.

要 旨

SAS Version8 よりリリースされた、エンドユーザ向け分析/レポートツール の Enterprise Guide の新バージョンである Enterprise Guide2.0 には様々な新機能が存在する。それらの新機能の一つとして、独自のタスクダイアログを作成し Enterprise Guide の機能としての利用、提供が可能になった。本稿では、利用方法と幾つかのサンプルを紹介しアドイン・カスタマイズ機能を説明する。

キーワード： Enterprise Guide2.0、COM アドイン、Visual Basic

はじめに

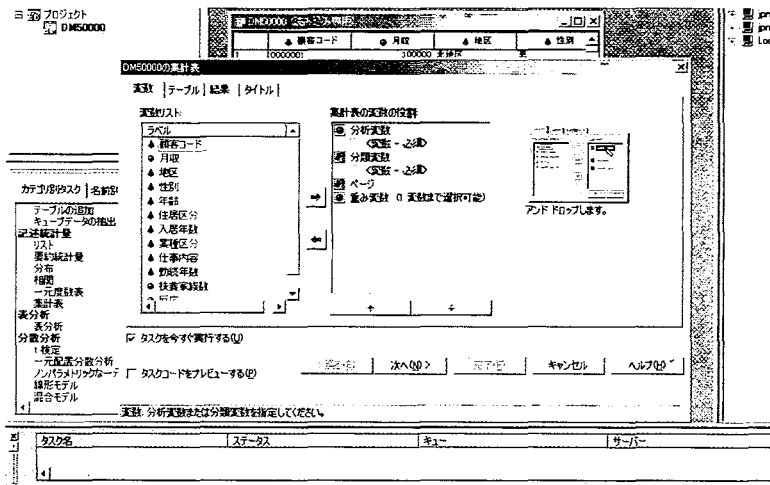
昨今、SAS システムがこれまでとは異なったユーザー層での導入事例が多くなってきている。これまでの SAS システムは、「Fat-Client」でのクライアント・サーバーでの利用がほとんどであったが、Enterprise Guide やその他のブラウザベース等でのクライアント・インターフェースでの SAS システムの利用が受け入れられてきた結果と考えられる。ブラウザ等の「Thin-Client」と呼ばれるクライアントでは、従来の「Fat-Client」の SAS システムの全ての機能を提供することは、コスト面やツールの習得面から考えても難しいが、逆に分析/レポートを行うユーザーの全てがそれらの機能を必要としているわけではない。この、ブラウザやシンプルなインターフェースでの SAS システムのエンジンの提供を行う事で、様々なニーズのユーザーに対しても SAS システムのメリットの提供が可能になった。Enterprise Guide では、グラフィカルでユーザーフレンドリーなインターフェースでの分析/レポート機能を提供しているが、新しいバージョンの Enterprise Guide2.0 のアドイン機能により、標準では搭載されていない機能や、業界や自社内のみで発生する処理などを SAS 外のアプリケーション開発言語(詳細については後に説明)で開発をして追加することが可能になる。この機能を利用することにより、また新たなユーザー層での SAS システムの利用が期待される。

第1章 Enterprise Guide とは

本章では簡単に、Enterprise Guide の機能と構成についての簡単な紹介をおこなう。

第1節 Enterprise Guide の基本的な機能紹介

まず Enterprise Guide はこれまでの SAS によるクライアント・サーバーシステム構築手段である、SAS/Connect による接続や、ブラウザインターフェースから CGI 経由でサーバー接続をおこなうものではなく、COM/DCOM 又は、IOM Bridge といった技術によりサーバーシステムへの通信を実現している。この COM/DCOM 又は、IOM Bridge による接続では、クライアントモジュールがそれらの技術に対応できているものであれば、サーバーモジュールである SAS/Integration Technologies によりサーバーの接続が可能になる。そのため、Enterprise Guide も SAS の SCL や HTML ではなく、Visual C++ で開発されているためより Windows ライクな操作性で親しみがあり、使いやすいインターフェースとなっている。



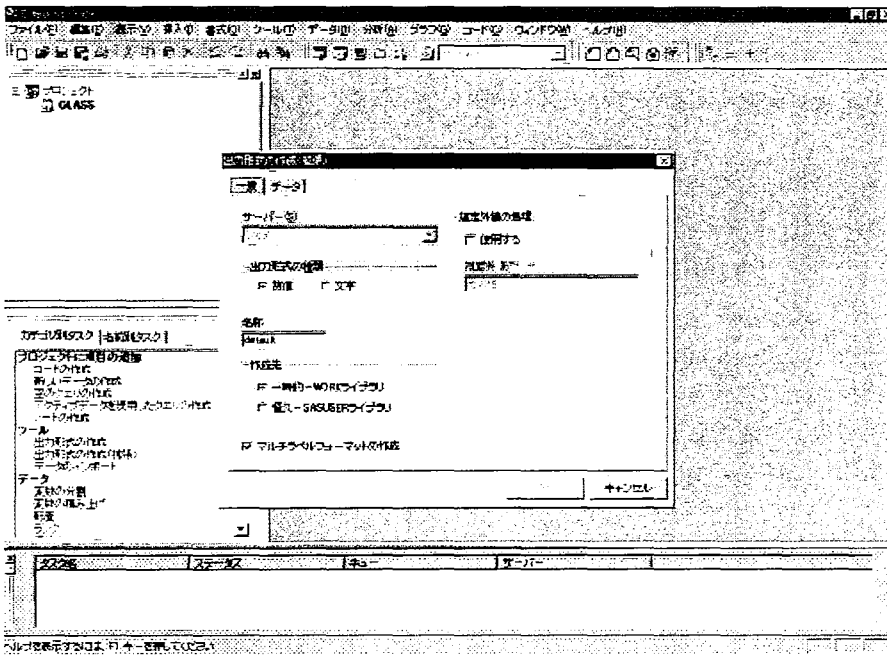
【ユーザーフレンドリーな操作画面】

操作方法としては、行いたい処理を選択して各種設定をおこなうためのウィザードを起動させる。ウィザード中で処理をおこなうために必要な変数、及びアウトプットのためオプション等選択して【完了】ボタンを選択する。以上の操作により、ノンプログラミングで対象データに対してアウトプットを得ることができる。アウトプットについては、帳票的なものは HTML 形式、PDF、リッチテキストフォーマットに出力が可能であり、グラフは G I F、JAVA、ActiveX に対応が可能である。また簡単なデータ加工は、クエリーウィンドウの機能を利用することによりノンプログラミングで実行が可能である。クエリーウィンドウでは、特定の条件によるデータの抽出や複数テーブルのマージ処理をサポートしている。複数のマージ処理も、設定することによりレフトジョイン、ライトジョイン等にも対応が可能になる。これらのデータ加工については、処理後のデータを SAS データセットのみではなく SAS データビュー形式で保存も可能であるため、全てデータセットにしなくても使用する状況にあわせてそれらを使い分け、リソースの管理をおこなうことも可能である。またこれまで紹介

した機能は全て GUI 操作で実現可能だが、GUI 操作により生成された SAS プログラムを保存しバッチジョブとしスケジューリング登録をすることも可能であり、編集をすることによって再利用することも可能である。

第2節 Enterprise Guide の応用的な機能

Enterprise Guide ではグラフィカルな GUI を利用し様々な処理をおこなうことができるが、標準でサポートされていないか、標準の機能でも実現は可能だが様々な処理を組み合わせる必要がある処理など、業務や業界に特化した機能を全ておこなうのは難しい。このような場合に、SAS の SCL 言語などではない一般的な開発言語で開発したアプリケーションを Enterprise Guide にアドオンすることにより、Enterprise Guide の一機能として実現することが、Enterprise Guide2.0 から可能になった。



【サンプルアドインモジュール起動時のイメージ】

上記のサンプルイメージのように、追加されたモジュールはあたかも Enterprise Guide2.0 上の一機能のように操作が可能で、データの加工や分析、レポート処理をおこなうことができる。この機能を利用することにより、ただ単に Enterprise Guide をツールとしてユーザーに提供するのではない、エンドユーザーが本当に求めるような形でソリューションパッケージとして Enterprise Guide を提供することが可能になった。

アドイン機能を利用するには、まず Enterprise Guide2.0 でアドインモジュールの登録作業をおこなう必要がある。アドインモジュールを登録して利用可能アドインリストに追加をおこなう。作業の手順としては、

1. [ツール] メニューの中の [ユーザー設定] を選択し、[アドイン] タブをクリック。
2. [追加] ボタンを選択すると表示される追加タスクの登録画面で、[ProgID] のコマンドラインに追加タスクのプログラム識別子を入力する。
3. 利用可能アドインリストで追加したアドインを選択して、Enterprise Guide2.0の再起動をおこなう。以上の操作で、Enterprise Guide2.0上へのアドインコンポーネントの追加登録がおこなうことができる。これらの詳細な手順については、後程説明をおこなう。

第2章 アドインコンポーネントとは

次に前章で紹介したアドインコンポーネントの具体的な内容と開発方法、及び実行例を紹介する。また紹介の際に提示するサンプルアプリケーションは、Visual Basicで開発されたものを利用するが、本稿では多言語の詳細なロジックの説明等は、ページの都合上割愛する。

第1節 アドインコンポーネントとは

Enterprise Guide2.0 で追加登録の可能なモジュールは様々な言語での開発を可能にしている。これらのアドインモジュール(COM アドイン)は最終的には、DLL(ダイナミック・リンク・ライブラリ)として用意をする必要がある。これらが可能な開発言語としては、Microsoft Visual Basic、Microsoft Visual C++、.Net、C#等が上げられる。これらの開発のために、開発言語で利用可能な Enterprise Guide との接続インターフェイス(API)と利用可能メソッドを用意している。これらの詳細な情報については、SAS 社サイトの

http://www.sas.com/technologies/bi/query_reporting/guide/segcustomize.chm に詳細な利用方法があるので参照されたい。上記にある詳細情報は Visual Basic をメインに書かれているが、それ以外の言語でも、アドインの作成は可能である。以下に簡単なプログラムのサンプルイメージを提示するが、このサンプルコンポーネントの全て記述すると、膨大な量になるためその中の一部のプログラムを記述する。

- サンプルプログラムイメージ

BEGIN

```
MultiUse = -1 'True
Persistable = 0 'NotPersistable
DataBindingBehavior = 0 'vbNone
DataSourceBehavior = 0 'vbNone
MTSTransactionMode = 0 'NotAnMTSObject
```

END

```
Attribute VB_Name = "SortOptions"
Attribute VB_GlobalNameSpace = False
Attribute VB_Creatable = True
Attribute VB_PredeclaredId = False
Attribute VB_Exposed = True
Attribute VB_Ext_KEY = "SavedWithClassBuilder6" ,"Yes"
```

```
Attribute VB_Ext_KEY = "Top_Level" ,"Yes"
```

```
Option Explicit
```

```
'local variable(s) to hold property value(s)
```

```
Private mvarSortOrder As String 'local copy
```

```
Public Property Let SortOrder(ByVal vData As String)
```

```
'used when assigning a value to the property, on the left side of an assignment.
```

```
'Syntax: X.SortOrder = 5
```

```
    mvarSortOrder = vData
```

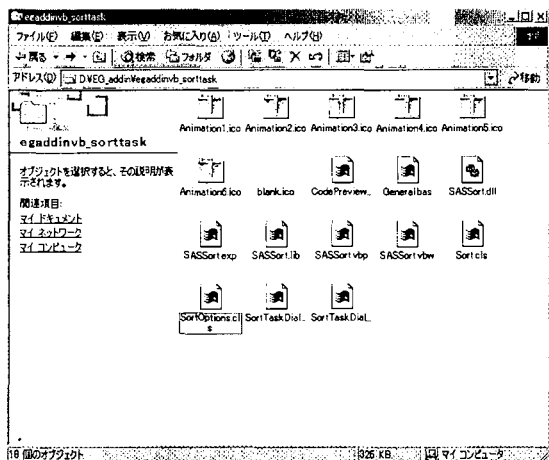
```
End Property
```

第2節 アドインコンポーネントの登録及び実行

アドインコンポーネントを作成してソースファイル、イメージファイル、DLL ファイルを作成した後に、Enterprise Guide2.0 で実際に登録をおこなう際には以下のような処理をクライアントサイドでおこなう必要がある。今回は登録のプロセスを行う際に、Sort 機能を追加したサンプルモジュールを利用しておこなってみたい。標準の Enterprise Guide 内で実際に Sort 処理のみをおこなう場合は、クエリウィンドウを利用して、SQL ベースで実行するしかなく(グラフ等の作成時に自動的におこなわれる、sort 処理は除く)データ加工時に従来の SAS 言語での Sort プロシジャの重複行削除オプションを利用することはできなくなっている。そのような際に、このような Sort プロシジャをおこなうモジュールをアドオンで提供することによって、ユーザーの更なるニーズを満たした使い方が可能になる。アドインコンポーネントの詳細な登録については、以下のようになる。

1.

まず、Sort 処理をおこなうモジュールを開発してそれらをローカルドライブに準備する。(今回は D:\¥EG_addin¥egaddinvb_sorttask 内に準備をおこなった。)



【モジュールフォルダの一覧イメージ】(但し、内容は用意するファンクションによって異なる。)

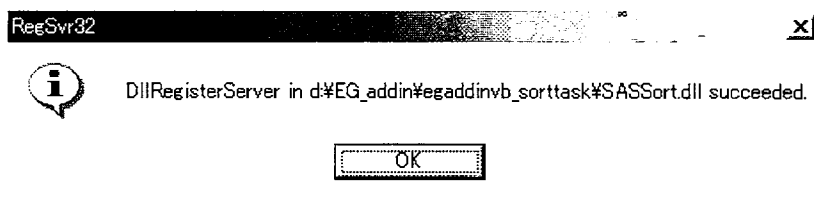
2.

次にサンプルモジュールのスクリプトの登録をおこなう。今回のモジュールのDLLファイルは”D:\¥EG_addin¥egaddinvb_sorttask”配下にSasSort.dllという名前であるために、次の様なコマンドにより登録をおこなう。登録はWindowsの[スタート]メニューから、[ファイル名を指定して実行]でコマンドの入力をおこなう。これはWindowsに搭載されているregsvr32.exeを利用し、<registration>の要素の情報を読み込み、クライアントマシン内にあるWindowsレジストリに登録をおこなうという作業にあたる。この作業により、クライアントマシン内にサンプルモジュールをの登録が完了される。

regsvr32 D:\¥EG_addin¥egaddinvb_sorttask¥ SasSort.dll

[ファイル名を指定して実行]でコマンドを入力し、登録が完了すると以下のようなメッセージが表示されるので[OK]を選択する。

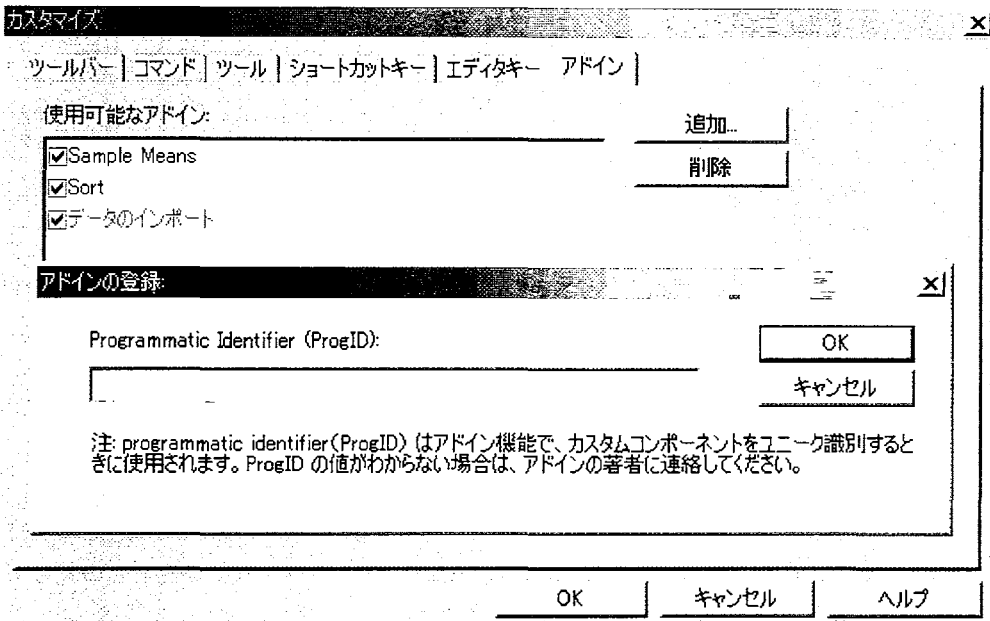
また多くのクライアントに対して、上記の作業をおこなうと煩雑さが増すために上記の作業をバッチファイルとして、ダブルクリックのみで実行が可能な状態にしてユーザーに配布することで、デリバリの作業の効率化も考えられる。



登録が完了すると上記のようなポップアップメニューが出現し、登録完了が確認される。

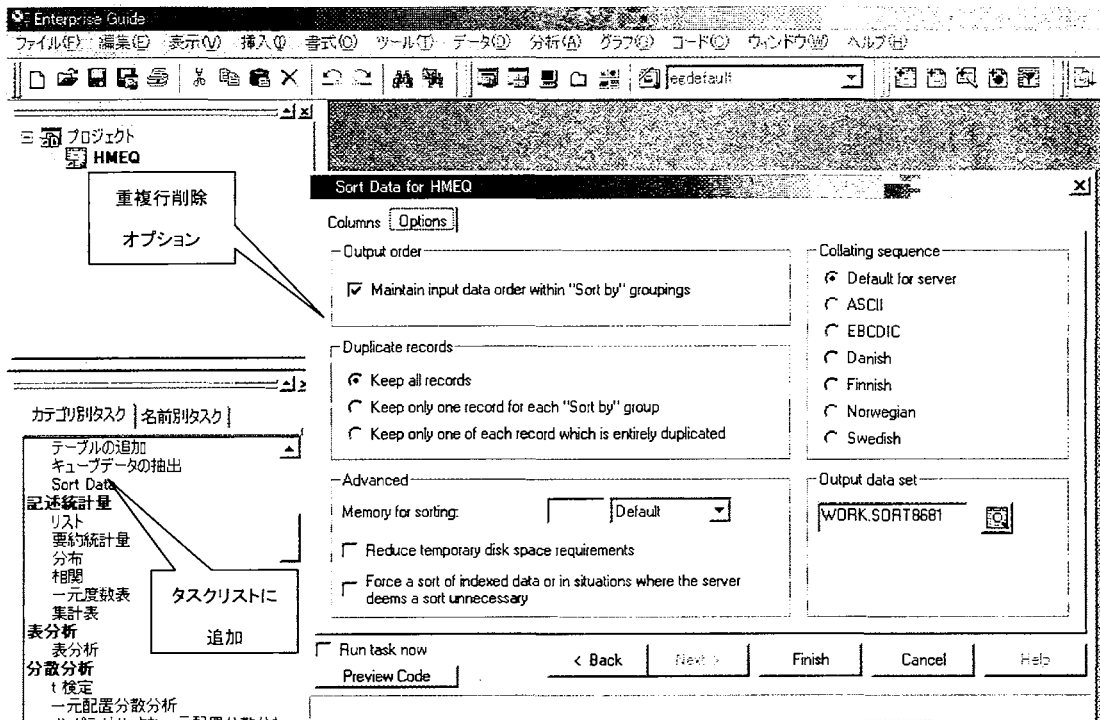
3.

サンプルモジュールを登録後に、Enterprise Guide内で登録をおこなったサンプルモジュールのProgidの指定をおこないます。このProgidは開発者がスクリプト コンポーネントを参照するために使用するテキスト名のため、各クライアントに登録をモジュール開発者以外がおこなう際は、事前に開発者からProgidの情報を得る必要がある。Progidが判明の後、[ツール]メニューの中の[ユーザー設定]を選択し、[アドイン]タブ中の[追加]を選択すると、Progid登録画面が表示される。このProgid登録画面にProgidを登録し、Enterprise Guideの再起動をおこなうとメニュー中にアドインのモジュールのメニューが登録されたのが確認できる。



4.

再起動後のEnterprise Guide2.0上には新しいタスクを確認することができ、新しい機能を使うことで標準のEnterprise Guide2.0では実現ができなかった機能を再現することが可能になる。



【Sortモジュールの操作画面】

まとめ

本論文では、サンプルアプリケーションを利用しユーザーニーズに合わせたアドインのEnterprise Guide2.0での実装方法について説明をおこなった。今回は、単純にSortプロシジャを実装させたサンプルアプリケーションになったが、開発次第では特定の項目の値を設定することにより元データより抽出をおこない、その後に定型のレポート処理までを行うモジュールを一つの機能として開発することも可能になる。この機能を利用すると、これまでその作業をおこなうためだけに開発を行い、社内業務に特化したアプリケーションなどをEnterprise Guide2.0内に吸収をして、同一SASシステムとしてメンテナンス、運用をおこなうことが可能である。それに伴いもう一つのメリットとして、SASの柔軟なデータハンドリング能力やデータ処理/分析能力を、これまでの対象ユーザーではなかったユーザーにも活用することができるようになってきている。

これらの機能はエンドユーザーにも利用可能にはなっているが、実際に活用が想定されるのはSASを利用したサービスの提供を検討している、System Integrator、コンサルティング会社などのソリューションプロバイダである。各会社が持っている、業務ノウハウ・コンサルティング力をベースとした業務に特化したアプリケーションを有効にアドイン機能として提供することによりツールにプラスαされた、サービス提供が考えられる。また開発の際も、Visual BasicやVisual C++などの一般的な開発言語を利用できるため、アプリケーション開発者にもハードルが低く開発をおこなっていただくことができる。

また今回、本論文中で利用したサンプルアプリケーションはSAS社サイトで自由にダウンロードが可能のため、一度開発をおこなう前にトライアルとしてダウンロードをして試してみることもできる。サイトのアドレスは以下のとおり、

http://www.sas.com/technologies/bi/query_reporting/guide/customtasks.html

日本SASユーザー会 (SUGI-J)

SAS/SHAREサーバーアクセスログの分析

中村 崇文

SAS Institute Japan 株式会社
カスタマーサービス本部プロフェッショナルサービス第1部

Analyzing the Access Log

Takatomo Nakamura

Professional Service Department 1 Customer Services Division
SAS Institute Japan Ltd.

要 旨

クライアント/サーバー型の接続におけるサーバー上の SAS システムにおいて、どのユーザーがどのデータに、何回アクセスしたのかといった情報を得る一例として、SAS/SHARE ソフトウェアによるログを利用する方法がある。本稿では、サンプルプログラム等をまじえながらその実装方法を紹介する。

キーワード： SAS/SHARE ソフトウェア SAS/CONNECT ソフトウェア アクセスログ

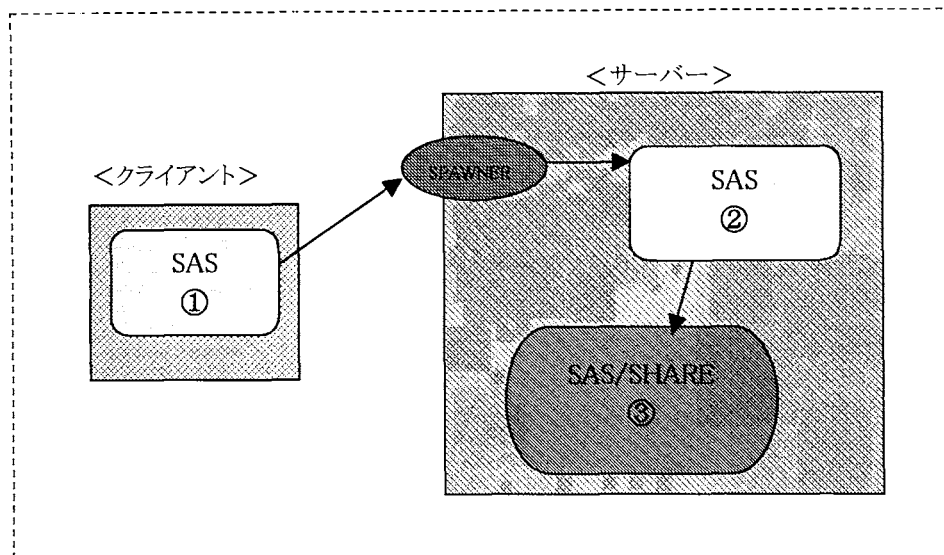
1. はじめに

SAS/SHARE ソフトウェアは排他制御の機能を提供するもので、サーバー上のデータに対して一貫した同時更新処理の実現を可能とするプロダクトである。サーバー上にこの SAS/SHARE ソフトウェアが導入されていれば、そのログを参照することにより、SAS/SHARE の配下にあるデータに対しての複数ユーザーからのアクセス状況を確認することが可能である。本稿では次ページに示すようなシステム構成を想定し、SAS/SHARE ソフトウェアによるログを外部ファイルに出力させ、さらに SAS が提供するサンプルプログラムを利用してそれを SAS データセット化し、分析可能な環境を作成する方法を紹介する。

2. システム構成例と操作手順の概要

本稿では次の【システム構成図】に示すようなシステム構成を想定する。

【システム構成図】



- ・ クライアントの SAS①からサーバー上の SAS②へ SAS/CONNECT により接続を行なう。
- ・ サーバー上の SAS の起動には SPAWNER を使用する (SAS/SHARE のログ取得とは関係ない)。
- ・ サーバー上で SAS/SHARE③を起動する。

なお、今回検証のために使用した環境は以下の通りである。

	OS	SAS のバージョン
サーバー	Solaris 8	Release 8.2 (TS2M0)
クライアント	Windows2000 Professional (SP2)	Release 8.2 (TS2M0)

※ サーバーにおいては !sasroot/sas_ja (!sasroot は SAS のルートディレクトリを指す) を用いて SAS を実行する。なお、本サーバーの !sasroot は /SAS/SAS_8.2 である。

※ サーバーのホスト名は「jpn1」である。

※ 接続時のユーザー名は「jpntcn」、事前にサーバーのホームディレクトリ/users/jpntcn に、分析用データの格納場所として「slogdata」を作成済みである。

また、SAS/SHARE によるログを取得してデータセット化を行なうまでの手順は以下の通りである。

- (1) SAS/SHARE の環境設定 (サーバー)
- (2) SAS/SHARE の起動 (③)
- (3) SAS/CONNECT を用いてサーバー上の SAS を起動 (①→②)
- (4) SAS/SHARE への接続 (①→②→③)
- (5) SAS/SHARE の停止、セッションの終了 (①→②→③)
- (6) SAS/SHARE のログファイルのデータセット化 (①→②)

3. SAS/SHARE の環境設定と起動

SAS/SHARE 起動のための環境設定として、サーバーの OS の Services ファイルへの登録が必要である。サーバーの OS の Services ファイルに以下を追記して、SAS/SHARE サーバー名とポート番号、プロトコルを定義する。

shr	5060/tcp	#SAS/SHARE サーバー
-----	----------	-----------------

※SAS/SHARE サーバー名「shr」を Services ファイルに登録

SAS/SHARE サーバーは1つの SAS セッション上で起動を行なう形となる。以下のコマンドにより、サーバー上の SAS を起動する。この SAS セッション起動時に、-log オプションを指定することにより、SAS/SHARE のログを外部ファイルに出力させる。

```
/SAS/SAS_8.2/sas_ja -log /users/jpntcn/share.log
```

※-log オプションを指定し、ログを外部ファイル/users/jpntcn/share.log(ファイル名は任意)に出力する。

SAS/SHARE サーバーを起動するために、上記で起動した SAS セッション上で次の SAS プログラムをサブミットする。

```
options comamid=tcp ;
%let tcpsec=_secure_ ;
proc server msgnumber server=shr ;
run ;

endsas ;
```

本稿で紹介するような、SAS が提供するサンプルプログラムを利用してログファイルをデータセット化する際には、ここで PROC ステートメントに必ず msgnumber オプションを指定する必要がある。また、上記のプログラム2行目のように「%let tcpsec=_secure_ ;」を記述することにより、SAS/SHARE サーバーへの接続時にユーザー認証を行なうことが可能である。認証を取らない場合は、「%let tcpsec=_secure_ ;」を記述せず、PROC ステートメントにおいて authenticate=optional を指定する。なお、デフォルト(記述無し)では authenticate=required となる。

4. SAS/SHARE サーバーへの接続

本稿では、2.の【システム構成図】で想定しているように、まずクライアントの SAS①から SAS/CONNECT によりサーバー上の SAS②を起動し、さらにこのサーバー上の SAS セッションから、同じサーバー内で起動している SAS/SHARE③へと接続を行なう。

SAS/CONNECT を使用してサーバー上の SAS を起動するには、クライアントの SAS で次のようなプログラムをサブミットする。

```
options comamid=tcp remote=jpn1 ;
filename rlink "d:%test%CONNECT_scr%tcpunix_sun450.scr" ;
signon ;
```

SAS/SHARE サーバーへの接続は、LIBNAME ステートメントを使用する。サーバー上の SAS に接続完了後、クライアントの SAS で以下のプログラムをサブミットすることで SAS/SHARE サーバーへの接続が開始される。

```
rsubmit ;
libname lib1_shr '/users/jpntcn/shrtest1' server=shr
                user=jpntcn password=XXXXXX ;
endrsubmit ;

libname lib1_shr slibref= lib1_shr server=jpn1 ;
```

なお、SAS/SHARE サーバー起動プログラムの PROC ステートメントにおいて、authenticate=optional を指定した場合には、LIBNAME ステートメントにて user=、password=オプションを指定する必要はない。

5. SAS/SHARE の終了とログファイルのデータセット化

SAS/SHARE の終了には OPERATE プロシジャを使用する。次のプログラムをクライアントの SAS でサブミットする。

```
rsubmit ;
proc operate server=shr user=jpntcn password=XXXXX ;
  stop server ;
run ;
endrsubmit ;
```

SAS/SHARE を起動した SAS セッションの起動時に-log オプションで出力先を指定した外部ファイルは、この SAS セッションの終了時に作成される。上記のプログラムをサブミットすることでサーバー上の SAS/SHARE が終了し、SAS/SHARE の起動プログラムの後に記述していた ENDSAS ステートメントが実行され、SAS セッションが終了する。このタイミングで-log オプションで指定したディレクトリに share.log というファイルが完成する。これは SAS/SHARE が生成するログが外部テキストファイルとして出力されたものである。

このログファイルを、SAS が提供する分析データ作成用サンプルプログラムを使用することによって、分析可能な形のデータセットを作成する。クライアントの SAS で次のプログラムをサブミットする。

```

rsubmit ;
filename INLOG '/users/jpntcn/share.log' ;
libname SLOGDATA '/users/jpntcn/slogdata' ;
filename pgm '/SAS/SAS_8.2/samples/share' ;

%include pgm(sltoolm.sas) ;
%include pgm(sltool1.sas) ;
%include pgm(sltool2.sas) ;
/*%include pgm(sltool3.sas) ;*/
/*%include pgm(sltool4.sas) ;*/
endrsubmit ;

libname SLOGDATA slibref=SLOGDATA server=jpn1 ;
libname S_WORK slibref=WORK server=jpn1 ;

```

上記プログラムでは sltoolm.sas、sltool1.sas、sltool2.sas という3つのサンプルプログラムを使用しているが、以下にそのサンプルプログラムの内容を記す。

分析データ作成用サンプルプログラム(格納場所/SAS/SAS_8.2/sample/share)

プログラム名	説明
sltool0.sas	以下の全てのプログラムを実行する。
sltoolm.sas	マクロ変数定義を行なう。
sltool1.sas	ログの読み込みを行なう。データはWORKに作成される。ログデータの2倍程度の空き容量が必要である。
sltool2.sas	sltool1.sasにより作られたデータセットを目的別に加工し、SLOGDATAライブラリ内に保存する。
sltool3.sas	サンプルプログラム
sltool4.sas	サンプルプログラム

上記プログラムの実行により、指定した SLOGDATA ライブラリ内に複数のデータセットが作成される。これらのデータセットについて内容を簡略に説明したものが次ページの表である。

SLOGDATA ライブラリに作成されるデータセット

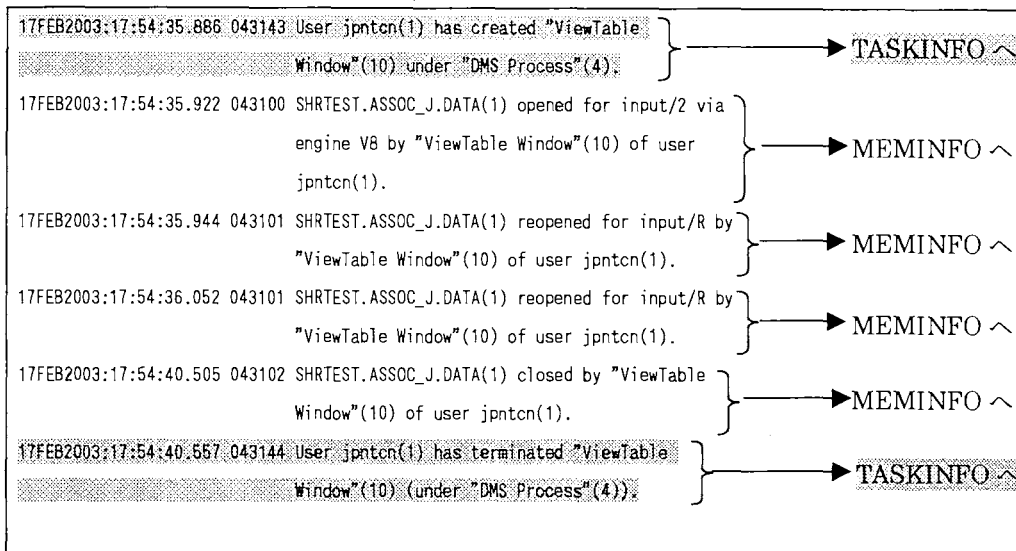
データセット名	説明
CONNINFO	ユーザーによる SHARE サーバーへの接続、切断を示す 「接続」とは、LIBNAME ステートメントによるライブラリ割り当て、 OPERATE プロシジャの実行を含む
CONNSUM	接続の総数を示す
TASKINFO	各処理(プロシジャ、DATASTEP、ViewTable Window など)の開始と 終了を示す 変数 USERID-TASKNMBR でユニークになる
TASKSUM	ユーザー別のタスク(上記 TASKINFO に出力されたもの)の数を示す
LIBINFO	ライブラリの割り当て、解除の時間を示す ユーザーごとライブラリごとに、割り当て時2オブザベーション、解除時 2オブザベーションの計 4 オブザベーション
PHYSINFO	ライブラリが割り当てられた実際のパスを示す
ENGSUM	ライブラリエンジンのリストを示す
MEMINFO	データセットへのアクセスを示す データセットへのアクセスについては、input、output モードでは OPEN と CLOSE の 2 オブザベーション、update モードでは OPEN、 REOPEN、CLOSE の 3 オブザベーション 1つのデータセットに対し ViewTable Window を開いて閉じる処理につ いては OPEN、REOPEN、REOPEN、CLOSE の4オブザベーション 変数 USERID-TASKNMBR-SLIBREF-MEMNAME でユニークになる ため、ここからデータセットへのアクセスの回数、アクセスしていた時間 等を求めることができる
OCMDINFO	OPERATE プロシジャの実行を示す
SERVINFO	SHARE サーバーの起動と停止の時間を示す

また、SAS/SHARE が生成するログの内容と、データセットへの出力は以下のようなイメージとなる。

例 1 TRANSPOSE プロシジャの実行

17FEB2003:17:54:18.505 043143 User jpnctn(1) has created "TRANSPOSE"(8) under "Program"(1).	} → TASKINFO へ
17FEB2003:17:54:18.521 043100 SHRTEST.ASSOC_J.DATA(1) opened for input/S via engine V8 by "TRANSPOSE"(8) of user jpnctn(1).	} → MEMINFO へ
17FEB2003:17:54:18.945 043102 SHRTEST.ASSOC_J.DATA(1) closed by "TRANSPOSE"(8) of user jpnctn(1).	} → MEMINFO へ
17FEB2003:17:54:19.023 043144 User jpnctn(1) has terminated "TRANSPOSE"(8) (under "Program"(1)).	} → TASKINFO へ

例2 ViewTable Windowのオープン、クローズ



6. データ加エプログラム例

ここでは、SLOGDATA ライブラリ内に作成されたデータセット TASKINFO、MEMINFO から、それぞれ処理に要した時間、データセットへのアクセス時間を算出するプログラムの例を紹介する。クライアントの SAS で以下のプログラムをサブミットする。

```

rsubmit ;

/*****タスクの処理時間*****/
/****タスクごとにユニークになるように並べ替え****/
proc sort data=SLOGDATA.TASKINFO out=WORK.S_TASKINFO ;
  by USERID TASKNMBR ;
run ;
/****タスク処理時間を計算****/
data WORK.T_TASKINFO ;
  set WORK.S_TASKINFO ;
  by USERID TASKNMBR ;
  format TASKTIME time13.2 ;
  retain START END ;
  drop START END ;
  if first.TASKNMBR then
    do ;
      START=DTSTAMP ;
      END =DTSTAMP ;
    end ;
  else
    do ;
      if START>DTSTAMP then START=DTSTAMP ;
      if END <DTSTAMP then END =DTSTAMP ;
    end ;
  if last.TASKNMBR then

```

```

do ;
    TASKTIME=END-START ; /*←タスク処理にかかった時間を計算*/
output ;
end ;
run ;

/*****データセットへのアクセス時間*****/
/****データセットのopen~closeでユニークになるように並べ替え****/
proc sort data=SLOGDATA.MEMINFO out=WORK.S_MEMINFO ;
    by USERID TASKNMBR SLIBREF MEMNAME ;
run ;
/****データセットへのアクセス時間(ACCESSTIME)を計算****/
data WORK.T_MEMINFO ;
    set WORK.S_MEMINFO ;
    by USERID TASKNMBR SLIBREF MEMNAME ;
    format ACCESSTIME time13.2 ;
    retain START END ;
    drop START END ;
    if first.MEMNAME then
        do ;
            START=DTSTAMP ;
            END =DTSTAMP ;
        end ;
    else
        do ;
            if START>DTSTAMP then START=DTSTAMP ;
            if END <DTSTAMP then END =DTSTAMP ;
        end ;
    if last.MEMNAME then
        do ;
            ACCESSTIME=END-START ; /*←データセットへのアクセス時間を計算*/
            output ;
        end ;
run ;

/****変数TASKTIME、ACCESSTIMEを含むデータセット作成*****/
data WORK.TIME ;
    keep USERID TASKNAME SLIBREF MEMNAME DTSTAMP ACCESSTIME TASKTIME ;
    merge WORK.T_TASKINFO WORK.T_MEMINFO (IN=d rename=(DTSTAMP=DTSTAMP1)) ;
    by USERID TASKNMBR ;
    if d ;
run ;

/**時間順に並べ替え**/
proc sort data=WORK.TIME out=WORK.TIME1 ;
    by DTSTAMP ;
run ;
endrsubmit ;

```


7. 参考

(1) ログ上の User ID とユーザーのアクセス権限

サーバー上の SAS セッション②を起動したユーザーと、その SAS セッションから SAS/SHARE サーバーに対しライブラリを割り当てたユーザーが異なる場合のログ上の User ID は、前者となる。ただし書き込み、読み込み権限は後者のものとなる。

user1・・・読み込み、書き込み権限あり

user2・・・読み込み、書き込み権限なし

		SAS セッション起動者		
		user1	user2	
libname ステートメント内のユーザー名	user1	ライブラリ割り当て	○	○
		データ参照	○	○
		データ作成	○	○
		ログ内 User ID	user1	user2
	user2	ライブラリ割り当て	○	○
		データ参照	×	×
		データ作成	×	×
		ログ内 User ID	user1	user2

(2) SAS/SHARE を Windows サーバー上で起動する際の権限設定

サーバーのOSがWindows2000(もしくはWindowsNT)の場合には、サーバー側で以下の方法により権限設定を行なう必要がある。

「管理ツール」→「ローカルセキュリティポリシー」→「セキュリティの設定」→「ローカルポリシー」→「ユーザー権利の割り当て」を開き、

- ・ 「オペレーティングシステムの一部として機能」にて、SHAREサーバーを起動するユーザーに対して権限を付与する。
- ・ 「バッチジョブとしてログオン」にて、「Authenticated Users」、ログオンするユーザーに権限を付与する。

8. おわりに

以上、本稿ではSAS/SHAREのログをもとに分析用のデータセットを作成するプロセスをサンプルのプログラムと共に紹介した。既にSAS/SHAREソフトウェアを導入済みで、かつサーバー上のデータに対するアクセス状況を把握したい場合には、是非とも参考にさせていただきたい。

参考文献

“SAS/SHARE User’s Guide, Version 8” P87～ Analyzing the Server Log

簡易 運用入門

弘田 貴

カスタマーサービス本部

SAS Institute Japan 株式会社

Simple guide to employment

Takashi Hirota

Customer Service Division,

SAS Institute Japan, Ltd.

要 旨

SAS Enterprise Guide や、SAS Enterprise Miner 等使用する際、元データの整備は不可欠である。非定型な利用で都度作成するデータであれば、特に考慮する必要はないものの、日単位、月単位で、定期的にレポート等作成する処理を考慮した上で、データ整備の為の運用を実施する際の注意点をまとめたものである。

キーワード： 運用、環境構築

はじめに

SAS Enterprise Guide や、SAS Enterprise Miner 等を使用する際、使用者は、使用する環境や、データの規模(大きいのか、小さいのか)を意識する事無く使用しがちで、結果、使用環境の許容範囲を越えるデータを処理しようとして環境を停止させたり、ネットワークに負荷をかけて、処理時間が予想に反したりすることがある。このような場合、使用者及び利用促進者には、予め使用するデータの規模や構造を意識したデータ加工プロセスが必要になることを、再認識していただくことをお勧する。ここでは、バッチモードによるデータ加工プロセスに着目して、最低限、運用する際の注意点、整備すべき資料を紹介する。

処理の流れを決める

データ加工プロセスを考える際に、最初に考えることは、大まかな処理の流れとなる。大まかな流れとして、元となるデータから、データ加工プロセスを経て、最終的な目的のデータを生成することを想定する。

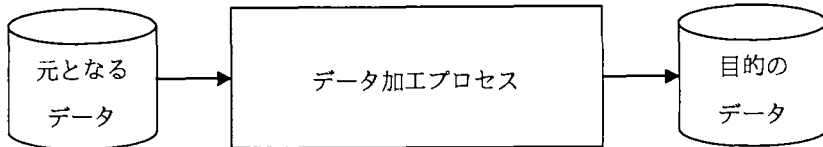


図1.

次に、元となるデータの種類、データ加工プロセス内の処理、目的のデータの種類へ落とし込んでいく。

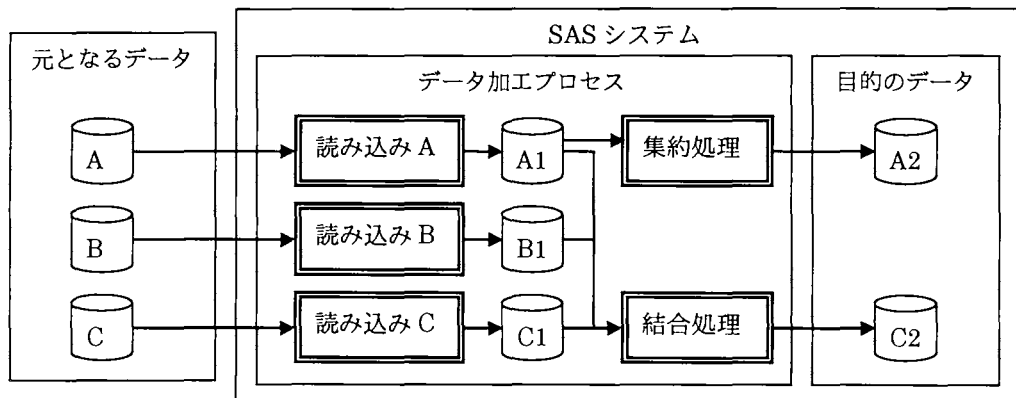


図2.

元となるデータの A,B,C はそれぞれ、異なる形式で任意の場所に保管されている。データ加工プロセスでは、それぞれのデータを SAS システム内に取り込む為に、読み込み処理 A,B,C を実施する。データ A1 を更に、任意のキーで集約した結果 A2 を作成する。また、データ A1,B1,C1 を、任意のキーで結合した結果 C2 を作成する。

次に、データ加工プロセス内の処理における確認項目に着目して考える。読み込み A,B,C は、元となるデータ A,B,C 無くしては、正常に SAS システム上のデータ A1,B1,C1 を作成できない。また、処理終了時に、データ A1,B1,C1 が生成できていない場合には、同様に正常処理が実施できていないことになる。したがって、処理:読み込み A,B,C では、起動時には元となるデータ A,B,C、処理終了時にはデータ A1,B1,C1 の存在確認が必要となる。集計処理、結合処理においても同様に、確認項目を予め想定しておくことと表1となる。

処理名	入力確認対象	出力確認対象	先行の処理	処理順	備考
読み込み A	元データ A	データ A1	—	①	①,②,③は、 並列実行でも可
読み込み B	元データ B	データ B1	—	②	
読み込み C	元データ C	データ C1	—	③	
集約処理	データ A1	データ A2	読み込み A	④	
結合処理	データ A1,B1,C1	データ C2	読み込み A,B,C	⑤	

表 1.

処理が正常に終了したのか、異常終了なのかの判定確認項目を予めまとめておくことをお勧めする。これは、次に説明する障害発生時の対応に大きく影響する。

障害対応

障害対応とは、処理が異常終了した際における、処理を正常に戻す為の対応処理を意図する。異常終了の定義は、様々な事象を想定して考慮する必要があるが、先の表1での入力確認対象及び出力確認対象となっているデータが存在しない場合に異常終了とまず定義する。

次に、出力確認対象のデータは確認でき、正常に終了したかに見えるが、プログラムの途中でエラーとなり、新しいデータが更新されていない。但し、過去に正常に作成されたデータが残った状態で、出力確認時には、正常と判断されてしまうような場合は、第1にプログラムが正常に実施されているかを確認する必要がある。第2に表1のデータ有無確認を有効にする為、過去に作成したデータを予め削除しておく、または、確認対象のデータ名に予め“年月日”という情報を付加して、確認時に年月日を意識する等の配慮が必要となる。また、作成するデータが累積型となる場合、予め対象データのバックアップ、障害回復時にリストア、再実行という流れも予め想定しておく事をお勧めする。

環境構築の為の準備

次に、環境構築の前段階として、以下の考慮点をあげる。

- ・ ハードウェア環境
- ・ ソフトウェア環境

ハードウェア環境

ハードウェア環境を検討する際、以下の資料を予めまとめておくと、後々の管理がし易くなる。

- ・ リソース一覧
- ・ メモリ容量算出一覧

リソース一覧

ハードウェア環境を検討する場合、ディスク容量の算出は欠かせないものである。最低限、以下の項目に関しては、予め想定し一覧表を作成しておくことをお勧めする。その際、容量に関しては、今後の拡張性を含めて任意係数を掛けて算出しておく事が望ましい。それは、システム稼動して、数ヶ月経過後に、ディスク容量が足りなくなった為、処理が停止するようなことは絶対に避けなければならぬ為である。その時点で、プログラム改修や、ディスク増設する場合と、予め任意係数を掛けた値で、ある程度余裕を持ってディスクを確保している場合とでは、投資費用面でかなりの差が生まれることは言うまでもない。

フォルダ/ ディレクトリ名	リソース名	ファイル 名	レコード 長 (BYTE)	件数/日 (最大値)	容量 (MB)	保存 期間 (月)	想定容量 (MB) x1.5
C:\¥LOW	A(元)	A.txt					
C:\¥LOW	B(元)	B.csv					
C:\¥LOW	C(元)	C.DAT					
C:\¥SAS_TMP	A1(中間)	A1 .sas7bdat					
C:\¥SAS_TMP	B1(中間)	B1 .sas7bdat					
C:\¥SAS_TMP	C1(中間)	C1 .sas7bdat					
C:\¥SAS_TEST1	A2	A2 .sas7bdat					
C:\¥SAS_TEST2	C2	C2 .sas7bdat					
C:\¥SASWORK	—	—	—	—	—	—	

表2.

注意すべきは、データの洗い替え(上書き)とする場合、対象データと新規データ分で、想定容量は、2倍以上を確保すること。また、SAS WORK ライブラリは、処理用途に合わせて3倍から5倍以上確保すること。尚、SAS データセットは、compress オプションを使用してデータセットを圧縮して使用することが可能なことは予め、押さえておく必要がある。

メモリ容量算出一覧

メモリ容量算出では、使用するアプリケーション(サーバ含む)の必須メモリを予め、一覧化し管理しておくこと、新規でハードウェアを購入する際および、現行ハードウェアでシステムが正常に稼動できるか否かを計る資料となる。以下の資料は、SAS システムにのみ着目した必須メモリ容量となるが、他社 DBMS や WEB サーバ、運用管理ツール、バックアップツール等、常時稼動するアプリケーションに関しての必須メモリは、同様に一覧に記載しておくこと良い。

また、WEBによる照会システムや、クライアント・サーバ環境において多数のユーザが、同時に1つのサーバ機に処理を集中させるようなシステム構成の場合には、1ユーザにおけるサーバ機の必須メモリ容量を予め想定し、最大同時アクセス数と掛け合わせた値分の容量を確保することが重要である。

WEB システム (AP サーバ用)	32MB×5 (最大起動数)	160MB
SAS/SHARE ソフトウェア	サーバ稼動	32MB
SAS システム稼動用		832MB
小計		1024MB

表3.

ソフトウェア環境

想定しうるソフトウェア全てのバージョン及び対応するオペレーションシステム(以降 OS と略す)を管理する。SAS システムにいたっては、システムのバージョン、TS レベル、HOTFIX 等のパッチ情報、プロダクト構成等を、予め控えておくと、後々 OS や SAS システムを含む他ソフトウェアのバージョンアップ時に効果を発揮する。

実装編

SAS プログラムの作成

ここでは、具体的な SAS プログラムを記さず、バッチモードにおける注意点を記す。バッチモードでの SAS システムの利用は、以下のように、作成した SAS プログラムをコマンドとして実行する。これにより、対話型ラインモードとは異なり、夜間等のバッチ実行が可能になる。

※ Windows 環境上で実施した場合のコマンド例

```
c:\sas\nls\ja\sas.exe -sysin "読み込み A.sas" -log "c:\log\読み込み A.log"
```

上記のように、実施する処理のプログラムを登録し、SAS システムの実行 LOG を別途出力指示しておくと、処理中にエラーが発生した際、“C:\log\読み込み A.log”という名称で、実行 LOG が出力される為、原因究明を実施する際、便利である。

バッチモードで SAS システムを使用する場合の注意点は、実行する SAS プログラムが終了した時点で、SAS WORK ライブラリ上で作成したデータが削除されるという、特性をもつことである。

障害発生の確認対象となっているデータが、WORK ライブラリで処理するようなプログラムは、固定のユーザライブラリで一時作成するように変更することが必要となる。

処理の実行

図2の処理を、実際に実行してみよう。表1を参考にして、各処理を順番に実行する。以下のように、順番に処理が実行されるように、新たにバッチファイル“test.bat”を作成する。

```
C:%test.bat -----  
c:%sas%nls%ja%sas.exe -sysin “読み込み A.sas” -log “c:%log%読み込み A.log”  
c:%sas%nls%ja%sas.exe -sysin “読み込み B.sas” -log “c:%log%読み込み B.log”  
c:%sas%nls%ja%sas.exe -sysin “読み込み C.sas” -log “c:%log%読み込み C.log”  
c:%sas%nls%ja%sas.exe -sysin “集計.sas” -log “c:%log%集計.log”  
c:%sas%nls%ja%sas.exe -sysin “結合.sas” -log “c:%log%結合.log”
```

Windows のコマンドにて以下を実行する。

```
c:%test.bat
```

また、以下のように Windows のショートカット機能を使って、処理毎に実行していく方法、及び、上記のバッチ処理をショートカットとしてリンク先に登録することで、手動での実行が可能となる。

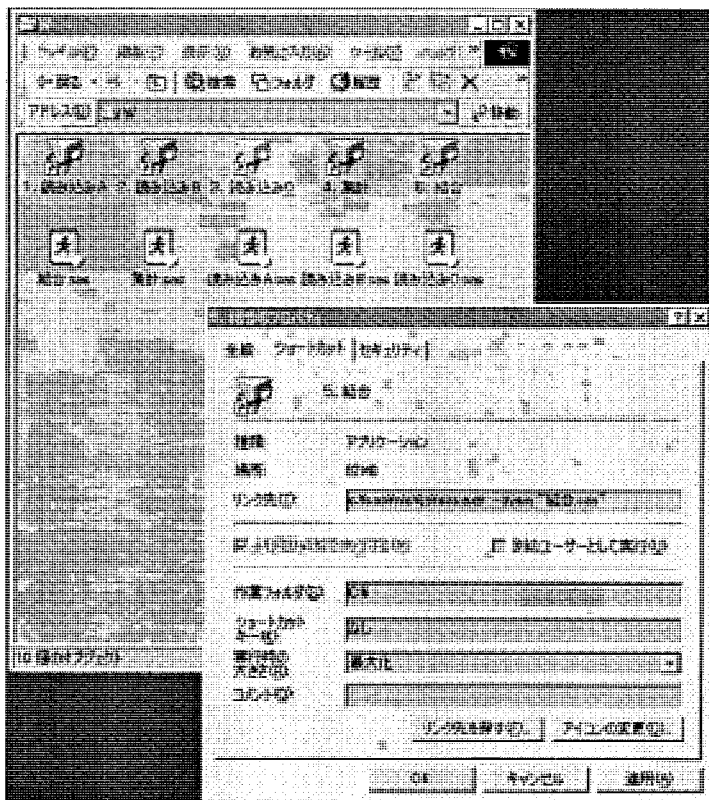


図3.

スケジューラソフトを使う

日次処理や、月次処理といったバッチ処理を自動運用したい場合、先に紹介したバッチモードによる実行コマンドをスケジューラソフト等に登録する場合と、障害発生時のエラーハンドリングを考慮した、バッチプログラムをスケジューラソフト等に登録する場合とがある。障害対応等を考慮すると後者をお勧めする。スケジューラソフトは、Windows の AT コマンドの使用でもスケジュール登録は可能だが、万人が使用し分かりやすいものを選択したほうが管理しやすいと言える。

サンプルとして、エラーハンドリングを考慮した図2の処理:読み込み A を対象に、WindowsNT 上での Microsoft-DOS で作成したサンプルを記す。尚、UNIX 環境では、シェル等で同様に作成し、スケジューラへ登録する。

サンプルでの SAS プログラムのエラーハンドリングは、SAS 実行 LOG より、“ERROR”という文字列が先頭に存在するか否か、“_ERROR_”という文字列が存在するか否かの 2 点で確認を実施している。共通定義や、固有定義は、別途バッチファイルを分けて行っても良い。また、処理の実施時間が把握できるように、時間という概念を加えても良い。このサンプルでは、メッセージを表示する形式を採用しているが、バッチ用の LOG ファイルに書き出して管理する形式に変更したり、実施年月日時分の情報をファイル名に付加し、任意期間を持って削除していくように変更すると尚良い。

WindowsNT:Microsoft-DOS プログラムサンプル

```
TEST_A.bat -----
@echo off

Rem -----
Rem 共通定義
Rem -----

set SASEXE=c:\Program Files\sas\nls\ja\sas.exe
set SASWORK=c:\saswork
set SASLOG=c:\test\log\test_a.log

Rem -----
Rem 定義
Rem -----

set SASPGM=C:\TEST\READ_A.sas
set INFILE=C:\LOW\A.txt
set OUTFILE=C:\SAS_TMP\A1.sas7bdat
```



```

Rem -----
Rem 入力ファイルチェック
Rem -----
if not exist %INFILE% goto INFILE_ERROR

Rem -----
Rem SASプログラム実行チェック
Rem -----
start /w %SASEXE% -sysin "%SASPGM%" -work "%SASWORK%" -log "%SASLOG%" -icon -nosplash
findstr /B "ERROR" %SASLOG% > nul
if not errorlevel 1 goto SASERR
find "_ERROR_=1" %SASLOG% > nul
if not errorlevel 1 goto SASERR

Rem -----
Rem 出力ファイルチェック
Rem -----
if not exist %OUTFILE% goto OUTFILE_ERROR
goto NORMAL

Rem -----
Rem エラー処理
Rem -----
:INFILE_ERROR
echo INPUT FILE NOT FOUND: %INFILE%
goto ERROR

:SASERR
echo SAS BATCH NG: %SASPGM%
goto ERROR

:OUTFILE_ERROR
echo OUTPUT FILE NOT FOUND: %OUTFILE%
goto ERROR

```

:ERROR

echo ERROR END

goto END

:NORMAL

echo NORMAL END

:END

echo ** End of TEST_A.bat **

日本SASユーザー会 (SUGI-J)

MEANS,TABULATE,DATASETS プロシジャの機能紹介

渋谷佳枝・〇檜皮孝史・迫田奈緒子

SAS Institute Japan 株式会社

カスタマーサービス本部プロフェッショナルサービス第1部

Useful Functions of MEANS,TABULATE,DATASETS Procedure

Yoshie Shibuya /Takafumi Hiwada /Noako Sakota

Professional Service No1 Department/SAS Institute Japan Ltd.

要 旨

本稿では Version6 から Version8 へのバージョンアップに伴い MEANS,TABULATE,DATASETS のそれぞれのプロシジャに追加されたオプションの中から便利だと思われるもの、上記のプロシジャの便利な使い方などを紹介する。これらの機能紹介がエンドユーザの利便性の向上に繋がれば幸いである。

キーワード：

MEANS,TABULATE,DATASETS

オプション、欠損値、一貫性制約

はじめに

SAS システムは、Version6 から Version8 へのバージョンアップにより様々な機能拡張が行われた。ODS に代表されるビジュアル面での向上や配信機能、DCOM や CORBA といった分散オブジェクトの準拠をはじめとしたよりオープンな環境での利用など、非常に多岐に渡り利便性の向上が図られている。また、新たなプロシジャの追加やオプションの追加など SAS 言語に対しても多くの拡張が行われている。そこで本稿ではエンドユーザの利用頻度が高い MEANS、TABULATE、DATASETS のそれぞれのプロシジャにおける拡張点と便利な使い方を紹介する。

1. MEANS プロシジャ

最初に MEANS プロシジャの拡張点と便利な使い方について説明する。MEANS プロシジャは要約統計量を算出する際に非常に便利なプロシジャであり、数多くある SAS のプロシジャの中でも比較的使用頻度が高いプロシジャであると思われる。

なお、ここで説明しているオプション等については SUMMARY プロシジャでも同様に使用することが可能となっている。

1.1 新たに出力可能な統計量

以下の統計量が新たに MEANS プロシジャで出力することが可能になった。これによりこれまで Univariate プロシジャで出力していた Median などの統計量も Means で出力することが可能となる。

MEDIAN(中央値)	P1(1 パーセント点)	P5(5 パーセント点)
P10(10 パーセント点)	P90(90 パーセント点)	P95(95 パーセント点)
Q1(25 パーセント点)	Q3(75 パーセント点)	QRANGE(Q1 と Q3 との差異)

1.2 Autaname, Autolabel オプション

次に、便利なオプションとして Autaname、Autolabel オプションを紹介する。これらは統計量をデータセットに出力する際に非常に便利なオプションとなっている。

Autaname オプションは、Output ステートメントで出力する統計量の変数名を明示的に指定していない時、自動的に変数名を作成し出力する。自動的に作成される変数名は、「分析変数名+統計量」で作成される。例えば、変数 Uriage に対して平均(mean)、中央値(median)を出力する場合、Autaname オプションを使用すると、出力される変数は「Uriage_Mean」、 「Uriage_Median」という変数名で作成される。

Autolabel オプションは、Autaname オプションと同様に Output ステートメントで出力する変数のラベルを自動的に作成する。ラベル名は、「分析変数のラベル+統計量」となり、また、分析変数にラベルがない場合は「分析変数名+統計量」となる。例えば、変数 Uriage(ラベル「売上」)に対して平均(mean)、中央値(median)を出力する場合、Autolabel オプションを使用すると、出力される変数のラベルは「売上_Mean」、「売上_Median」というラベルで作成される。

これまで 2 つ以上の分析変数を対象に複数の統計量を出力する場合、出力する変数名を指定しないと、全ての統計量を出力することができなかった。そのため、分析変数と統計量が数多くなればなるほど、記述する変数名の数も増加してしまう。しかしながら、Autaname オプションを使用すると、変数名を自動的に作成してくれるため、そういった

変数名の記述という煩雑な作業を軽減することが出来る。以下は変数名を指定した場合と Autaname オプションを使用した場合のプログラムの比較である(アウトプットの結果は同じ)。出力する変数が増えるにしたいが、Autaname オプションを使用する事でよりプログラムの効率化が図れると思われる。

```

PROC MEANS DATA = demo.bento NWAY NOPRINT;
  VAR point nedan total ;
  CLASS shop ;
  /*変数名を指定したOutputステートメント*/
  OUTPUT OUT = summary1 MEAN = heikin1 heikin2 heikin3 MEDIAN = chuo1 chuo2 chuo3 ;
  /*Autanameオプションを使用したOutputステートメント*/
  OUTPUT OUT = summary2 MEAN = MEDIAN = / AUTONAME ;
RUN ;

```

しかしながら、分析変数にラベルが貼付されている場合、上記のプログラムにより作成されたデータセットに出力された統計量には、分析変数のラベルがそのまま適用されるため、同じラベルの変数が複数作成されてしまう。これではどれがどの変数のどの統計量を出力したものか一見しただけでは理解しづらい。また、Means プロシジャ中で新たに作成される変数に対してラベルをつけようとしても Warning が表示されてしまう。

VIEWTABLE: Work Summary 2						
	満足度	値段	合計	満足度	値段	合計
1	73.528822055	433	866.3659179	72	430	860
2	73.999	446	890	70	450	900
3	73.491					860
4	73.559					860
5	73.818181818	409	817.9020979	73	445	890

ラベルが同じためにどの変数のどの統計量が分かりづらい

このような場合に、Autolabel オプションは非常に便利なオプションとなってくる。Output ステートメントの最後に「Autolabel」と付け加えるだけで、自動的に重複する事のないラベル名を作成し、以下のような理解しやすいデータを作成することが可能となる。

VIEWTABLE: Work Summary 2						
	満足度 Mean	値段 Mean	合計 Mean	満足度 Median	値段 Median	合計 Median
1	73.528822055	433	866.3659179	72	430	860
2	73.999	446	890	70	450	900
3	73.491484				430	860
4	73.559808				430	860
5	73.818181818	409	817.9020979	73	445	890

Autolabel によって自動的に作成されたラベル


1.3 CLASSDATA オプション

次に Version8 から新たに追加された CLASSDATA オプションについて説明する。CLASSDATA オプションを使用すると、元データの分類変数に存在しない値でも CLASSDATA で指定したデータに存在していれば、MEANS プロシジャの出力結果には CLASSDATA での分類変数での処理結果が出力されるようになる。以下の例を参考にしていきたい。

データ「gender」に対して変数「性別」の件数の集計を行いたい。しかしながら、結果には、「女性」「不明」以外に0件の「男性」という値も出力したい。通常、このデータを Means プロシジャで処理した場合、男性という値は性別に入っていないため出力されない。

	id	性別
1	1	女性
2	2	女性
3	3	女性
4	4	不明
5	5	女性

「gender」



	性別	kensu
1	女性	4
2	不明	1

```


PROC MEANS DATA = gender NWAY NOPRINT ;
  CLASS sex ;
  OUTPUT OUT = result2(DROP=_type_ _freq_) N = kensu ;
RUN ;

```

しかしながら、出力したい分類変数の値を格納したデータ(下図「sex」)を作成し、そのデータを CLASSDATA オプションで指定することにより、任意の値を出力することが可能となる。

	sex
1	男性
2	女性
3	不明

「sex」



CLASSDATA に sex を指定

	性別	kensu
1	女性	4
2	不明	1
3	男性	0

```

PROC MEANS DATA = seibetsu NWAY NOPRINT CLASSDATA = sex ORDER = freq;
  CLASS sex ;
  OUTPUT OUT = result(DROP=_type_ _freq_) N = kensu ;
RUN ;

```

また、上記のプログラムでは ORDER オプションも指定している。ORDER オプションも Version8 から追加されたオプションで、出力されるデータの出力順を指定することが出来る(上記のプログラムでは、出力は件数の多い順)。他にも分類変数の組み合わせを制御する TYPES ステートメントや分類変数の組み合わせの数を制御する WAY ステートメントなどの便利なステートメントも新たに追加された。紙面の都合上、ここでは省略させていただくが、TYPES ステートメントや WAY ステートメントの詳細やそれ以外の拡張点についてはオンラインヘルプ等を参考にしていきたい。

2. TABULATE プロシジャ

次に TABULATE プロシジャを用い集計表を作成する場合のデータに含まれる欠損値の

取扱について説明する。変数内の欠損値がどのようにアウトプットに影響するかは、Tabulate プロシジャ内でどのように変数を使用するかに影響される。下記の表は Tabulate プロシジャがどのように欠損値を取り扱うかについて記している

	デフォルトの扱い	表示を変更するには
1. 分類変数に欠損値が含まれる場合	テーブルより除外する	TABULATE ステートメント、もしくは CLASS ステートメントにて MISSING を指定
2. 特定のセルに該当するオブザベーションの分析変数がすべて欠損値の場合	(N と NMISS 以外の) すべての統計量において欠損値が表示される	TABLE ステートメントにて MISSTEXT= を指定
3. 特定の水準のデータが存在しない場合	結果テーブルに水準が表示されない	TABULATE ステートメントにて CLASSDATA= を指定

下記で、オプションの有無によってどのように欠損値が取り扱われるかを詳しく説明する。欠損値がないデータにて出力される結果は以下のようになる。

[オプションなしの場合]

```
PROC TABULATE DATA = class ;
  CLASS age sex ;
  VAR height weight ;
  TABLES age * (height * sum weight * sum), sex ;
RUN ;
```

			Sex	
			F	M
Age	Height	Sum		
	Weight	Sum		
11	Height	Sum	51.90	57.50
	Weight	Sum	50.50	65.00
12	Height	Sum	116.10	181.10
	Weight	Sum	161.50	310.50
13	Height	Sum	121.80	62.50
	Weight	Sum	182.00	84.00
14	Height	Sum	127.10	132.50
	Weight	Sum	182.50	215.00
15	Height	Sum	129.00	133.50
	Weight	Sum	224.50	245.00
16	Height	Sum	.	72.00
	Weight	Sum	.	150.00

2.1 分類変数に欠損値が含まれる場合

分類変数に欠損値が含まれる場合、その値は出力より除外される。欠損値を一つの水準として集計を行いたい場合は MISSING オプションを指定する。このオプションを指定する事により、欠損値が一つの水準として扱われ、集計結果に反映される。

例) 年齢が不明な人を有効な水準として集計する場合

[MISSINGオプションを使用した場合]

```
PROC TABULATE DATA = class ;
  CLASS age sex / MISSING ;
  VAR height weight ;
  TABLES age * (height * sum weight * sum), sex ;
RUN ;
```

			Sex	
			F	M
.	Height	Sum	127.10	132.50
	Weight	Sum	192.50	215.00
11	Height	Sum	51.90	57.50
	Weight	Sum	50.50	95.00
12	Height	Sum	116.10	181.10
	Weight	Sum	110.50	10.50
13	Height	Sum	.	82.50
	Weight	Sum	.	84.00
15	Height	Sum	.	133.50
	Weight	Sum	224.50	245.00
16	Height	Sum	.	72.00
	Weight	Sum	.	150.00

欠損値が一つの水準として集計

2.2 特定のセルに該当する分析変数の値がすべて欠損値の場合

特定のセルに該当する分析変数の値がすべて欠損値だった場合、すべての統計量（N と NMISSED 以外）に欠損値が表示される。結果表示に欠損値以外の値を使用したい場合は、MISSTEXT オプションを使用する。これにより、MISSTEXT オプションで指定した値が欠損値の代わりに使用される。

例) 特定のセルに分類されたオブザベーションの身長、体重の値が欠損値だった場合、欠損値の代わりに“未測定”と表示させたい。

[MISSTEXTオプションを使用した場合]

```
PROC TABULATE DATA = class ;
  CLASS age sex / MISSING ;
  VAR height weight ;
  TABLES age * (height * sum weight * sum), sex
  / MISSTEXT = "未測定" ;
RUN ;
```

			Sex	
			F	M
.	Height	Sum	127.10	132.50
	Weight	Sum	192.50	215.00
11	Height	Sum	51.90	57.50
	Weight	Sum	50.50	95.00
12	Height	Sum	116.10	181.10
	Weight	Sum	110.50	10.50
13	Height	Sum	.	82.50
	Weight	Sum	.	84.00
15	Height	Sum	.	133.50
	Weight	Sum	224.50	245.00
16	Height	Sum	未測定	72.00
	Weight	Sum	未測定	150.00

欠損値の代わりに指定した文字列が表示

2.3 特定の水準に該当するデータがない場合

特定の水準に該当するデータがなかった場合、その水準は結果テーブルに表示されない。データの有無に関わらず、特定の水準数での結果を出力したい場合は、MEANS プロシジャでも記述した CLASSDATA オプション、もしくは PRELOADFORMAT オプションを使用する。出力させる表に含める分類変数の組み合わせをデータにしておき、CLASSDATA で指

定することにより、特定の水準数での表を出力する事ができる。また、CLASSDATA オプションと共に EXCLUSIVE オプションを使用すると、CLASSDATA に含まれない分類変数の組み合わせを除外した表を出力する事が可能である。CLASSDATA、EXCLUSIVE オプションは Version8 より拡張された機能である。

例) 14 歳のデータが欠落しているが、11 歳から 16 歳までの各年齢を結果テーブルに出力したい、また 11 歳以下のデータがある場合はそれを除外したい場合。

```
PROC TABULATE DATA = class CLASSDATA = age_data EXCLUSIVE ;
  CLASS age sex ;
  VAR height weight ;
  TABLES age * (height * sum weight * sum), sex / MISSTEXT = "未測定" ;
RUN;
```

<オプション無し>

Age				
9	Height	Sum	60.00	71.00
	Weight	Sum	83.00	98.00
11	Height	Sum	51.30	57.50
	Weight	Sum	50.50	85.00
12	Height	Sum	116.10	181.10
	Weight	Sum	161.50	310.50
13	Height	Sum	121.80	62.50
	Weight	Sum	182.00	84.00
15	Height	Sum	128.00	133.50
	Weight	Sum	224.50	245.00
16	Height	Sum	未測定	72.00
	Weight	Sum	未測定	150.00

出力対象でない値だが、データが存在するため結果に出力される

<オプションあり>

Age			Sex	
			F	M
11	Height	Sum	51.30	57.50
	Weight	Sum	50.50	85.00
12	Height	Sum	116.10	181.10
	Weight	Sum	181.50	310.50
13	Height	Sum	121.80	62.50
	Weight	Sum	182.00	84.00
14	Height	Sum	未測定	未測定
	Weight	Sum	未測定	未測定
16	Height	Sum	128.00	133.50
	Weight	Sum	224.50	245.00
	Height	Sum	未測定	72.00
	Weight	Sum	未測定	150.00



データの欠落より、出力されていなかった「14」が出力される

CLASSDATA オプションと同じように、出力したい水準のフォーマットを作成し、それを指定する事でデータの有無に関わらず出力したい組み合わせを出力する事も可能である。その場合は PRINTMISS オプションと PRELOADFMT オプションを指定し、フォーマットステートメントにて作成したフォーマットを指定する。

```
PROC FORMAT :
  VALUE agefmt
  11=11
  12=12
  13=13
  14=14
  15=15
  16=16;
RUN ;

PROC TABULATE DATA = class ;
  CLASS sex;
  CLASS age / PRELOADFMT ;
  VAR height weight ;
  FORMAT age agefmt. ;
  TABLES age*(height*sum weight*sum),sex
  / PRINTMISS MISSTEXT = "未測定" ;
RUN;
```

3. DATASETS プロシジャ

最後に DATASETS プロシジャの拡張機能と便利な使い方について説明する。DATASETS プロシジャは、SAS データライブラリ中の SAS ファイル一覧の作成、名前の変更、コピーや削除などを行うプロシジャで柔軟なファイル操作を可能とする。

3.1 ライブラリ内の全てのメンバを削除

DATASETS プロシジャでは KILL オプションを指定することにより、ライブラリ内の全てのメンバを削除することができる。アプリケーションなどで最後にこのオプションをつけたプログラムを組み込むだけで、一時的に作成されたデータを一括削除でき非常に便利である。以下のサンプルコードでは WORK ライブラリ内を全て削除している。LIB=オプションを省略した場合は WORK ライブラリが削除され、さらに NOLIST オプションを指定しているため、SAS ログにメンバリストが出力されない。

```
PROC DATASETS LIB = work NOLIST KILL ;  
QUIT ;
```

3.2 指定したメンバ以外の削除

DATASETS プロシジャの SAVE ステートメントを使用すると、指定したメンバ以外の全てのメンバが削除される。このステートメントは多くのデータが格納されたライブラリから少数のデータのみを保持しておきたい場合に非常に便利である。

```
PROC DATASETS LIB = work NOLIST ;  
    SAVE result ;  
QUIT ;
```

3.3 データセット名の変更

DATASETS プロシジャの CHANGE ステートメントを使用すると、ライブラリ内のメンバ名の変更が可能となる。以下のサンプルコードでは AIRLINE ライブラリの STAFF を NEWSTAFF に変更している。

```
PROC DATASETS LIB = airline NOLIST ;  
    CHANGE staff = newstaff ;  
QUIT ;
```

3.4 インデックスの作成

DATASETS プロシジャの INDEX CREATE ステートメントを使用すると、データのインデックスを容易に作成することが可能となる。インデックスを作成することにより、以

下の2つの点でパフォーマンスの向上が望める。

- WHERE 式が含まれているプログラムを実行した場合、オブザベーションのサブセットへの素早いアクセスが可能
- SORT プロシジャによるソート処理を事前に行うことなく、BY グループ処理によるインデックス順のデータ取り出しが可能

以下のサンプルコードでは、airline.mechanic というデータセットに単一インデックス employeeidnumber と複合インデックス addid (変数「state」と「city」)を作成する。さらに UNIQUE オプションを使用することにより、変数 employeeidnumber に同一の値の組み合わせが存在しないよう指定している。

```
PROC DATASETS LIB = airline NOLIST ;
  MODIFY mechanic;
  /*重複を許可しないインデックスの作成*/
  INDEX CREATE employeeidnumber / UNIQUE ;
  /*複合インデックスの作成*/
  INDEX CREATE addid = (state city) ;
QUIT;
```

3.5 一貫性制約の作成

DATASETS プロシジャの IC CREATE ステートメントを使用することで、Version8 から一貫性制約を作成することが可能となった。また、MESSAGE=オプションとの併用でエラー時のメッセージの設定も行うことが可能である。一貫性制約を使用することにより、データの矛盾や間違いを未然に防ぐことが出来、よりデータの整合性が保たれる。以下のサンプルコードでは、3つの一貫性制約 (ok_job、ok_cost、nnull_id) を作成している。

```
PROC DATASETS LIB = airline NOLIST ;
  MODIFY mechanic;
  /*変数「jobclass」には"ME1","ME2","ME3"のみが格納可能*/
  IC CREATE ok job = CHECK (WHERE=(jobclass IN ('ME1' 'ME2' 'ME3')))
  MESSAGE = 'Job Class must be ME1, ME2, or ME3';
  /*変数「ANNUALSALARY」には100000以下の値のみが格納可能*/
  IC CREATE ok cost = CHECK (WHERE=(annualsalary < 100000))
  MESSAGE = 'Annual Salary must be less than 100000';
  /*変数「employeeidnumber」にはNULL値を許可しない*/
  IC CREATE nnull id = NOT NULL (employeeidnumber)
  MESSAGE = 'You must provide an Employee ID Number';
QUIT;
```

変数「jobclass」に値「ME4」を代入しようとすると、一貫性制約「ok_job」に適合しないため、エラーとなり、以下のようなメッセージがログに出力される。

```
ERROR: Job Class must be ME1, ME2, or ME3 データセット WORK.MECHANIC
への追加/更新に失敗しました。データ値が一貫性制約 ok_job に適合しません。
NOTE: This insert failed while attempting to add data from VALUES clause 1 to the data set.
NOTE: テーブルを矛盾のない状態に戻すため上記のエラー 前の挿入を削除します。
```

一貫性制約は SQL プロシジャ、DATASETS プロシジャ、SCL でのみ生成・追加・削除が可能であり、DATA ステップでは取り扱うことはできない。なお、作成した一貫性制約の削除には、DATASETS プロシジャの IC DELETE ステートメントを使用する。

3.6 DATA ステップとのパフォーマンス比較

以下のサンプルコードでは、同一データ(件数約 370 万件)を対象に DATA ステップと DATASETS プロシジャを使用してフォーマットを割り当てる処理を行った際のパフォーマンスを比較している。DATA ステップでは、フォーマットを適用する場合でも、1 オプザベーションずつデータ読み込んでいくのに対して、DATASETS プロシジャではディスクリブタ部の情報を読み込み、書き換えるだけなので処理時間が大幅に短縮出来る。そのため、フォーマットの適用や変数名の変更などの処理では DATASETS プロシジャを使用することが非常に有効である。

```
/*データステップによるフォーマット処理*/  
DATA sample ;  
    SET sample ;  
    FORMAT birthday YYMMDD8. ;  
RUN ;
```



NOTE: DATA ステートメント 処理 :	
処理時間	4:57.16
CPU 時間	12.00 秒

```
/*DATASETSプロシジャによるフォーマット処理  
PROC DATASETS LIB = work NOLIST ;  
    MODIFY sample ;  
    FORMAT birthday YYMMDD8. ;  
QUIT ;
```



NOTE: PROCEDURE DATASETS 処理 :	
処理時間	0.03 秒
CPU 時間	0.02 秒

4. まとめ

今回の論文では、紙面の都合上一部のオプションや機能の紹介に留まっているが、今回紹介した以外にも多数の拡張が施されている。また、本稿だけでは詳細な説明まで至らなかったため、興味を持たれた方はぜひオンラインヘルプやマニュアルの方も参考にして頂ければ幸いである。Version8 へのバージョンアップに伴い、オープンな環境で SAS システムの利用が可能となり、DCOM/COM といったアーキテクチャーを利用した GUI ツールもリリースされているが、依然として SAS の言語体系は強力なものであり、多くのユーザから支持されている。本稿がそういったユーザにとっての一助となれば幸いである。

なお本稿についての質問、意見などがあれば、下記まで。

Yoshie.Shibuya@sas.com

Takafumi.Hiwada@sas.com

Naoko.Sakota@sas.com

口頭論文発表
経営・経済

日本SASユーザー会 (SUGI-J)

SAS ソフトウェアを利用した CIR++モデルの パラメータ推定と金利パス生成

岸田 則生

株式会社 CRC ソリューションズ
金融システム部

Parameter Estimation and Path Generation of Interest Rates for CIR++ Model with SAS Software

Norio Kishida

CRC Solutions Corp.
Financial Systems Development Dept.

要 旨

瞬間金利モデルの一つである CIR++モデルの日本市場におけるパラメータ値を SAS/ETS ソフトウェアに含まれる非線形最小自乗法を用いて推定した。非観測量である瞬間金利の標本値には、1、2、3 ヶ月もの短期金利を採用し、パラメータ値の比較を試みた。金利モデルは短期金利ばかりでなく、長期金利を含む金利の期間構造の将来予測にも利用されるので、期間構造を標本値とするパラメータ推定も行った。しかし、パラメータに課せられた制限値内の推定値を得ることは出来なかった。CIR モデル同様、生成金利パスが正值であることが確認された。

キーワード： SAS/ETS ソフトウェア、CIR++モデル、金利、平均回帰過程、確率微分方程式、時系列モデル、自己回帰モデル、非線形最小自乗法

1. 緒言

銀行経営におけるバンキング勘定の金利変動リスクを補足するための Earning at Risk 手法や、金利を原資産とする派生証券の価格評価には、将来金利が時間的にどのように変動するかを記述する金利の期間構造モデルが使用される。期間構造モデルには大きく分けると確率微分方程式で定式化される確率変動モデルと自己回帰式で記述される時系列分析モデルとがある。Brigo-Mercurio による CIR++モデル¹⁾の元になった CIR(Cox-Ingersoll-Ross) モデル²⁾は前者に属する。CIR モデルは金利時間変動の性質である正值性と平均回帰性を併せ持つが、現時点の金利の期間構造を反映できない点で問題がある。この問題点を克服するモデルとして Hull-White による拡張 CIR モデル³⁾と Brigo-Mercurio による CIR++モデルが知られている。拡張 CIR モデルは CIR モデルのパラメータを時間依存とする事により現時点における期間構造を再現するようにしたため、期間構造の解析解は得られず、また数值的に解くのも極めて困難であり、実務上ほとんど採用されていない。それに対して Brigo-Mercurio による CIR++モデルは解析解が得られるので、実務上魅力的なモデルである。

CIR++モデルは金利市場で直接観測できない瞬間的なゼロ・レートを記述する確率変動金利モデルで

ある。モデルに内在するパラメータを推定するには、ゼロ・レートの市場データあるいはゼロ・レートから導かれる割引債価格が必要である。本論文ではまずゼロ・レートとして1、2、3ヶ月ものゼロ・レートを標本値としてモデル・パラメータを推定した。パラメータの推定は最尤法が望ましいが、CIR++モデルに従うゼロ・レートの確率密度関数を求めるのは困難なので尤度関数の導出も難しい。そこで、今回はCIR++モデル確率微分方程式の離散化から得られる差分方程式である自己回帰式にSAS/ETSソフトウェアの非線形最小自乗法を適用してパラメータ推定を行った。さらに推定パラメータを用いて金利の時系列的な生成を行い、金利が負にならないことを確認した。現今の本邦の金利状況において、Hull-Whiteによる拡張 Vasicek モデルのようなブラウン運動から導かれるモデルでは、生成した金利バスの半分程度が負になってしまい、Earning at Risk の計算に使用するには問題がある。この問題が生じないCIR++モデルのパラメータが推定できたことは大いに意味があると考えられる。

Earning at Risk の算出では将来時点での長期金利が必要となる。長期金利の情報が反映されない短期金利の市場データのみで推定したパラメータを用いて算出した長期金利が、Earning at Risk の算出に最適かどうかはかなり疑問がある。そこで長期金利情報を含むゼロ・イールド・カーブ・データからパラメータを推定できることが望ましい。幸いCIR++モデルでは割引債価格の解析式が求まるので、ゼロ・イールド・カーブから求めた割引率を標本値としてパラメータ推定が行える。本論文では短期金利から推定したパラメータ値と金利の期間構造の情報をすべて含んだゼロ・イールド・カーブから求めたパラメータ値との比較も行った。

2. CIR++モデル

CIRモデルは瞬間的なゼロ・レート(スワップ・レート) x_t が以下の確率微分方程式に従って時間変動していると仮定する金利変動モデルである。

$$dx_t = k(\theta - x_t)dt + \sigma \sqrt{x_t}dW_t \quad (1)$$

ここで、 k 、 θ 、 σ がモデル・パラメーターである。これらのパラメーターが正の値を取る場合、このモデルは市場で観測される短期金利が持つ経験的な性質である、(1) 金利は負にならない、(2) 金利は長期的に見るとある平均的な金利の周りを変動するという平均回帰性を有している。そのため、 k を平均回帰速度パラメーター、 θ を平均回帰レベル・パラメーターと呼んでいる。また、 σ は金利のふらつきの度合いを表すパラメーターである。しかし、このモデルは x_t の初期値 x_0 のみで初期イールド・カーブ(金利の期間構造)が決まってしまうので、市場で観測されるイールド・カーブに適合しないという重大な欠陥を持つ。

CIR++モデルは確率変動しない時間に確定的な関数 $\varphi(t)$ を x_t に加えた r_t 、すなわち

$$r_t = x_t + \varphi(t) \quad (2)$$

が瞬間的なゼロ・レートの動きを表すとしたモデルである。ここで導入した関数 $\varphi(t)$ により、CIR++モデルは市場で観測される初期イールド・カーブに適合可能となる。確率微分に関する伊藤の公式を利用すると、 r_t に関する確率微分方程式

$$dr_t = \left[k\theta + k\varphi(t) + \frac{d\varphi(t)}{dt} - kr_t \right] dt + \sigma \sqrt{r_t - \varphi(t)}dW_t \quad (3)$$

を得る。

Brigo-Mercurio は r_t から導かれる割引債価格 (割引関数) あるいはそれと等価なゼロ・レート の期間構造が市場価格に適合するためにはシフト関数 $\varphi(t)$ が

$$\varphi(t) = f^M(0, t) - f(0, t) \quad (4)$$

$$f(0, t) = 2k\theta \frac{e^{th} - 1}{2h + (k+h)(e^{th} - 1)} + x_0 \frac{4h^2 e^{th}}{[2h + (k+h)(e^{th} - 1)]^2} \quad (5)$$

$$h = \sqrt{k^2 + 2\sigma^2} \quad (6)$$

でなければならないことを示した。ここで、 $f^M(0, t)$ は現時点における瞬時的な市場フォワード・レートである。瞬時的なフォワード・レートは市場で取引される量ではないので、実際は市場データから得られる割引関数 $P^M(0, T)$ の期間構造から

$$f^M(0, t) = -\frac{\partial \ln P^M(0, t)}{\partial t} \quad (7)$$

の関係を用いて計算する。

確率微分方程式 (1) の x_t の解が非心カイ二乗分布でありその解析解が既知であることを利用して、Brigo-Mercurio は時点 t 、満期時点 T の割引債価格 $P(t, T)$ とゼロ・レート $R(t, T)$ が以下の式で与えられることを示した。

$$P(t, T) = \frac{P^M(0, T)A(0, t)e^{-B(0, t)x_0}}{P^M(0, t)A(0, T)e^{-B(0, T)x_0}} A(t, T)e^{-B(t, T)\{r_t - \varphi(t)\}} \quad (8)$$

$$R(t, T) = \frac{1}{T-t} \left[\ln \frac{P^M(0, T)A(0, t)e^{-B(0, t)x_0}}{P^M(0, t)A(0, T)e^{-B(0, T)x_0}} - \ln A(t, T) + B(t, T)\{r_t - \varphi(t)\} \right] \quad (9)$$

ここで

$$A(t, T) = \left[\frac{2he^{(k+h)(T-t)/2}}{2h + (k+h)(e^{th} - 1)} \right]^{2k\theta/\sigma^2} \quad (10)$$

$$B(t, T) = \frac{2(e^{th} - 1)}{2h + (k+h)(e^{th} - 1)} \quad (11)$$

である。

3. モデル・パラメーター推定法

CIR++モデルのパラメーター推定法として、(1) 市場で観測される短期金利を瞬時的なゼロ・レート の代用金利として用いる方法と、(2) ゼロ・レート の期間構造を用いる方法を試みる。

3.1 短期金利法

確率微分方程式 (1) に従う x_t の解は既知で非心カイ二乗分布に従うことが知られているので、条件付き推移確率密度 $p(x_{t+dt}|x_t)$ も良く知られている。従って、 $r_t = x_t + \varphi(t)$ の関係で結ばれる r_t の条件付き推移確率密度 $p(r_{t+dt}|r_t)$ も原理的には導出可能である。しかし、その推移確率密度から尤度関数を導いて最尤法でパラメーター推定を行うのは、尤度関数が非常に複雑になり現実的には困難である。そのため、通常は r_t が従う確率微分方程式 (3) を離散近似することによって得られる自己回帰式から導かれる

正規分布に対する尤度関数を用いてパラメーター推定を行う。離散近似では微小時間 Δt に関して一次の Euler 近似および二次の Milstein 近似がよく使用される。確率微分方程式のドリフト関数および拡散関数が定数でないとき、Euler 近似はあまりよい近似でないことは知られているが、確率微分方程式 (3) の Milstein 近似に基づく尤度関数を導くのは困難なので、本論文では Euler 近似を採用する。

式 (3) を Euler 近似して離散化すると

$$r_{t+\Delta t} = r_t + \mu(t) \Delta t + \Sigma(t) \sqrt{\Delta t} \varepsilon(t) \quad (12)$$

となる。ただし、

$$\mu(t) = k\theta + k\varphi(t) + \frac{d\varphi(t)}{dt} - kr_t \quad (13)$$

$$\Sigma(t) = \sigma \sqrt{r_t - \varphi(t)} \quad (14)$$

と置いた。また、 $\varepsilon(t)$ は全ての t に関して独立な標準正規分布をする確率変数である。すなわち、平均 μ 、分散 σ^2 の正規分布を $N(\mu, \sigma^2)$ と表すと推移確率密度 $p(r_{t+\Delta t}|r_t)$ は

$$p(r_{t+\Delta t}|r_t) \sim N\left(r_t + \mu(T), \Sigma(t)^2 \Delta t\right) \quad (15)$$

与えられる。これから $N+1$ 個の時点 $t=0, 1, 2, \dots, N$ に対する尤度関数は容易に導けるが、 $\mu(t)$ および $\Sigma(t)$ が定数ではないため、最尤法でパラメーターを推定するには、SAS/IML ソフトウェアに含まれる非線形最適化法を用いてプログラミングしなければならない。そこで、本論文ではもっとプログラミングが簡単な、SAS/ETS ソフトウェアを用いた時系列標本に対する近似的パラメーター推定問題に変換する。

r_t に対する自己回帰式 (12) を時点 $t=0, 1, 2$ に対して書き下すと

$$r_1 = r_0 + \left[\frac{df^M(0, t)}{dt} \right]_{t=0} \Delta t + \sigma \sqrt{x_0} \sqrt{\Delta t} \varepsilon(0) \quad (16)$$

$$r_2 = r_1 + \left[k\theta - k\varphi(1) + \left[\frac{df^M(0, t)}{dt} \right]_{t=1} - kr_1 \right] \Delta t + \sigma \sqrt{r_1 - \varphi(1)} \sqrt{\Delta t} \varepsilon(1) \quad (17)$$

となる。時点 $t=0, 1, 2, \dots, N$ の $N+1$ 個の一連の標本値 $r_0, r_1, r_2, \dots, r_N$ から、相隣りあう 3 時点に対して時間移動的に上式を適用し、最小自乗法によってパラメーター推定を行う。本来、 $\varphi(t)$ は時間依存の関数なので、 $\varphi(3)$ と $\varphi(N)$ の値は当然異なる。従って、相隣りあう 3 時点の標本値が多数あるとしてパラメーター推定を行うのは、一組の $r_0, r_1, r_2, \dots, r_N$ に対して最尤法でパラメーター推定を行う場合の近似に過ぎない。

$\varepsilon(0)$ と $\varepsilon(1)$ は独立な正規分布であることを仮定しているの、それとは独立の正規分布 ε を導入すると、上式はまとめて

$$r_2 = r_0 + \left[k(\theta - k\varphi(1) - r_1) + \left[\frac{df^M(0, t)}{dt} \right]_{t=0} + \left[\frac{df^M(0, t)}{dt} \right]_{t=1} \right] \Delta t + \sigma \sqrt{x_0 + r_1 - \varphi(1)} \sqrt{\Delta t} \varepsilon \quad (18)$$

ともかける。

自己回帰式 (16) と (17) あるいは式 (18) は、SAS/ETS ソフトウェアを用いるとパラメーターの最小自乗推定がきわめて容易に実行できる。

3.2 ゼロ・レート期間構造法

CIR++モデルのゼロ・レート期間構造は式(9)で与えられるが、現時点の期間構造を再現するモデルなので、 $R(0, T) = R^M(0, T)$ になっている。従って、 $R(0, T)$ からはパラメーター推定可能な表現式は得られない。そこで、 $R(\Delta t, \Delta t + T) - R(0, T)$ という量を作ると

$$\begin{aligned} R(\Delta t, \Delta t + T) - R(0, T) &= \frac{1}{T} \left[(\Delta t + T)R^M(0, \Delta t + T) - TR^M(0, T) - \Delta t R^M(0, \Delta t) \right. \\ &\quad \left. + \ln \frac{A(0, \Delta t + T)}{A(0, \Delta t)A(\Delta t, \Delta t + T)} - \{B(0, \Delta t) - B(0, \Delta t + T)\}x_0 \right] \\ &\quad + \frac{1}{T}B(\Delta t, \Delta t + T)x_{\Delta t} \end{aligned} \quad (19)$$

となる。 $x_{\Delta t}$ は非心カイ二乗分布に従い、その平均 $E[x_{\Delta t}]$ と分散 $V[x_{\Delta t}]$ は

$$E[x_{\Delta t}] = \theta + (x_0 - \theta)e^{-k\Delta t} \quad (20)$$

$$V[x_{\Delta t}] = \frac{\sigma^2}{k} \left[x_0 e^{-k\Delta t} + \frac{\theta}{2}(1 - e^{-k\Delta t}) \right] (1 - e^{-k\Delta t}) \quad (21)$$

で与えられることが知られている。従って、 $\Delta R = R(\Delta t, \Delta t + T) - R(0, T)$ の平均 $E[\Delta R]$ と分散 $V[\Delta R]$ は容易に得ることができ、

$$\begin{aligned} E[\Delta R] &= \frac{1}{T} \left[(\Delta t + T)E[R^M(0, \Delta t + T)] - TE[R^M(0, T)] - \Delta t E[R^M(0, \Delta t)] \right. \\ &\quad \left. + \ln \frac{A(0, \Delta t + T)}{A(0, \Delta t)A(\Delta t, \Delta t + T)} - \{B(0, \Delta t) - B(0, \Delta t + T)\}x_0 \right] \\ &\quad + \frac{1}{T}B(\Delta t, \Delta t + T)E[x_{\Delta t}] \end{aligned} \quad (22)$$

$$V[\Delta R] = \frac{1}{T^2}B(\Delta t, \Delta t + T)^2 V[x_{\Delta t}] \quad (23)$$

となる。パラメータの推定はゼロ・レートの標本値から各満期毎の標本平均と標本分散を算出し、それを式(22)と式(23)で最小自乗適合すれば実行できる。分散だけでなく平均も用いるのは、分散に対する θ の感応度が小さく分散だけからでは θ を決定できないからである。式(22)と式(23)による最小自乗推定も、SAS/ETSソフトウェアを用いると簡単に実行できる。

4. パラメーター推定結果

パラメーター推定に使用した金利は日本市場における2002年10月から2003年3月までの半年間の満期1、2、3ヶ月もの金利を連続複利のゼロ・レートに変換したものである。連続複利を用いるのは式(3)が連続複利を仮定して定式化されているからである。実際の推定にはSAS/ETSソフトウェア中のMODEL Procedureを使用した。推定すべきパラメーターは k 、 θ 、 σ 、 x_0 の4パラメーターである。また、パラメーター推定には以下の各条件を課した。

$$0 < k \quad (24)$$

$$0 < \theta \quad (25)$$

$$0 < \sigma \quad (26)$$

$$\sigma^2 < 2k\theta \quad (27)$$

$$0 < x_0 \quad (28)$$

$$\theta < x_0 \quad (29)$$

最初の 5 条件はオリジナルの CIR モデルに従う金利 x_t が正でかつ平均回帰性を持つための条件であり、最後の条件は CIR++モデルに従う金利 r_t が正であるための条件である。

まず、短期金利を瞬時的なゼロ・レートの代用金利とした場合の推定結果について述べる。表 1 に式 (16)、(17) で推定した結果を、表 2 に式 (18) で推定した結果を示す。なお、時間の単位は年に取っている。

表 1: 式 (16)、(17) による推定値。

代用金利	k	θ	σ	x_0
1 ヶ月もの	196.9	0.000552	0.47	0.000552
2 ヶ月もの	123.3	0.000654	0.40	0.000654
3 ヶ月もの	66.6	0.000752	0.32	0.000752

表 2: 式 (18) による推定値。

代用金利	k	θ	σ	x_0
1 ヶ月もの	139.8	0.000551	0.39	0.000551
2 ヶ月もの	94.1	0.000647	0.35	0.000647
3 ヶ月もの	68.0	0.000745	0.32	0.000745

表 1、2 を見ると代用金利の取り方によって平均回帰速度 k が大きく変化していることが分かる。一方、平均回帰金利 θ と拡散パラメータ σ は代用金利の違いによる差はそれほど大きくない。平均回帰金利の値は推定に使用した標本金利の平均値に近い値が得られている。表に示した有効桁数では $\theta = x_0$ のように見えるが、実際は条件 $\theta < x_0$ を満たしている。表 1 と 2 の間で大きく値が異なるのは 1 ヶ月もの と 2 ヶ月もの の平均回帰速度である。式 (18) は誤差分布の独立性を仮定して導いた式なので、この違いは相隣りあう時点間の誤差が独立ではないことを示唆している可能性がある。

次に、ゼロ・レートの期間構造を使用した場合の推定結果について述べる。表 3 に Δt を 1 日、5 日、25 日にとった場合の推定結果を示す。この推定法では $0 < k$ および $0 < \theta$ の範囲内の推定値を得られなかった。図 1 に ΔR の標本標準偏差と CIR++モデルの推定標準偏差を満期を変数にとって示す。図から分かるようにその再現性はかなり良い。 $0 < k\theta$ なので平均回帰金利は正となるが、平均回帰速度が負値なので、金利が時間的に平均回帰せず発散してしまう。従って、この推定値を実際の金利生成に使用することは出来ない。また、この推定値は短期金利から推定した表 1、2 の値とも全く異なる。短期金

利の推定値を用いて ΔR の標準偏差を算出した場合、標本標準偏差を全く再現しない。従って、ゼロ・レート・期間構造に基づく推定モデルの構成法には何らかの問題があるのかもしれない。

表 3: 式 (22)、(23) による推定値。

Δt	k	θ	σ	x_0
1 日	-0.180	-0.8465	0.051	1.0E-8
5 日	-0.190	-0.1592	0.045	1.0E-8
25 日	-0.213	-0.0134	0.042	1.0E-8

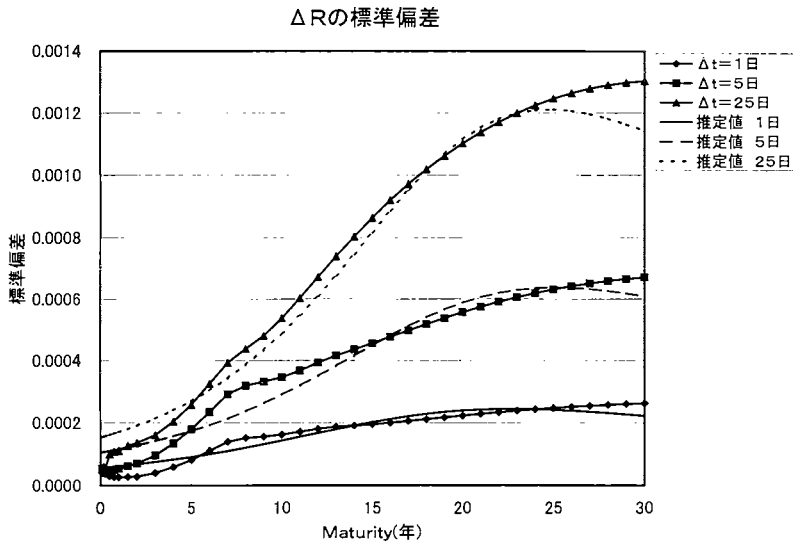


図 1: ΔR の標本標準偏差と CIR++モデルの推定標準偏差。

5. 金利パス生成

図 2 に表 1 の 1 ヶ月もの代用金利から推定したパラメーター値と式 (1) を用いて生成した CIR モデルの 1800 日間に渡る金利パスを示す。CIR モデルが持つ金利の正值性と平均回帰性が満たされていることが見てとれる (実際の金利生成では無限小時間 dt を有限時間 Δt で置き換えるので、負金利の発生が起こるが、平均回帰性により、すぐに正金利に戻る。この図からは負金利を除いてある)。

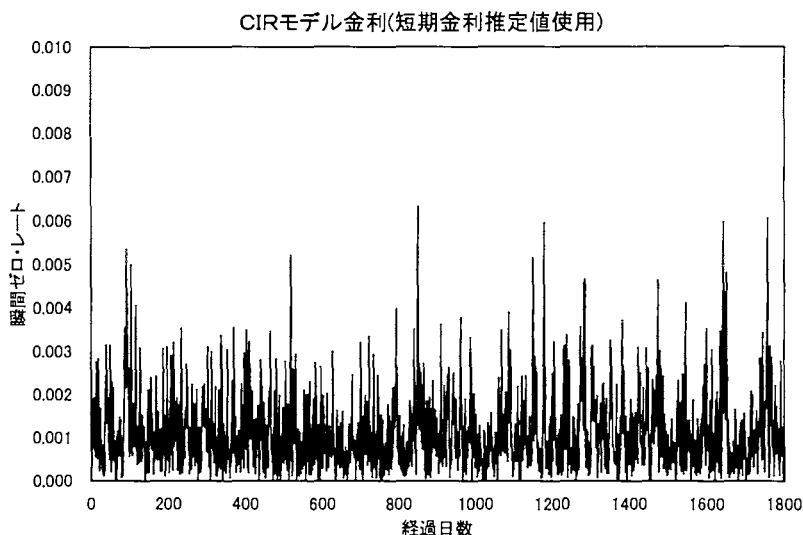


図 2: 短期金利から推定したパラメーター値による CIR モデルの金利パス。

図 3 に表 3 の Δt を 1 日にとったときのパラメーター値を使用した CIR モデルの金利パスを示す。平均回帰速度が負値なので金利が発散してしまっているのがわかる。

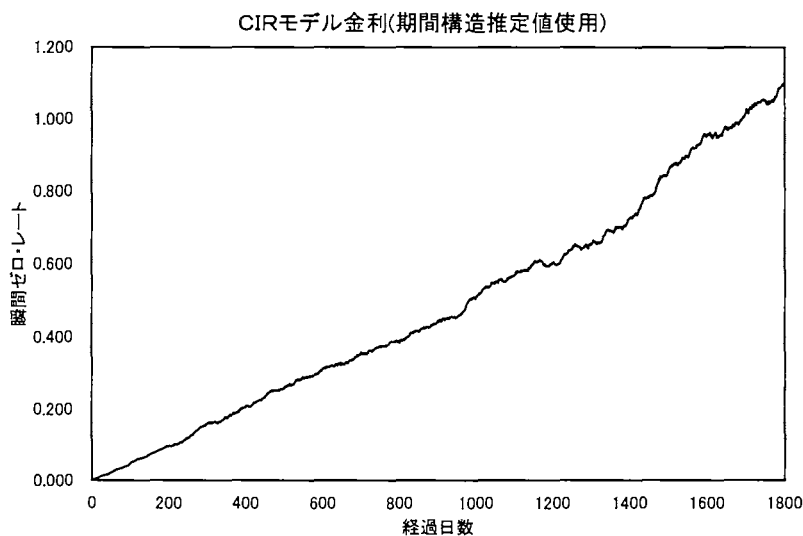


図 3: 金利の期間構造から推定したパラメーター値による CIR モデルの金利パス。

図 4 に表 1 の 1 ヶ月もの代用金利から推定したパラメーター値と式 (1)、(2)、(9) を用いて生成した満期 1 ヶ月の CIR++モデルによるゼロ・レート時間変化を示す。比較のために CIR モデルによる金利

パスも示した。CIR モデルによる金利パスが低下傾向にあるのに、CIR++モデルによる金利が上昇するのは、CIR++モデルでは初期期間構造から求まるインプライド・フォワード・イールドが金利の下限値を決めているからである。

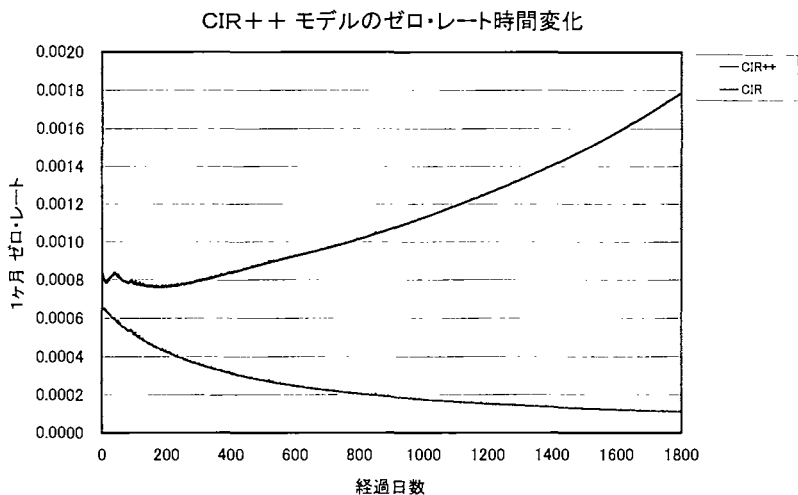


図 4: 金利の期間構造から推定したパラメーター値による CIR モデルの金利パス。

6. 結言

CIR++モデルのパラメーター推定を2種類の方法で行った。一つは短期金利を瞬間ゼロ・レートの代用金利として用い、SAS/ETS ソフトウェアの時系列モデルの取り扱いでパラメーターを推定した。もう一つは金利の期間構造情報を用いる方法であるが、パラメーターに課せられた制限域内で推定値を得ることが出来なかった。その原因としては、パラメーター推定法の誤りが考えられる。推定パラメーターを使用して瞬間ゼロ・レートを生成したところ、金利の正值性と平均回帰性が確認できた。さらに、満期1ヶ月のゼロ・レートを生成したところ、インプライド・フォワード・イールドに下支えされて金利が変動するという、理論的な事実を数値的に確かめられた。SAS ソフトウェアにより、ほとんどプログラミングすることなく非線形問題のパラメーターを推定出来たのは、金融工学分野への SAS ソフトウェアの適応可能性を示したといえる。

参考文献

- 1) D. Brigo and F. Mercurio : “*A deterministic-shift extension of analytically-tractable and time-homogeneous short-rate models.*”, Finance and Stochastics, **5**, 369-388(2001).
- 2) J. C. Cox, J. E. Ingersoll and S. Ross : “*A theory of the term structure of interest rates*”, Econometrica, **2**, 385-407(1985).
- 3) J. Hull and A. White : “*Pricing interest-rate-derivative securities*”, The Review of Financial Studies, **4**, 573-592(1990).

要旨

本稿は利益性を高めた新しいモデル“Profitable Model”を提案する。このモデルは従来の Default Model に Profit Model の要素を取り入れた与信モデルである。特徴は精度をわずかに犠牲にして、従来モデルよりも高い利益を目指す点にある。構造は分析途中で目標変数が変化する 2 段階モデリングである。1 段階のモデルはデフォルト/正常顧客の 2 値フラグを目標変数にして、精度が定常になる途中で分割を止める。2 段階のモデルは収益/損失の 2 値フラグを目標変数に変更し、1 段階で分類されたグループごとに分類器を適用する。Profitable Model の精度は 1 段階モデルで確定させ、2 段階モデルで収益/損失の改善を目指す。本稿はキャッシングローンの初期与信モデルに Profitable Model を適用した事例を報告する。その結果、従来の Default Model より収益性が優れていることが確認された。

キーワード：データマイニング, Profit Model, Profitable Model, Default Model

1. Default Model と Profit Model について

日本のコンシューマ・クレジット会社は、与信審査を短時間にかつ客観的に行うために 1997 年頃からデータマイニングを用いた与信モデルが開発している。与信モデルは『性別・年齢・職業・地位・家族状況などのデモグラフィック属性』と『自社の利用状況』と『個人信用情報機関のデータ』に基づいて個人の信用度合を客観的に算出する。日本の与信モデルは、顧客のデフォルト/正常を判別する Default Model (デフォルトモデル) である。それに対して顧客の収益/損失¹を判別する Profit Model (利益モデル) がある。顧客がデフォルトすれば損失が発生するため、両者は強い相関がある。

Profit Model は Default Model より戦略的に優れているが、その実現性やコンプライアンス上に問題がある。日本の金融業では個人行動を捉えるリアルデータが存在しないために、安定した Profit Model (利益モデル) の開発は難しいと言われている²。例えば家の購入・結婚・出産・転職・昇給時には多額の個人資金が動き、大きな収益の機会をもたらすが、予兆となるデータは金融データベースに存在しない。外部の情報販売会社からこれら情報を購入すれば収益よりもコストがかかり、何よりもプライバシーにかかわるため個人情報保護法に触れる可能性がある。また Profit Model はデフォルトリスクの高い顧客に一定期間の貸し出す戦略を採用できるが、コンプライアンス上に問題が生じる。さらに専門家は『デフォルト直前の顧客は、キャッシングローンやショッピングを限度額まで使うため収益性が高い。もし個人属性の変化を捉えるリアルデータがない状態で Profit Model を構築すれば、優良顧客の中にデフォルトリスクの高い顧客が紛れ込む可能性がある』と指摘する。

これらの理由により日本の与信モデルは利益を直接予測するのではなく、利益と相関関係が高いデフォ

¹ 利益は収益から損失 (コストとデフォルト額) を引いたものである。

² 詳細は佐々木(2002)を参照。

ルト顧客を予測し、事後的に顧客セグメントの利益を計算する構築手順を踏む。与信モデルは高い精度が必要であるため、専門家は加工変数の生成・分類器の改良・ハイブリッドモデルの導入により5%以上の向上を試みる。しかしそれらを用いても与信モデルの精度の向上には限界がある。本稿のProfitable Model (利益指向モデル) はDefault Modelのフレームワークで使わない顧客の収益/損失データを利用し、利益の改善を目指す。Profitable Modelは、従来のDefault ModelにProfit Modelの要素を取り入れた与信モデルである。なおProfitable Modelの名称はProfit Modelほど利益を目標にしていないが、Default Modelより利益を得られる可能性があるという意味で名づけた。

2. Profitable Model (利益指向モデル) の提案

Profitable Modelはカスケードモデリング³の一種である。カスケードモデリングは複数の分類器⁴を組み合わせて精度を高める構築法の総称である。ハイブリッドモデル⁵はカスケードの具体的な形状や分類器の組み合わせを具体的に実装している。ハイブリッドモデルでは多段階の分類器において同じ目標変数が使われ、1段階の分類器で計算された確信度を2段階の分類器の説明変数に組み込むことで精度を高める。なお確信度はデフォルト率(確率)と違う。デフォルト率は確信度、実績デフォルト率、ブラック/ホワイトのサンプリング数から計算する。

それに対してProfitable Modelは相関のある目標変数を分類器ごとに別々に定義し、2方向でデータマイニングを行なう。特徴はモデルの精度をわずかに犠牲にして、従来のDefault Modelよりも高い利益を目指す点にある。Profitable Modelの1段階の目標変数はデフォルト/正常顧客の2値フラグであり、2段階は収益/損失の2値フラグである。与信モデルの精度は1段階のモデルで確定され、2段階のモデルは精度に関係せず、利益性を高めるために使われる。このモデルのミソは、たとえ運用時に利益予想が機能しなくとも1段階のモデル精度が保障されているので、最小限の利益は確保される点である。Profitable Modelはカスケードモデリングの延長にすぎないが、顧客利益を取り入れた点において金融業の専門家に受け入れやすいモデルである。なお2方向の2段階データマイニングが実務分析で用いられることは初めてである。

3. Profitable Model の理論

3.1. 1段階モデルの分類器

分類器にはカテゴリーを明示できるものと確信度のみでカテゴリーを表示できないものがある。本稿の分類器は決定木を採用するが、後者のMBRやニューラルネットワークの場合は確信度順に並べた顧客を等分割してグループを作成する。

決定木は分割基準値が最小となる説明属性を発見し、高い反応属性順の顧客セグメントの判別ツリーを作成する逐次型の分類アルゴリズムである。決定木は分割(分岐)するたびに各属性の分割基準値の差が小さくなり、最終的に分割できなくなる。デフォルト/正常と収益/損失の2値フラグは相関が高いので、1段階で多く分岐すると2段階の分類器の精度が劣化する。そのためProfitable Modelの1段階の決定木はモデル精度が定常になる前に分岐を止める。その点においてProfitable Modelは従来のDefault Modelより精度が劣る。

本稿の分類器は分割基準値がGini値である決定木CARTを用いる。Gini値の定義を以下に述べる。データ集合Sに、j個のカテゴリー値をもつ目標属性が存在し、集合S内にi個番階の値をもつデータがそれぞれ $X_i(S)$ 個 ($i=1, \dots, j$) があると仮定する。ルールRで S_1 と S_2 に2分割し、部分集合 S_1 内のi番階の値の分布比率を $P_i(S_i)=X_i(S_i)/|S_i|$ とすると、Gini値は次式で求まる。

$$\begin{aligned} Gini(R) &= Gini(x(S_i)) \\ &= (1 - \sum_{i=1}^k p_i(S)^2) - \frac{|S_1|}{|S|} (1 - \sum_{i=1}^k p_i(S_1)^2) - \frac{|S_2|}{|S|} (1 - \sum_{i=1}^k p_i(S_2)^2) \end{aligned}$$

SAS/EMの決定木はCARTのアルゴリズムに準拠していないが、Gini値を用いた決定木を本稿はCARTと呼ぶ。

³ 英語の“Cascade”は階段上に連続分岐する滝を意味する。カスケードモデリングの名称は滝の形状に由来する。

⁴ 人工知能ではデータマイニングの分析手法を分類器という。

⁵ 詳細は小野(2001)を参照。

3.2. 2段階モデルの分類器

1段階のモデルでグループに分けられ、それぞれに2段階のモデルを適用する。1グループに属するデータは分割情報量が少ないため、単独の分類器では精度が低く、わずかな判別力しか得られない。そこで本稿は10個の決定木のアンサンブル学習を用いて精度を高める。

アンサンブル(Ensemble)学習は、あまり精度が高くない分類器の仮説集合(アンサンブル)に投票権を与えて、投票原理に基づいて判定する。アンサンブル学習には、Bagging(バギング)、Boosting(ブースティング)という2種類の代表的なモデルが存在する。Baggingは複数のランダム・サンプリングデータに同じ分類器を適用し、推定値の等ウェイト合計を求める。それに対してBoostingは判定不能データが多く出現するサンプリングを行い、代表推定値はウェイト付き合計で求める⁶。本稿は移植性が高いBaggingを採用する。SAS/EMはBaggingとBoostingの学習法を有する。

2段階のモデルで算出される利益性の確信度は、1段階のデフォルトの確信度と絶対基準で比較できない。そこで利益の確信度は同じデフォルト率を有する顧客の予想利益の相対基準(順位付け)に利用する。実務運用ではさらに同一グループ内の利益の偏りを利益に結びつかせる戦略(ストラテジー)を適用する。例えば利用限度額ストラテジーは、デフォルトリスクに応じて利用限度額の増減を行う戦略である。Profitable Modelにこの戦略を適用すれば、同じデフォルトリスクでも利益予想が大きい顧客の利用限度額を引き上げることができる。

3.3. モデルの比較

Profitable Modelの選択は、1段階のモデルがリフトチャートを用いて精度を比較し、2段階のモデルが累積利益チャートを用いて利益金額を比較する。重要なことは利益が大きくても1段階の精度が良くなければ再現性が乏しいことである。そのため選択モデルは精度の高い1段階のモデルの中で、累積利益が高い2段階のモデルを選択する。

モデルの精度の比較は、正反応割合チャートと正反応補足割合チャートという2種類のリフトチャートを用いる。リフトチャートはデフォルトリスクが大きい順に並べ、デフォルト顧客の的中確率を表す。正反応補足割合チャートの縦軸は累積のデフォルト顧客の中率、横軸は顧客総数(累積セグメント)に対する割合である。モデルの精度は正反応補足割合が大きいほど、あるいは正反応割合が小さいほど優れている。

累積利益チャートは以下のように作成する。デフォルトリスクの確信度が低い順に、同じデフォルトリスクならば利益が多い順に並べる。次に全顧客を20等分し、それぞれのカテゴリーごとに収益/損失を合計する。累積利益チャートはデフォルトリスクが低い順にカテゴリーの収益/損失を累積することで得られる。曲線はデフォルトリスクが高くなるにつれて損失が増加するため、凸型曲線になる。累積曲線の頂点がモデルの最大利益となる。顧客の収益/損失額はBlackとWhiteのサンプリングに応じた調整係数を乗じて現実の数字に近づけたものを用いる。

4. 分析結果

4.1. 分析データ

対象者は信販会社のキャッシングローンを利用した顧客である。サンプリング数はデフォルト顧客1726人(35.36%)、正常顧客3155人(64.64%)、合計4881人である。デフォルト顧客の定義は2年間に3ヵ月以上の延滞が発生した顧客であり、正常顧客は延滞なしである。

データ項目は個人信用情報機関の情報(借入件数、借入残高、照会件数、契約金額等)、個人属性(職業、業種、勤続年数、年齢、居住年数、年収等)、その他の計150項目を用いた。実際にProfitable Modelで使われる項目は20~40項目程度である。実務のモデル開発では精度を高めるために多数の加工データを用いるが説明を要するため、本稿は最小限の加工データしか採用しない。

4.2. 1段階の決定木の分析結果

一般にコンシューマー・クレジット業の与信は、個人信用情報機関のデータが最も有効であると言われる。その事実が図1の判別ツリーから読み取れる。本稿のDefault ModelとProfit Modelの判別ツリーの上層部はすべて図1の左図と同じ分岐を有する。分岐条件は第1層が「個人信用情報機関の借入全

⁶ 詳細はフロイド、シャピロ(1999)を参照。

件数]であり、次層が「他の個人信用情報機関に登録があるかどうか」、「個人信用情報機関の合計残高」、「クレジット等の登録が2枚以上あるかどうか」のいずれかである。

図1の右図はモデルの精度が21枚(グループ)で定常になることを示す。2本の線は学習データと訓練データのモデルの精度であり、両者が一致するほど良い。SAS/EMは精度が定常になる21枚を自動探索できる。

図2はモデル精度の観点からDefault Modelの葉数の影響を比較する。葉数は最終分岐のグループ数である。右図の正反応補足割合チャートのモデル曲線は座標点(横軸50%, 縦軸65%)を通る。それは全顧客50%を選択した場合に全デフォルト顧客65%を的中するモデルを意味する。図2は共にSAS/EMが探索した最適枚数21枚のモデルが、8枚よりわずかに精度が高いことを示す。

図3は判別ツリーの構造を示す。左上図の2個はDefault Modelの決定木の最終葉数8枚と21枚である。判別ツリーの構造は、中心円から外円に向かって分割されていくことで表現する。色が濃いほどBlackとWhiteの分離がうまくいき、区分面積が人数を表わし、最終外円の区分が判別ツリーの葉数を表わす。内部の区分構造が複雑になるにつれて、判別ツリーの構造は複雑になる。一般に決定木の短所の一つは巨大グループを作成することである。実務のモデル開発では細かい施策を適用させやすくするために、業務知識を反映させながら葉数30~50枚になるまで強制分割させる。

本稿のProfitable Modelの1段階のモデルは最終葉数10枚を採用し、10グループごとに次の分類器を適用する。葉数10枚で分岐を止めた理由は、これ以上分岐するとグループ内のデータ数が少なくなるためである。

4.3. Profit Model と Default Model の比較

Profit Modelは目標変数の収益額が問題になる。少額ローンの顧客は収益からコストを引くとマイナスの利益となるため、そのままではブラックと判定されてしまう。カード入会から2年間のコストは、カード作成費用を含めて約3000円以上になる。この事実に基づいて複数の収益の境界値を設定しモデルを比較する。

下表1は境界値とデフォルト/正常顧客の人数割合を記載する。境界値が1円、3500円では、デフォルト/正常顧客の人数割合はそれほど変化しない。

図3はProfit Modelの収益/損失の境界値が1円、3500円、6000円、8000円、10000円である判別ツリーが含まれる。Profit Modelは境界値が変動しても大きく判別ツリーの構造が変化しない。注目点は図3のProfit Modelの中に円の分割面積の約45%を占める巨大なグループ存在することである。この部分は「他社借入が0また1件で、かつクレジットカードを2枚以上保有する顧客層」に相応する。一方、Default Model葉数10枚以上では、この顧客層を分割できる。つまりProfit ModelはDefault Modelと同じ精度、同じ葉数を有しても判別ツリーの構造が相違する。一般に密に分割された部分はオーバーフィッティングの可能性が高く、しかも経済変化等に弱い可能性がある。

図4はProfit ModelとDefault Modelの精度の比較を示す。両者の差異は右図の正反応補足割合チャートからわかりづらいが、Default Modelが最大5%程度優れている。左図の正反応割合チャートと下表の誤差集計表から、モデル精度が優れている順序はDefault Model最終葉数21枚、Profit境界値1円、Default最終葉数8枚、Profit境界値3500円、6000円、8000円、10000円である。つまりDefault ModelはProfit Modelより精度が優れており、Profit Modelの境界値が大きくなるほど予測が難しくなる。

図5は収益面からモデルを比較したものである。左図はDefault Modelの葉数21枚が8枚より収益が大きいことを示す。中図はProfit Modelの境界値による収益の影響が少ないことを示す。右図はリスクが少ない場合はDefault Modelの収益が上回り、リスクが大きい場合はProfit Modelの収益が大きくなる。つまりProfit Modelは高リスク顧客の収益を予想している可能性がある。本稿はサンプリング条件が複雑なため、誤解が持たれないように累積利益チャートの金額単位を付けていない。

表1 境界値とデフォルト/正常顧客の人数割合

収益の境界		境界値1円		境界値3500円		境界値6000円		境界値8000円		境界値10000円	
		未満	以上	未満	以上	未満	以上	未満	以上	未満	以上
不良	35.36%	34.58%	0.80%	34.85%	0.51%	34.99%	0.37%	35.10%	0.27%	35.18%	0.18%
正常	64.64%	1.45%	63.18%	3.87%	62.14%	3.46%	61.18%	4.16%	60.48%	4.92%	59.72%

4.4. 2 段階の分類器の分析結果

図 6 は 1 段階のグループの一つに 2 段階の分類器を適用したときのモデル精度の結果である。上から決定木 CART の Bagging, 決定木 CART の Boosting, 決定木 CART 単体の順である。CART 単体では精度の向上が見込まないので決定木 CART の Bagging を採用する。またアンサンブル学習を試みても精度が上がらないグループも存在した。

図 7 の累積利益チャートは Default Model と Profitable Model と Profit Model の利益比較である。リスクが低い場合は Profitable Model の利益が大きく、リスクが大きい場合は Default Model, Profitable Model, Profit Model の順位に利益が大きくなる。以上から Profitable Model は従来の Default Model よりも利益性が高いと言える。ただモデル再現性は他会社データの検証が必要になる。

5. 考察

本稿は Profitable Model, Profit Model, Default Model をモデルの構造・精度・利益から比較した。

① Profit Model と Default Model の比較

Profit Model は Default Model と比べてモデルの精度は低いが、リスクが高いときは収益が高い。Profit Model のモデル構造には全体の 45%弱におよぶ「他社借入が 0 また 1 件で、かつクレジットカードを 2 枚以上保有する顧客層」が存在する。Default Model はこの顧客層を分割できるため、Profit Model よりもモデルの安定性があると推測される。Profit Model は Default Model とほぼ同じ精度かつ最終枚数が同じでも経済変化に伴う精度の劣化が予想される。

② Profitable Model

Profitable Model は Default Model と比べて、モデルの精度はわずかに低いが、収益性は優れている。しかし実務運用へ移行するには経済変化によるモデルの精度・利益の影響を検証しなければならない。本稿では過去データが不足しているため時系列変化が検証できず、今後の課題である。また Profitable Model が必ず Default Model より優れているわけでない。1 段階のモデルでの最適分割数は試行錯誤で求めるが、それが Profitable Model の収益性に大きく影響する。

Profitable Model はアンサンブル学習を用いるため、キャッシングローンモデルを運用するホストコンピュータへの移植が簡単でなく、実用化には運用面の課題が残る。

6. おわりに

Profitable Model を開発するために、Profit Model と Default Model の比較を試みたが、この研究も価値のある結果が得られた。今まで Profit Model は使ってはならないと言われていたが、本稿が初めて Profit Model の欠点を定量的に追求した。Profit Model が分割不能の顧客層をさらに分割するには、顧客のリアルデータが必要であろう。

この事実をキャッシングローンの業務担当者に告げたところ、次の発言が得られ業務上の裏づけがとれた。『他社借入が 0 また 1 件で、かつクレジットカードを 2 枚以上保有する顧客層は、キャッシングローン顧客の中でリスクが小さい安全顧客に相当する。安全顧客のデフォルトは勤務先、勤続年数、住居形態等からもう少し深く推測できる。しかし安全顧客の収益は個人信用情報を参考にしても予想が難しい。逆にリスクの高い顧客の収益は予想できるかもしれない』と。

顧客の予想利益モデルの開発は研究者の夢であるが、顧客属性のリアルデータの取得が困難な状況下で、前述のモデル開発は時期尚早であろう。限られた顧客属性データで、与信モデルの精度や収益性を向上させるには、顧客の収益/損失データか、顧客のリアル取引データの利用が考えられる。本稿は今後の与信モデルの改良に新しいアイデアを提供したと位置づけられる。

本稿は個人的見解で書かれており、所属する UFJ 銀行の意見をあらわすものではありません。

7. 参考文献

- 佐々木研, “リスク管理とそれに必要な要素”, 第 21 回日本 SAS ユーザ会研究発表論文集, pp349-353, SAS Institute Japan, 2002.
- コアブ・フロインド, ロバート・シャピリ, 訳: 阿倍直樹, “ブースティング入門”, 人工知能学会, vol.14 No.5, pp771-780, 1999.
- 小野潔, “データマイニングを利用した融資モデルの現状と課題”, 人工知能学会研究会資料 SIG-J-A004, pp49-54, 2001.
- 小野潔, “ハイブリッド・コンポーネントの構築”, 第 20 回日本 SAS ユーザ会研究発表論文集, pp269-327, SAS Institute Japan, 2001.

図1 Default Model の判別ツリーと精度

左図：判別ツリー 右図：モデル精度比較(上から学習データ, 訓練データ)

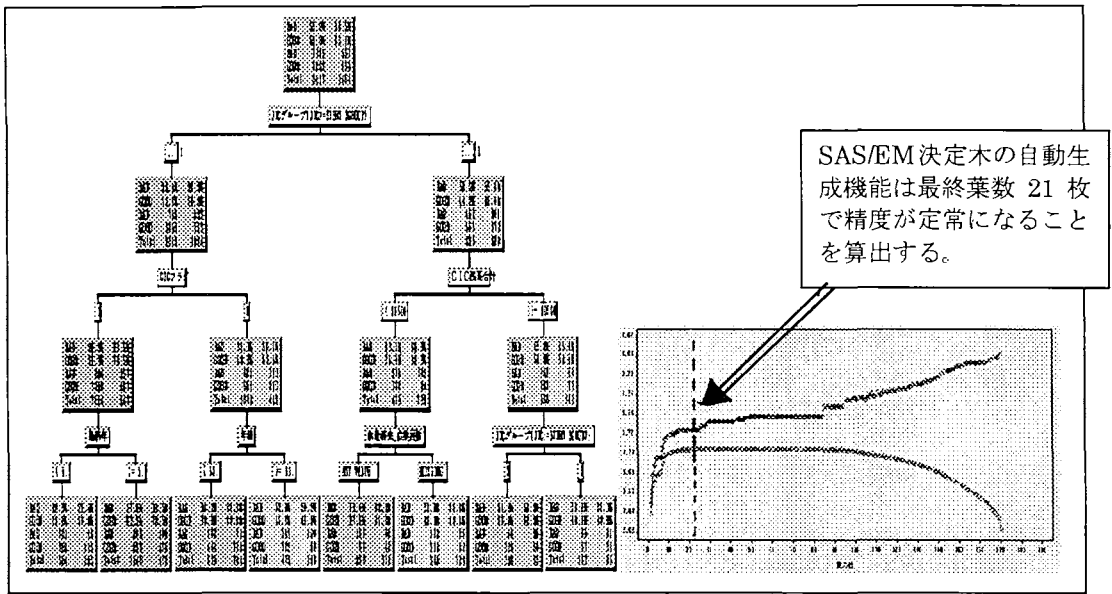


図2 Default Model の葉数の影響

左図：正反応割合チャート (上から, Default 葉数 21 枚, 8 枚, ランダム曲線)

右図：正反応補足割合チャート (上から, 理想曲線, Default 21 枚, 8 枚, ランダム曲線)

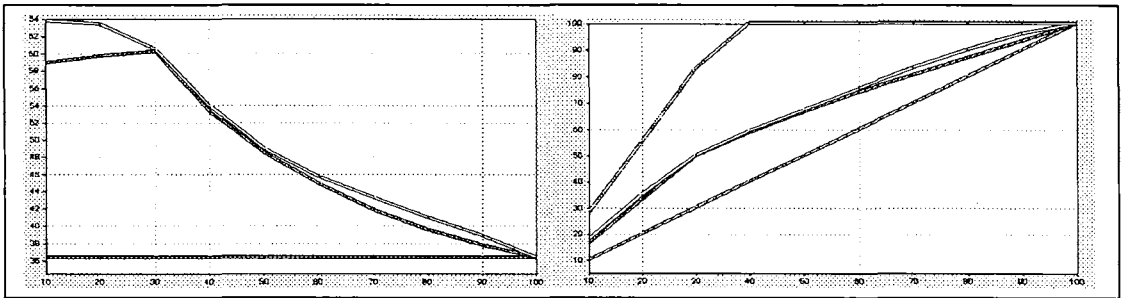


図3 判別ツリー (左上から Default 葉数 8 枚, Default 葉数 21 枚, Profit 境界値 1 円, Profit 境界 3500 円, Profit 境界 6000 円, Profit 境界 8000 円, Profit 境界 10000 円)

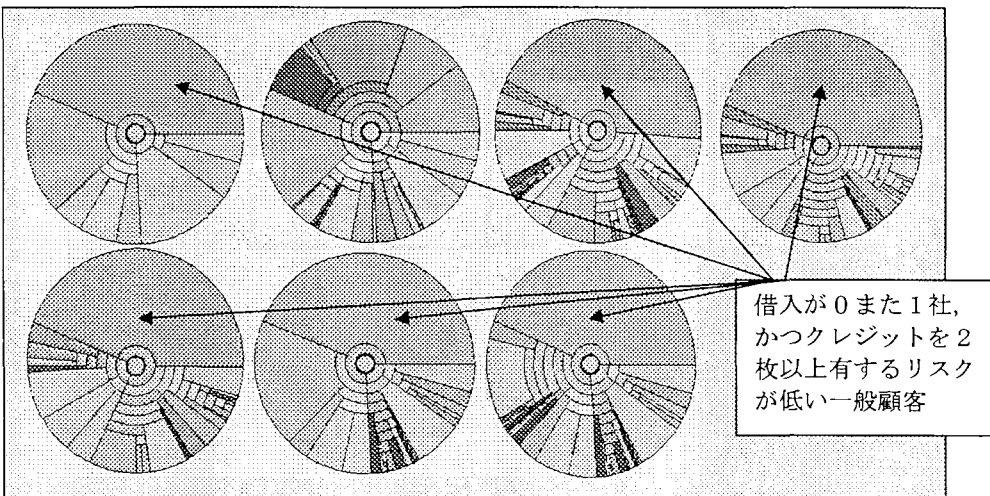


図4 Profit Model の境界値変化と Default Model の精度比較

上左図：正反応補足割合チャート

上右図：正反応割合チャート

(上から Profit10000 円, 8000 円, 6000 円, 3500 円, 1 円, Default21 枚, 8 枚)

下表：誤差集計表 (上から Default8 枚, 21 枚,

Profit Model 境界値 1 円, 3500 円, 6000 円, 8000 円, 10000 円)

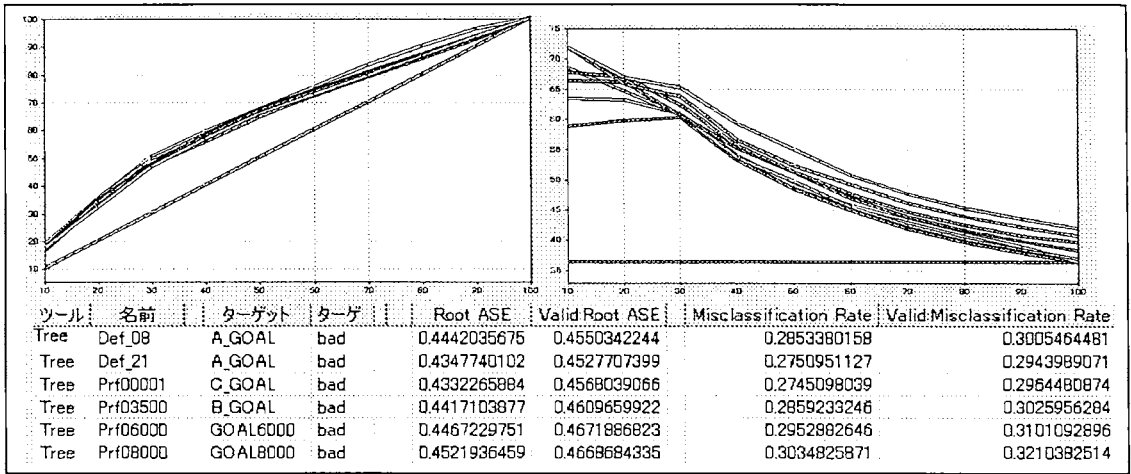


図5 累積利益チャート (横軸：リスクランク(右にいくほどリスク大), 縦軸：累積利益)

左図：上から Default Model 21 枚, 8 枚

中図：Profit Model 境界 1 円, 3500 円, 6000 円, 8000 円, 10000 円は共にほぼ同じ曲線

右図：リスクが小のときは Default 121 枚が上, リスクが大のときは Profit 3500 円が上

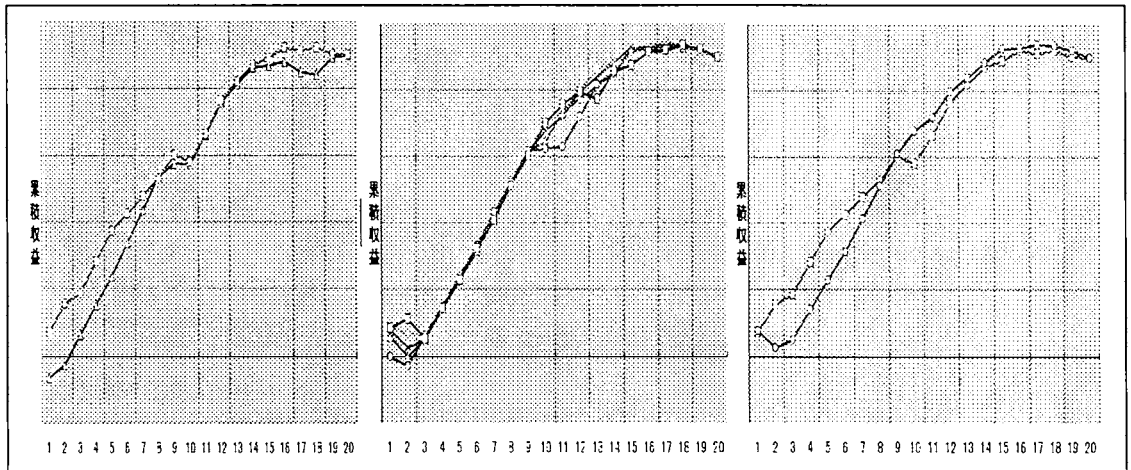


図6 2段階モデルの正反応補足割合チャート

(上から, 決定木の Bagging, 決定木の Boosting, 決定木単体)

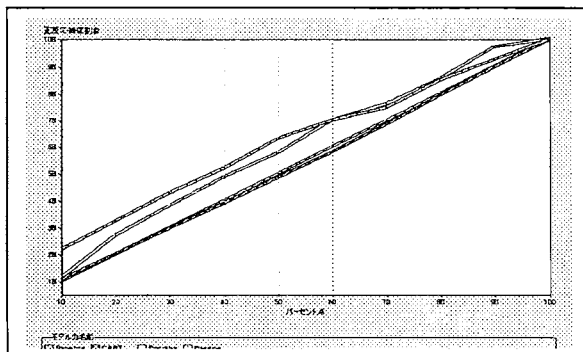
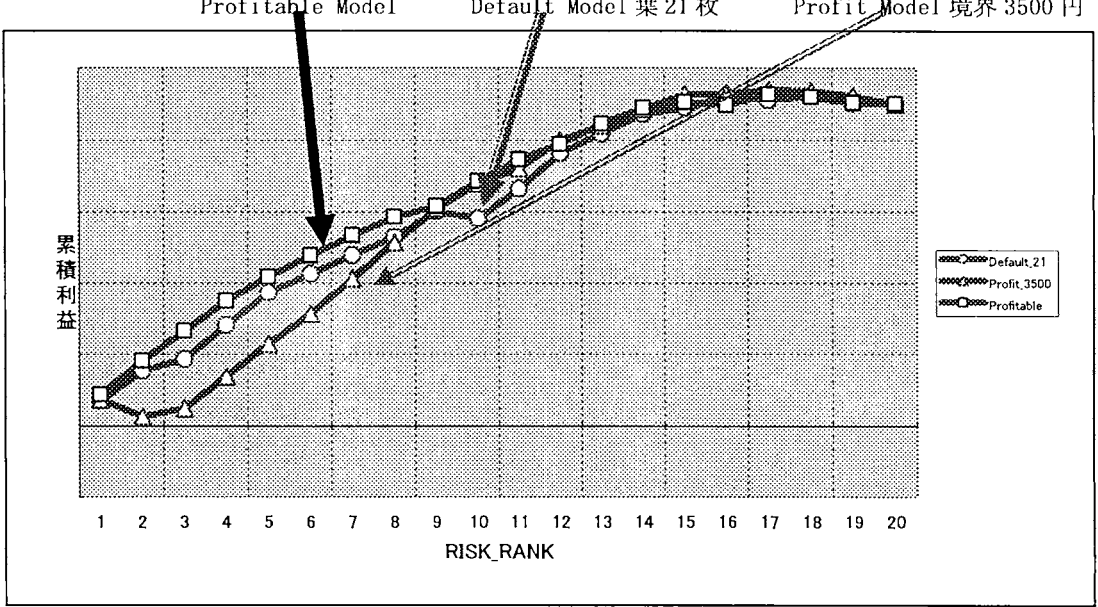


図7 Profitable Modelの累積利益チャート



日本SASユーザー会 (SUGI-J)

非補償型ロジットモデルを用いた企業倒産確率の予測モデル

～NLP Procedure による非補償型ロジットモデルに対するパラメータ推定～

坂巻 英一

株式会社金融工学研究所

東京工業大学大学院

社会理工学研究科価値システム専攻博士後期過程1年

Risk Scoring Model with Non-Compensatory Logit Model

～Parameter Estimation for Non-Compensatory Logit Model by NLP Procedure～

Sakamaki, Yoshikazu

Financial Technology Research Institute Inc.

Tokyo Institute of Technology

Graduate School of Decision Science and Technology

要 旨

現在、我が国の金融機関における最大の課題である金融システムの安定化を実現するために、信用リスクの測定と管理の効率化・精度の向上が求められている。

信用リスク管理は今日統計モデルを活用したアプローチが主流になっているが中でも二項ロジットモデルを用いたものは現在多くの金融機関で利用されている。本論では、従来用いられてきた線形補償型二項ロジットモデルを企業のデフォルト予測に用いる際の問題点を指摘すると共に、従来型モデルの持つ問題点を改善することを目的とし、非補償型ロジットモデルを用いた企業デフォルト確率の予測モデルを提案する。併せて、SAS NLP プロシージャを用いた最尤推定を行う方法を紹介すると共に、非補償型ロジットモデルを信用リスクの分野へ応用する際の今後の課題と展望について考察する。

キーワード： 信用リスクモデル・非補償型ロジットモデル・NLP プロシージャ

1. はじめに

経済が低迷を続ける中、わが国における金融システムと金融行政に対する信頼を回復し、世界から評価される金融市場を作ること、今日の金融市場に課された大きな課題である。こうした中金融行政は、平成16年度に主要行の不良債権比率を現状の半分程度に低下させ、問題の正常化を図るとともに、構造改革を支えるより強固な金融システムの構築を目指すとしており、現在全ての金融機関・投資家個人・事業会社に対し、信用リスクの存在を十分に認識しそれを測定・管理することが求められている。

信用リスク分析の基本として現在でも広く行われている手法の一つに財務諸表分析がある。財務諸表分析は19世紀末に金融機関が融資対象企業における信用調査を目的として開発された手法である。しかしながら、財務諸表に基づく分析はアナリストの分析経験の深さに大きく左右されることが多く、その結果、分析結果にばらつきが生じることが多い。このような主観的な分析手法のみではリスク測定に関するコンシステンシーを保つことが難しい上、大量の信用分析を精度良く行う上で費用も時間も掛かるといった問題点が指摘されていた。

こうした問題点に対し、数理統計的手法取り分け多変量解析的手法を駆使したモデルを用いることで財務諸表分析を科学的な方法論によって導出しようという試みが W.H.Beaver(1967)以降盛んに行われるようになってきた。中でも最も有名なものが、E.I.Altman(1968)(1979)(1971)(1976)の発表したZ値モデル或いはZスコアモデルと呼ばれる手法である。

本論では Altman(1968)以降、これまで先行研究として行われてきた信用リスク管理における統

計的アプローチの流れを概観すると共に、現在広く用いられている線形補償型二項ロジットモデルを基礎とした確率モデルによる企業デフォルト確率の推定方法が持つ問題点を指摘すると共に、これらの問題点を解決しデフォルト確率の推定精度を向上させるための非補償型二項ロジットモデルを基礎としたスコアリングモデルの改善提案を行うことを試みる。

2. 先行研究における信用リスク分析モデルの紹介

ここでは統計的アプローチに基づく信用リスク測定手法についてデフォルト企業の予測を例に先行研究としてこれまで行われてきた研究を紹介する。前述した通り、企業のデフォルト傾向に関する研究は、Altman(1968)の判別分析モデルに遡ることができ、その研究成果はZ値モデルとして知られ、これまで企業のデフォルト傾向分析において最も広く使用されてきた手法の一つであると言える。

ここで、判別分析を用いた信用リスク測定方法について簡単に説明する。属性を示すリスクファクターが個社*i*毎にそれぞれ存在するとし、これらのリスクファクターに対してウエイトを課し、加重平均することでデフォルトに対するリスクファクター ($Z_i^{(1)}$)、非デフォルトに対するリスクファクター ($Z_i^{(2)}$) を合成する。

$$\begin{aligned} z_i^{(1)} &= \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi} \\ z_i^{(2)} &= \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi} \end{aligned} \quad (1)$$

ここで合成されたリスクファクターのそれぞれのグループにおける平均を

$$\bar{z}^{(1)}, \bar{z}^{(2)} \quad (2)$$

とする。また、合成されたリスクファクターのサンプル全体における分散を $Var(Z)$ とする。さらに、未知のパラメータであるウエイトは、下式を最大化することによって決定される。

$$\eta^2 = \frac{(\bar{z}^{(1)} - \bar{z}^{(2)})^2}{Var(z)} \quad (3)$$

このようにして得られたパラメータをウエイトとし、複数のリスクファクターの加重平均の値の水準によって、任意の与信がデフォルト・非デフォルトのどちらに属するかが決定されることになる。

ここで紹介した線形判別関数に基づく信用リスク分析は計算が容易であることと伝統的ないわゆるスコアカード方式に代表される信用評点分析に似た点が多いため、扱いやすく現在広く一般的に利用されている。しかしながら、判別分析を信用分析に適用する場合には、その統計的仮定、信用分析に特有の問題点など以下の点に十分注意を要する必要がある。(森平;1999)

《1》 正規性と等分散性

判別分析を実施するにあたり分析に用いられる変数の正規性及び等分散性の仮定が満たされない場合、判別係数の統計的検定や判別評点によるデフォルト予測が正しく行われなくなる可能性がある。仮に、正規性と等分散性の仮定が満たされない場合には、誤差項の確率分布に関する仮定のみを要求する定性的従属変数モデルや利用するリスクファクターに関する仮定を全く必要としないツリー分析等のノンパラメトリック手法が適していると考えられる。

《2》 母集団におけるデフォルト・非デフォルト確率の均一性

判別分析では通常、デフォルト・非デフォルト企業を判別する場合、デフォルト企業と非デフォルト企業の割合に関して何ら事前情報がなく、デフォルト企業と非デフォルト企業の割合は等しいという仮定の基でモデルの構築が行われていた。しかしながら、通常、デフォルト企業と非デフォルト企業との割合は、地域毎・銀行毎等によって異なるのが普通でありこれらの仮定は現実と反することになる。

《3》 算出された重み係数の持つ意味

推定された判別係数は回帰分析における推定回帰係数と異なり、その絶対値の大きさは

ユニークに決定されるのではなく、その相対的な比率のみがユニークに決定される。従って、判別係数から推定された判別評点もその絶対値はユニークに決定されない。判別評点の値そのものの絶対値についても同様である。つまり、判別点からの各企業の判別評点の偏差のみが意味を持つことになる。

特に判別評点（Zスコア）を用いて企業のデフォルト・非デフォルトを分類するためには、リスクファクター（独立変数）が多変量正規分布に従うこと、デフォルト企業と非デフォルト企業の独立変数の分散・共分散行列が等しいという二つの強い仮定が必要不可欠であり、分析手法が広く一般的に利用されている半面、統計的に扱いにくいといった問題点も含んでいる。

また、リスクを数値化するためにはいつどのくらいの確率で企業がデフォルトするかを明らかにする必要がある。即ち、Z値スコアに代表される判別分析的アプローチでは比較的良好・比較悪いといった相対評価としての尺度を与えることが出来たととしても、絶対評価としてどのくらいの確率で企業がデフォルトするかを把握することは出来ない。

こうした問題点を解決する手法としてやがて定性的従属変数モデルが代用されるようになってきた。定性的従属変数モデルの代表的なものとして、線形回帰分析モデルが上げられる。即ち、

- y_i : 0（正常）或いは 1（デフォルト）を取る確率変数
- x_{ij} : 個社 i の持つ j 番目の信用リスクファクター
- β_j : j 番目のリスクファクターに対する推定パラメータ
- ε_i : 個社 i の誤差項

とした時、

$$y_i = \beta_0 + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i \quad (4)$$

となりこの回帰モデルから得られる従属変数の期待値がデフォルト確率の推定値を表すことになる。しかしながら、線形回帰モデルを用いた手法では、リスクファクターの合計値が大きいところでは 1 より大きくなり、一方、リスクファクターの合計値が小さいところでは 0 より小さくなることから、確率の定義に反することになる。この問題を解決するために、従属変数の期待値が 0 より小さい時にはデフォルト確率を 0、1 より大きい時にはデフォルト確率を 1 とするといった推定方法が取られることがあるが、推定デフォルト確率を 0 と 1 の間に恣意的に押し込めることは自然ではないと考えられる。

こうした問題点を克服するための手法として、ロジスティック回帰分析が用いられるようになってきた(Martin;1977)。

ロジスティック回帰分析では、リスクファクターの合計値を

$$Z_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} \quad (5)$$

と定義した場合、個社 i がデフォルトする確率 p_i を

$$p_i = \frac{1}{1 + \exp(-Z_i)} \quad (6)$$

により与えるというものである。

(6)式により企業 i のデフォルトが起こる確率 p_i が定義された場合、係数ベクトル β に対する尤度関数は $L_i = L(\beta | x_{i1}, x_{i2}, \dots, x_{im})$ であることから

$$\begin{aligned} L(\beta) &= \prod_{i=1}^I L_i \\ &= \prod_{i=1}^I p_i^{y_i} (1 - p_i)^{1 - y_i} \end{aligned} \quad (7)$$

で与えられる。
ここで、

$y_i=1$: 個社 i がデフォルトした時
 $y_i=0$: 個社 i がデフォルトしなかった時

である。また、(7)式の両辺に対し対数を取ることで対数尤度関数を

$$l(\beta) = \sum_{i=1}^n \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\} \quad (8)$$

と定義できる。一般的に、パラメータの推定には通常最尤推定法が用いられ、(8)式が最大になるような係数ベクトル β を決定する。

3. 従来型モデルの持つ問題点

ロジスティック回帰による企業デフォルト確率の推定は Altman(1968)の古典的信用リスクモデルに比べデフォルト確率を絶対尺度により測定することを可能にしたという点において意義のあるものであると言える。しかし、このモデルはいわゆる線形補償型二項ロジットモデルを基礎としており、モデルの中で使用されているある特定のリスクファクターだけがたまたま大きなデフォルト傾向を示したことにより、実際にはその企業はそれ程デフォルト傾向が強くないにも関わらず、高い推定デフォルト確率を与えられてしまう可能性を含んでいる。また、経験則からデフォルト直前の企業はほぼ全ての財務指標について平均値に比べてデフォルト傾向に傾いていると考えられる。表.1 は日経新聞社が提供する BULK システムに搭載された財務データ公開企業における、デフォルト企業と正常企業それぞれについて代表的な 7 つの財務指標の平均値を算出した結果であるが、この表からもデフォルト企業の全ての財務指標は正常企業のそれと比べ企業経営を悪化させる方向に傾いていることは明らかである。こうした事実を鑑みた場合、モデルの中で使用されているリスクファクターが総合的に強いデフォルト傾向を示した場合に推定デフォルト確率が上昇するとした、従来型の線形補償型二項ロジットモデルではデフォルト企業の捕捉に限界があるのではないかと考えられる。

表.1 正常企業・デフォルト企業それぞれにおける財務指標の基礎統計量
MEANS プロシジャ

DEFAULT=0 (正常企業)						
変数	ラベル	N	平均値	標準偏差	最小値	最大値
Xa	自己資本比率	2830	0.3893053	0.2118216	-2.0314880	0.9667590
Xb	借入金依存度	2830	0.3520748	0.2466985	0.0021947	2.5595341
Xc	売上高支払利息・割引料	2830	0.0096626	0.0177726	0	0.4604436
Xd	当座比率	2830	0.8918016	0.5655448	0.000317524	4.9957143
Xe	預貸率	2830	0.5905650	0.8625469	-0.0214242	4.9326155
Xf	売上高営業利益率	2830	0.0417913	0.0953923	-2.5071885	0.8874064
Xg	支払準備率	2830	0.2985706	0.3504356	-0.0069265	4.4315789

DEFAULT=1 (デフォルト企業)						
変数	ラベル	N	平均値	標準偏差	最小値	最大値
Xa	自己資本比率	84	-0.0073144	0.4437821	-2.2886805	0.6112871
Xb	借入金依存度	84	0.8092553	0.5584754	0.0423840	4.6631497
Xc	売上高支払利息・割引料	84	0.0286113	0.0267828	0	0.1592179
Xd	当座比率	84	0.4086666	0.3706348	0.0086844	1.8037889
Xe	預貸率	84	0.1039228	0.1418799	-0.0271460	1.0553471
Xf	売上高営業利益率	84	-0.0421947	0.1249830	-0.6001036	0.0882768

変数	ラベル	N	平均値	標準偏差	最小値	最大値
Xg	支払準備率	84	0.0981869	0.0964766	-0.0564024	0.5730717

こうした問題点を解決する上で、片平ら(1998)はマーケティングサイエンスの分野において興味深い研究を行っている。マーケティングサイエンスの世界においては、多くの場合線形補償型二項ロジットモデルにより消費者の意志決定が行われることを仮定しモデル構築が行われる。しかし、どんなに価格が安い商品であっても品質が一定の基準を満たさなければ消費者にその商品が受け入れられないといったように、いわゆる非補償型の意志決定が行われる場面は少なくない。この点に関して、ある条件のもとでは非補償型のモデルを補償型のモデルで近似できることが、Dawes and Corrigan (1974), Johnson and Meyer (1984)等によって示されているものの、そもそも消費者が非補償型の意志決定を行う頻度が高いという指摘がある上(Bettman and Jacoby (1976), Payne and Ragsdale (1978)), 属性間に負の相関がある場合にはそのような非補償型の意志決定を補償型モデルによって近似できないことが示されている。(Johnson, Meyer and Ghose (1989)).

これを企業のデフォルトに当てはめて考えてみた場合、特にデフォルト直前の企業についてみると、殆ど全ての財務指標が企業の経営を悪化させる方向に傾いている。また、従来企業デフォルト確率予測モデルにおいてはモデルの説明変数として企業の財務指標が利用されることが多いが、財務指標は一般に相互に強い負の相関を有していることが多く、Johnson, et al.(1989)の報告を考慮するならば、企業デフォルト確率の予測モデルとして従来行われている線形補償型の二項ロジットモデルを適用することは統計的な観点から適切ではないと考えられる。こうした点を鑑みた場合、従来型の線形補償型二項ロジットモデルよりもむしろ、モデルの中で使用されている全てのリスクファクターが強いデフォルト傾向を示した場合にのみデフォルト確率が上昇するといったいわゆる非補償型二項ロジットモデルを用いた方が現実には則しているのではないかと考えることができる。そこで、本論では従来型の補償型二項ロジットモデルを基礎とした信用リスクモデル(スコアリングモデル)の問題点を解決し、企業デフォルト確率の推定精度を向上させることを目的として、非補償型二項ロジットモデルを用いたモデルの改善提案を行うこととする。

また、本論における提案モデルの妥当性を検証するために、実データをモデルに適用することによりモデルの妥当性を検証するとともに本論における提案モデルである非補償型モデルが従来型の線形補償型二項ロジットモデルに比べ高い推定力を持つことを示す。

3-2 本論における提案モデルの説明

ここで本論における提案モデルとして、片平ら(1998)を基礎とし、企業のデフォルト確率推定モデルとして、以下に示す「連結型」と「分離型」の二つのモデルを示す。

<連結型モデル>

モデルの中で考慮される全ての属性が閾値を越えたときに、企業デフォルトが発生することを仮定したモデルである。ある企業*i*が与えられた時この企業がデフォルトする確率を

$$P_i = \prod_{k=1}^K \frac{1}{1 + \exp(-\beta_k(x_{ik} - \tau_k))} \quad (9)$$

により定式化する。

ただし、

x_k : 企業*i*の*k*番目の属性に対する推定パラメータ
($i=1,2,\dots,I; k=1,2,\dots,K$)

β_k : モデルで使用される*k*番目の属性に対する推定パラメータ

τ_k : *k*番目の属性についての閾値($k=1,2,\dots,K$)

とする。

<分離型モデル>

モデルの中で考慮される何れかの属性が閾値を越えたときに、企業デフォルトが発生することを仮定したモデルである。ある企業*i*が与えられた時この企業がデフォルトする確率を

$$P_i = 1 - \prod_{k=1}^K \left(1 - \frac{1}{1 + \exp(-\beta_k(x_{i,k} - \tau_k))} \right) \quad (10)$$

により定式化する。

ただし、

x_k : 企業 i の k 番目の属性に対する推定パラメータ

($i=1,2,\dots,I; k=1,2,\dots,K$)

β_k : モデルで使用される k 番目の属性に対する推定パラメータ

τ_k : k 番目の属性についての閾値($k=1,2,\dots,K$)

とする。

3-3 比較対象モデル

本論における比較対象モデルとしては、(6)式で示された通常の二項ロジットモデルを用いる。

3-4 パラメータの推定

パラメータの推定は最尤推定法により行い、(11)式を最大にするパラメータ (β) を求める。

$$L(\beta) = \prod_{i=1}^I p_i^{y_i} (1 - p_i)^{1-y_i} \quad (11)$$

3-5 モデルの適合度に関する検証

本論では従来線の線形補償型二項ロジットモデル、非補償型ロジットモデルの適合度指標として、対数尤度、AICの各統計指標に加え、Kolmogorov-Smirnov統計量(Chakravarti et al.(1967), Harter et al.(1984), Khamis (1990)(1992)(2000)) (以下K-S Distanceと記述する)、divergenceの各指標をモデル評価に利用した。ここで、K-S Distanceとは二つの分布をそれぞれスコアに基づき累積してゆき、その百分率をとった場合の最大の差である。また、divergenceとは正常企業に対して付与されたスコアとデフォルト企業に対して付与されたスコアの期待値と分散からそれぞれの分布がどれだけ離れているかを表す指標であり

μ_A : 正常企業に付与されたスコアの期待値

μ_B : デフォルト企業に付与されたスコアの期待値

V_A : 正常企業に付与されたスコアの分散

V_B : デフォルト企業に付与されたスコアの分散

とした時、

$$Divergence = \frac{2(\mu_A - \mu_B)^2}{V_A + V_B} \quad (12)$$

によって計算される。

4. モデルの実データへの適用

ここで、本論における提案モデルの妥当性を検証するために、提案モデルに対し、実データを適用することを試みる。検証用データとしては、日本経済新聞社が提供する BULK システムに掲載されている公開企業情報を使用した。

4-1 検証用データ概要

1. 2000年4月から2001年3月までの各社決算データ
2. データボリューム：モデルで使用される財務データに関して欠損値を含まないもの2,914件
3. デフォルト件数：決算書公開から3年以内にデフォルトした場合をデフォルトと見なしモデルを構築 (デフォルト確率：84/2,914=2.88%)

4. モデル内で使用された財務指標

実務経験に基づき企業のデフォルトに大きく影響すると考えられる

- 自己資本比率 (X_a)
- 借入金依存度 (X_b)
- 売上高支払利息・割引料 (X_c)
- 当座比率 (X_d)
- 預貸率 (X_e)
- 売上高営業利益率 (X_f)
- 支払準備率 (X_g)

を財務指標としてモデル内で使用した。

ただし、企業の財務データに基づく各財務指標の算出過程は以下の通りである。

自己資本比率	=	資本金合計/資産合計
借入金依存度	=	(長短期借入金+割引手形+CP) / (資産合計+割引手形)
売上高支払利息・割引料	=	(支払利息・割引料)/(売上高・営業利益)
当座比率	=	(現預金+受取手形・売掛金+有価証券)/流動負債合計
預貸率	=	(現金・預金)/(長短期借入金+受取手形割引高)
売上高営業利益率	=	(営業利益)/(売上高・営業利益)
支払準備率	=	(現金・預金)/(流動負債合計)

またモデル検証用として使用した検証用データの基礎統計量を以下に示す。

表.2 検証用データの基礎統計量
MEANS プロシジャ

変数	ラベル	N	平均値	標準偏差	最小値	最大値
Xa	自己資本比率	2914	0.3778722	0.2314982	-2.2886805	0.9667590
Xb	借入金依存度	2914	0.3652537	0.2717452	0.0021947	4.6631497
Xc	売上高支払利息・割引料	2914	0.0102088	0.0183644	0	0.4604436
Xd	当座比率	2914	0.8778746	0.5666294	0.000317524	4.9957143
Xe	預貸率	2914	0.5765369	0.8542477	-0.0271460	4.9326155
Xf	売上高営業利益率	2914	0.0393703	0.0973648	-2.5071885	0.8874064
Xg	支払準備率	2914	0.2927942	0.3473522	-0.0564024	4.4315789

4-2 パラメータの推定

このデータを基に本論における連結型・分離型双方の非補償型モデル並びに通常の線形補償型二項ロジットモデルを用いたパラメータ推定を最尤推定法により実施した。

パラメータ推定には SAS の NLP Procedure を使用した。推定に使用したプログラムの一部を参考資料として付録に添付する。ただし、非補償型モデルに関し変数選択法により効率的にパラメータ推計を実施することができる SAS モジュールは現在のところ市販されていないため、本論では全てのモデルに対しパラメータを総当り法により推定している。推定に当たっては、探索結果が局所最適解(Local Minimum)に落ちることのないよう、其々の変数の組合せに対して 20 回ずつ初期値を発生させ、其々の初期値を基にした最尤推定法による探索を行い、AIC 並びに対数尤度を最小にするパラメータの組合せを其々のモデルにおけるパラメータの推定結果として採用した。

表.3 パラメータの推定結果

パラメータ	従来型線形補償型ロジットモデル	p-値	非補償型モデル(連結型)	p-値	非補償型モデル(分離型)	p-値
B_0 定数項	-2.393	0.00001	-	-	-	-
B_a 自己資本比率	-1.660	0.01260	-6.612	0.00006	-	-
μ_a	-	-	0.098	0.00019	-	-
B_b 借入金依存度	1.291	0.02334	6.456	0.01522	3.610	0.00000
μ_b	-	-	0.566	0.00828	2.946	0.00000
B_c 売上高支払利息・割引料	-2.023	0.18282	-	-	-	-
μ_c	-	-	-	-	-	-
B_d 当座比率	-1.257	0.00827	-	-	-6.376	0.00004
μ_d	-	-	-	-	0.799	0.07183
B_e 預貸率	-2.918	0.01545	-	-	-	-
μ_e	-	-	-	-	-	-
B_f 売上高営業利益率	-2.477	0.00027	-32.168	0.14076	-1.540	0.09848
μ_f	-	-	0.029	0.00000	9.619	0.00003
B_g 支払準備率	-	-	-5.966	0.18674	-	-
μ_g	-	-	0.225	0.01899	-	-
対数尤度(-2logL)	569.10		507.12		603.07	
AIC	583.10		523.12		615.07	
Divergence	0.146		0.159		0.099	
K-S Distance	59.18%		64.74%		56.84%	

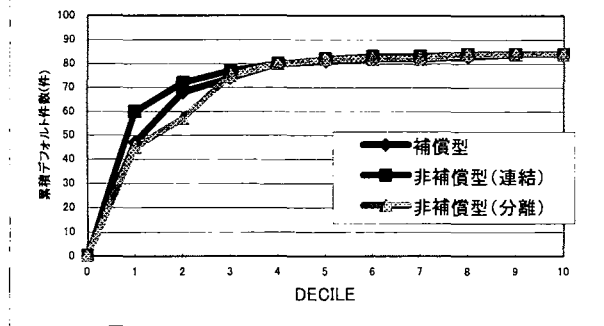
5. 考察

バーゼル合意に基づく自己資本規制の改定に向けた今日の動きに見られるように信用リスクの計量化に当たり、信用格付を基礎とする考え方が注目されるようになってきた。例えば、BIS 規制の新提案では銀行の抱える信用リスク量を把握するにあたり「内部格付手法」を利用する場合には銀行独自の格付を利用することが認められているが、そもそも各銀行の抱えるリスク量に相応しい自己資本の積み増しを行うためにも、現時点で銀行各々が抱えるリスク量を正確に把握できる精緻な確率モデルが必要となることは言うまでもない。

これに対し、従来広く一般に用いられてきた手法の一つに線形補償型二項ロジットモデルに基づくデフォルト確率の予測が挙げられるが、冒頭で述べたように、企業のデフォルト傾向と各財務指標との間には、補償型というよりはむしろ非補償型の関係が成り立つ傾向が強いことが経験則から推測される。一方で、信用リスクモデルにおいては、例えば自己資本比率と負債比率との関係のように相互に強い負の相関を持つ財務指標がリスクファクターとして用いられることが多いが、こうした変数を線形補償型モデルに適用することは、前述した通り確率的な問題点がある。そこで本論では従来型モデルの代替案として非補償型ロジットモデルを用いたデフォルト確率予測モデルを構築することでこうした問題点を克服することを試み、併せて本論における提案モデルに実データを適用することで本論における提案モデルが従来型モデルに比べ優位性を有することを示した。

図.1 は従来の線形補償型モデル、非補償型モデルそれぞれによるデフォルト企業の補足曲線(リフトカーブ)を描いたものである。このグラフからもわかるように、本論で提案した非補償型モデルによるデフォルト企業の補足力についてみると、分離型モデルでは従来の線形補償型モデルに比べ、補足力が落ちるものの、連結型モデルでは従来型モデルに比べ非常に高い補足力を示しており、非補償型(分離型)モデルを仮定した本論における提案モデルは、従来型モデルに対しモデルの改善効果が認められると考えられる。このことは、前章において算出された対数尤度、AIC 統計量からも明らかである。

図.1 累積デフォルト件数の推移



最後に、本研究の応用領域と今後の課題について言及する。

本論では、財務指標のみを用いてモデルの妥当性を検証してきた。しかし、一般的に企業のデフォルト確率予測モデルを運用するに当たっては特に対象となる企業規模が小さくなるに従い財務データの粉飾や意図的なデータ操作が行われることも多く、財務データそのものに対する信頼性が低下するといった問題点が指摘されており、こうした問題点を解決するために定性的な項目をモデルの中で考慮することが近年求められている。その際、どのようにして、あるいはどの程度の割合で定性要因をモデルの中に反映させてゆくのかといった点が問題となる。実務的には、定性要因と財務データを同時に線形補償型二項ロジットモデルの中で考慮するという方法や、定性要因のみからのパラメータ推定と財務データのみからのパラメータ推定を独立に行い、経験則に基づきある一定の割合でそれぞれを加味するという方法がとられることが多いが本論における非補償型の提案モデルを利用することにより、確率的にもより適切に定性要因と財務指標をモデルの中へ同時に取り込むことが可能になるのではないかと期待できる。

本論では企業のデフォルト確率を予測するに当たり、従来型の線形補償型二項ロジットモデルに代えて非補償型ロジットモデルを使用することによりモデルの改良を図ることを試みた。パラメータを推定する際、現在 SAS システムでは LOGISTIC プロシージャを利用することにより容易に二項ロジットモデルにおけるパラメータ推定を実行することが可能である。パラメータ推定のアルゴリズムとしても総当り法以外に変数選択法(Stepwise 法)等を使用することも可能である。一方で、本論で提案した非補償型ロジットモデルについて見るとその応用可能領域が、近年徐々に注目されつつある。このモデルは企業デフォルト傾向の予測のみならず、マーケティングリサーチにおける消費者選択行動の予測・土木工学における旅行者の交通手段選択の予測等幅広い分野で応用可能であることが先行研究によって報告されており、今後その活用領域はますます広がることが予想される。しかしながら、現在の SAS システムにおいて、非補償型ロジットモデルのパラメータ予測を行う際には、NLP プロシージャ等を利用した細かなプログラミングが必要とされるため、非補償型モデルに対してもより簡便な利用環境が開発されることを期待する。また、本論ではモデルにとって最適な利用変数を決定するに当たり、総当り法による変数選択を実施したが、この方法は必ずしも最適な変数の組合せを選択するとは限らない。またモデルの中で使用される変数の数が増加するに従いモデルの中で考慮しなければならない変数の組合せ数が急激に増加してしまい、最適なパラメータを推定するまでにかなりの時間が必要とされるといった問題点が予てより指摘されている。こうした問題点に対処するために、これまで一般の線形モデルや線形補償型二項ロジットモデルでは、変数選択法(Stepwise/Backward/Forward)による変数選択アルゴリズムが考え出されてきたが、同様の変数選択アルゴリズムを非補償型ロジットモデルに適用するための統計的手法の開発もあわせて必要になってくると考えられ、これらの研究課題を今後の研究に委ねたいところである。

以上

6. 謝辞

本論文を掲載するに当たり、論文発表の機会を与えて頂きました SAS Institute Japan 社様に対し、この場を御借り致しまして深く御礼申し上げます。

付録1 連結型モデルにおけるパラメータ推定を実施するための SAS プログラムの例

本論における連結型モデルによるパラメータ推定を行う際に使用した SAS プログラムを以下に示す。

パラメータの推定には分離型同様全ての変数の組合せを考慮した上で対数尤度が最も小さくなる変数の組合せを見つけ出す総当たり法を採用した。

```
/* SAS プログラム */
/* 最尤推定法におけるパラメータ初期値の乱数発生 */
Data _null ;
  call symput("Ba",rannor(0));
  call symput("Bb",rannor(0));
  call symput("Bf",rannor(0));
  call symput("Bg",rannor(0));
  call symput("Ua",rannor(0));
  call symput("Ub",rannor(0));
  call symput("Uf",rannor(0));
  call symput("Ug",rannor(0));
Run;

/* NLPプロシージャによるパラメータ推定 */

Proc NLP Data=dataset
  Tech=Newrap OUT=OUT1 outest=outest1
  cov=2 vardef=n pcov pstderr;
  Parms Ba=&Ba, Bb=&Bb, Bf=&Bf, Bg=&Bg,
  Ua=&Ua, Ub=&Ub, Uf=&Uf, Ug=&Ug;

  Va=Ba*(Xa-Ua);
  Vb=Bb*(Xb-Ub);
  Vf=Bf*(Xf-Uf);
  Vg=Bg*(Xg-Ug);

  Pa=1/(1+exp(-Va));
  Pb=1/(1+exp(-Vb));
  Pf=1/(1+exp(-Vf));
  Pg=1/(1+exp(-Vg));

  P=      ((Pa))*
          ((Pb))*
          ((Pf))*
          ((Pg));

  LL=Default*log(P) +(1-Default)*log(1-P);
  Max LL;
  profile Ba Bb Bf Bg Ua Ub Uf Ug / alpha=0.05;

Run;
```

付録. 2 分離型モデルにおけるパラメータ推定を実施するための SAS プログラムの例

本論における分離型モデルによるパラメータ推定を行う際に使用した SAS プログラムを以下に示す。

パラメータの推定には全ての変数の組合せを考慮した上で対数尤度が最も小さくなる変数の組合せを見つけ出す総当り法を採用した。

```
/* SAS プログラム */
```

```
/* 最尤推定法におけるパラメータ初期値の乱数発生 */
```

```
  Data _null_;  
    call symput("Ba",rannor(0));  
    call symput("Bb",rannor(0));  
    call symput("Bc",rannor(0));  
    call symput("Bd",rannor(0));  
    call symput("Be",rannor(0));  
    call symput("Bf",rannor(0));  
    call symput("Bg",rannor(0));  
    call symput("Ua",rannor(0));  
    call symput("Ub",rannor(0));  
    call symput("Uc",rannor(0));  
    call symput("Ud",rannor(0));  
    call symput("Ue",rannor(0));  
    call symput("Uf",rannor(0));  
    call symput("Ug",rannor(0));  
  Run;
```

```
/* NLP プロシージャによるパラメータ推定 */
```

```
  Proc NLP Data=dataset  
    Tech=Newrap OUT=OUT1 outest=outest1  
      cov=2 vardef=n pcov pstderr;  
    Parms Bb=&Bb, Bf=&Bf, Bd=&Bd,  
          Ub=&Ub, Uf=&Uf, Ud=&Ud;  
  
          Vb=Bb*Xb-Ub;  
          Vf=Bf*Xf-Uf;  
          Vd=Bd*Xd-Ud;  
  
          Pb=1/(1+exp(-Vb));  
          Pf=1/(1+exp(-Vf));  
          Pd=1/(1+exp(-Vd));  
  
          P=1-(1-Pb)*(1-Pd)*(1-Pf);  
  
    LL=Default*log(P) +(1-Default)*log(1-P);  
    Max LL;  
  
    profile Bb Bf Bd Ub Uf Ud / alpha=0.05;  
  
  Run;
```

- 参考文献 -

- Altman,Edward,I,Financial Ratios (1968), “Discriminant Analysis and The Prediction of Corporate Bankruptcy”, *Journal of Finance*, 23(4),589-609
- Altman,Edward,I (1970), “Ratio Analysis and The Prediction of Firm Failure: Reply”, *Journal of Finance*, 25(5), 1169-1172
- Altman,Edward,I (1971), “Railroad Bankruptcy Propensity”, *Journal of Finance*, 26(2), 333-345
- Altman,Edward,I,(1976), “A Financial Early Warning System for Over-The-Counter Broler-Dealers”, *Journal of Finance*, 31(4), 1201-1224
- Beaver, William (1966), “Financial Ratios As Predictors of Failure”, *Journal of Accounting Research*, 4(Supp),71-111
- Bettman, James R. and Jacob Jacoby (1976),”Patterns of Processing in Consumer Information Acuisition”, *Advances in Consumer Research*,3,315-320
- Chakravarti, Laha, and Roy, (1967). Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, pp. 392-394.
- Dawes, Robin M and Bernard Corrigan(1974), “Linear Models in Decision Making”, *Psychological Bulletin*, 81 (March), 95-106
- Harter, H.L., Khamis, H.J. and Lamb, R.E.(1984), “Modified Kolmogorov-Smirnov tests of goodness of fit”, *Communications in Statistics, Simulation and Computation*, Vol. 13, No. 3, 293-323.
- Johnson, Eric.J. and Robert J. Meyer (1984),”Compensatory Choice Models of Noncompensatory Processes: The Effect of Varying Context”,*Journal of ConsumerResearch*, Vol.11, June,pp.528-541
- Johnson, Eric.J., Robert Meyer and Sanjoy Ghose (1989),”When Choice Models Fail: Compensatory Models in Negatively Correlated Environments”, *Journal of Marketing Research*, 26(August), 255-270
- Khamis, H.J.(1992), “The delta-corrected Kolmogorov-Smirnov test with estimated parameters”, *Journal of Nonparametric Statistics*, Vol. 2, 17-27.
- Khamis, H.J. (1990), “The delta-corrected Kolmogorov-Smirnov test for goodness of fit”, *Journal of Statistical Planning and Inference*, Vol. 24, 317-335.
- Khamis, H.J.(2000), “The two-stage delta-corrected Kolmogorov-Smirnov test”, *Journal of Applied Statistics*, Vol. 27, No. 4, 439-450.
- Martin,Daniel (1977), “Early Warning of Bank Failure: A Logit Regression Approach”, *Journal of Banking and Finance*,1(3),249-276
- Payne, John W. and E.K.Easton Ragsdale (1978), “Verbal Protocols and Direct Observation of Supermarket Shopping Behavior: Some Findings and a Discussion of Methods”, in *Advances in Consumer Research*, 5,571-577
- 片平秀貴(1998), “ロジット分析を用いた満足化モデル”,消費者選択行動のニューディレクションズ,関西学院大
- 森平爽一郎(1999), “信用リスクの測定と管理”, 証券アナリストジャーナル,99.9
- 森平爽一郎(1999), “信用リスクの測定と管理”, 証券アナリストジャーナル,99.11

日本SASユーザー会（SUGI-J）

SAS Risk Dimensionsによる統合リスク分析のご紹介

○嘉陽亜希子 鬼頭拓郎 尾高雅代 田中愛
カスタマーサービス本部プロフェッショナルサービス第1部
SAS Institute Japan 株式会社

Introduction of integrated risk analysis using SAS Risk Dimensions

Akiko Kayo Takuro kito Masayo Odaka Ai Tanaka
Professional Service Department 1 Customer Services Division
SAS Institute Japan Ltd.

要 旨 金融分野においてリスク管理という場合、一般的には3つのリスクが考えられる。マーケットリスク、クレジットリスク、オペレーショナルリスクである。マーケットリスクとは、金融マーケットの変動によるポートフォリオの価値変化(とくに起こりうる損失)の量であり、クレジットリスクはカウンターパーティの格付悪化やデフォルトといった状況変化によるポートフォリオの価値変化(とくに起こりうる損失)の量である。マーケットリスクおよびクレジットリスク以外の不特定なリスクを総じてオペレーショナルリスクという。

この発表では、SAS Risk Dimensions の環境で統合リスク量の計測を行なう方法について説明する。SAS Risk Dimensions における分析環境の特長、リスクファクターのモデル化、プライシングロジックの登録、ポートフォリオのリスク分析プロジェクトの構築、レポート機能など、具体的な一連の手順を説明する。

キーワード： リスク計測 SAS Risk Dimensions 統合リスク管理

1. リスクとは

1-1. リスクの考え方

金融分野における代表的なリスクとして、マーケットリスク、クレジットリスク、オペレーショナルリスクがある。マーケットリスクとは、金融マーケットの変動によるポートフォリオの価値変化(とくに起こりうる損失)の量であり、例えば金利カーブの変化や株価指標の動向によって保有している債権の価格や株価などは変動する。クレジットリスクとは、カウンターパーティの格付悪化やデフォルトといった状況変化によるポートフォリオの価値変化(とくに起こりうる損失)の量である。例えば格付が下がれば金利スプレッドの影響により評価価格は下がり、また、デフォルトの場合は保有商品の価値自体が無くなることになる。マーケットリスクおよびクレジットリスク以外の不特定なリスク、例えばシステム障害などによる損失量を総じてオペレーショナルリスクという。

このように、企業は様々な要因により損失を被る可能性を有しており、これらの複雑に絡み合ったリスクを正確に測定するのは至難の業である。SAS Risk Dimensions では、複雑なリスク計測

を系統的に管理することで、リスク要因の把握や様々な角度からのリスク量の把握をより簡便にわかりやすく行うことができる。

1-2. リスクの基礎的な構成要素

リスクの基礎的な構成要素は、マーケットデータなどから算出するリスクファクター値、リスクファクターの変動を予測するモデル、各金融商品に対応するプライシング、ポジションデータの4つである。

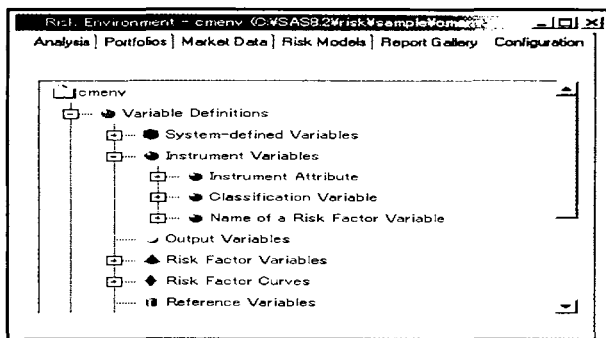
リスクファクターとは、金利や格付などのようにマーケット情報から取得されるものであり、日々変動するマーケットにおいてはリスクファクターも一定ではなく変動する。その変動によりリスク量も変化することになる。これらのリスクファクターの変動をモデル化できれば、将来における商品の価値変動を予測することができる。リスクファクターに関するモデルはモンテカルロシミュレーションなどの分析において用いられる。予測モデルからの各リスクファクターの予測値をプライシングアルゴリズムにインプットとして与え、各商品の価値を計算する。得られた商品の価値とポートフォリオの保有比率から損益分布を描きリスク量を計測することになる。

「2. Risk Dimensions におけるリスク構成要素の取扱い」では、同ソフトウェアにおけるこれら4つのリスク計測のための構成要素の取扱いについて記述する。

Risk Dimensions 利用の準備作業

分析環境において、リスク計測に必要な情報およびプライシング方法を系統的に登録し、データを用意した後、実装されている分析エンジンを利用することで、様々な角度からのリスク計測が可能となる。この分析環境で使用する変数名、マトリクス等は一元的に登録、管理される。これらの登録された変数と入力データの変数名とを紐付けることで、データソースの形式にとらわれない分析が可能となる。

登録変数の種類はシステム変数、商品変数、リスクファクター変数、リスクファクターカーブ、参照変数の5つに大きく分類される。



2. Risk Dimensions におけるリスク構成要素の取扱い

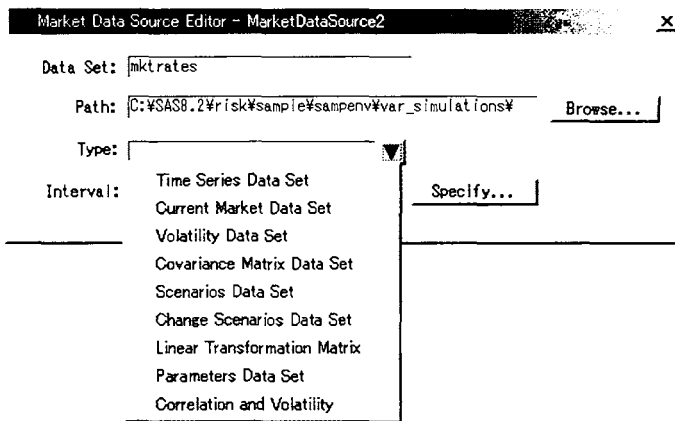
2-1. マーケットデータの取扱い

Risk Dimensions におけるマーケットデータの登録は、ソースファイルの登録、パラメータ行列の指定、および値を変換するためのプログラム指定から成り立つ。

マーケットデータソースファイルの登録

[Market Data]タブ→[Market Data Sources]で、分析のインプットとして Risk Dimensions で使用するマーケットデータを含んだ SAS データセットまたは SAS データビューを登録する。登録できるマーケットデータの種類は次の 9 種類である。

- Time Series : 時系列データ
- Current Market : カレントマーケットデータ
- Volatility : リスクファクターのボラティリティ推定データ
- Covariance Matrix : 分散共分散行列
- Scenarios : シナリオデータ
- Change Scenarios : 変化率シナリオデータ
- Linear Transformation Matrix : リスクファクターベクトルの線形変換行列
- Parameters : パラメータ行列
- Correlation and Volatility : 相関行列および対応する標準偏差ベクトル

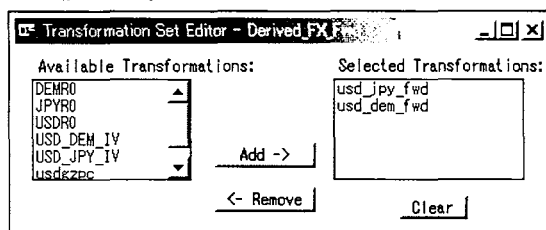


パラメータ行列の指定

[Market Data]タブ→[Parameter Matrices]では、[Market Data Sources]の登録において「Parameters」および「Linear Transformation Matrix」と指定したデータソースから必要な行および列のみを抜き出し、関数やサブルーチン、プライシングプログラム等で利用可能な行列として定義する。

リスクファクター値算出プログラムの指定

[Market Data]タブ→[Transformation Sets]は、[Configuration]タブ→[Method Program Library]→[Risk Factor Transformation]で登録されたプログラムのリストである。登録したリストを[Analysis]→[Analysis Project]で指定することで、選択されたマーケットデータから Analysis Project で扱うリスクファクターの値を自動的に計算し、分析結果に反映させることが可能となる。



2-2. 関数、サブルーチン、リスクファクター値算出およびプライシングに用いるプログラム

Risk Dimensions では、プライシングおよびマーケットデータからリスクファクター値への変換等に用いる関数、サブルーチン、プログラムを一元的に登録・管理する。下記では、それぞれのコーディング例および登録方法について説明する。

関数およびサブルーチンの登録

[Configuration]タブ→[Function Library]では SAS もしくは C 言語による関数およびサブルーチンを登録できる。登録した関数およびサブルーチンは、[Configuration]タブ→[Method Program Library]において、商品タイプ別のプライシング等に使用するプログラムから呼出して利用する。

関数登録の例 『Money Market Deposits のプライシング関数』

```
/*- 関数の開始宣言 -----*/  
function mmdepprc( valdate, amount, conrate, matdate, spotval[*], maturity[*] )  
    label = "Money Market Deposit Pricing" ;  
  
    npoint = dim(spotval);  
    if valdate > matdate then return(0);  
  
    /*- 評価日から満期までの時間を算出 -----*/  
    time = intck( 'day', valdate, matdate)/360;  
  
    /*- 線形補完によるスポットレート取得 -----*/  
    if ( time <= maturity{1} ) then spotrate = spotval{1};  
    if ( time >= maturity(npoin) ) then spotrate = spotval(npoin);
```

```

if (time > maturity(1)) and (time < maturity(npoint)) then do;
  do j=1 tonpoint-1;
    if (time > maturity(j)) and (time < maturity(j+1)) then do;
      spotrate = spotval[j]+((spotval[j+1]-spotval[j])*(time-maturity(j))
        /(maturity[j+1]-maturity(j)));
      go to prc;
    end;
  end;
end;
/*- 評価日時点の価格 -----*/
prc:
prc = amount * (1. + contrate)* exp (-(spotrate * time));
return( prc );
/*- 関数の終了宣言 -----*/
endsub;

```

上記関数の場合、下記のようにプログラムを書くことで戻り値 prc を取得する。

```

_value_ = mmdeprc( 引数を指定 );

```

プライシングプログラム、リスクファクター値算出プログラムの登録の登録

[Configuration]タブ→[Method Program Library]では商品タイプ別のプライシングに必要なデータ項目の加工(Instrument Input)、各商品のプライシング(Instrument Pricing)、マーケットデータのリスクファクター値への変換(Risk Factor Transformation)を行うためのプログラム登録を行う。SAS によるプログラム、および Function Library で登録した SAS もしくは C 言語による関数やサブルーチンを使用することができる。[Configuration]タブ→[Instrument Types]の登録では、商品タイプごとにプライシングのための「Method Program」を指定する必要があり、必要に応じて「Instrument Input」を指定できる。また、[Market Data]タブ→[Transformation Sets]では登録した「Risk Factor Transformation」を利用できる。

Instrument Input の登録例(Govbond_lookup 関数は Function Library にて定義済とする)

```

method Gov_Bond_Input desc= "Gov Bond Input Lookup"
  kind= input
  "call govbond_lookup( iss_type, coupfreq );";

```



```

Govbond_lookup サブルーチン
subroutine govbond_lookup (iss_type$, freq) kind=input;
  outargs freq;
  if iss_type = "BTAN"      then freq = 12;
  if iss_type = "OAT"       then freq = 12;
  |
  if iss_type = "Treasury" then freq 6= ;
endsub;

```

Instrument Pricing の登録例

(Govbondprc 関数は Function Library にて定義済とする)

```

method Gov_Bond_PF desc= "Gov Bond PF"
  kind=price
  "_VALUE = GOVBONDPRC( _date_, par_lc, coupfreq, coupon, "
  "mat_date, zcurve, zcurve.MAT );"
;

```

Risk Factor Transformation の登録例

```

method USD_DEM_FWD desc= "USD/DEM Forwards by inverse"
  kind= trans;
  usddem1m = 1./ demusd1m ;
  usddem3m = 1./ demusd3m ;
  usddem6m = 1./ demusd6m ;
endmethod;

```

2-3. リスクファクターのモデリング

リスクを計測するには、特定のマーケット環境における金融商品の価値を予測する必要がある。RiskDimensions では、目的の商品に関する情報とマーケット情報から商品の将来価格を求めするために、リスクファクターモデルを登録する機能が備わっている。リスクファクターモデルとして Cox Ingersoll Ross モデルや Vasicek モデル、幾何ブラウン運動、ARCH や GARCH などの時系列モデル等様々なモデルを登録できる。登録は[RiskModels]タブにて行なう。

以下に幾何ブラウン運動と GARCH モデルの登録例を示す。

幾何ブラウン運動

モデル式

$$\begin{cases} x_t = x_{t-1} + \mu x_{t-1} + \eta_t \\ \eta_t = \sqrt{h_t} \times \varepsilon_t \\ h_t = \sigma^2 \times x_{t-1}^2 \end{cases}$$

```

Model Program Editor - GBMSOURCE
00001
00002   endogenous x ;
00003   params mu sigma ;
00004
00005
00006   x = lag(x) + mu * lag(x);
00007   h.x = sigma * sigma * lag(x) * lag(x);
00008
00009   label sigma = "Diffusion Parameter";
00010   label mu = "Drift Parameter";
00011
00012
  
```

h.x: x の分散を定義している
この定義がない場合
x の分散は σ^2 乗で一定となる

◆ GARCH モデル

モデル式

$$\begin{cases} ret_t = mean + \eta_t \\ \eta_t = \sqrt{h_t} \times \varepsilon_t \\ h_t = \omega + \sum_{i=1}^q \alpha_i \times \eta_{t-i}^2 + \sum_{j=1}^p h_{t-j} \end{cases}$$

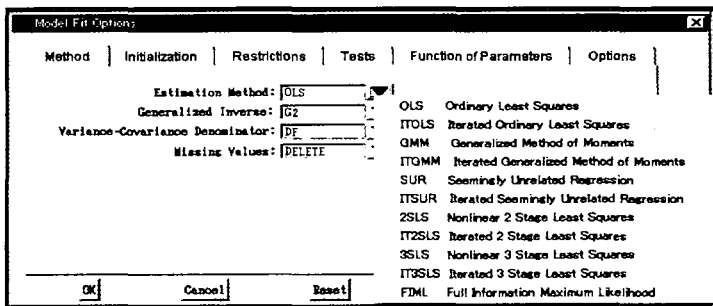
```

Model Program Editor - GARCH11.SOURCE
00001
00002   ret = mean;
00003   h.ret = arch0 + arch1 * zlag( resid.ret * resid.ret ) + garch1 * zlag(h.ret) ;
00004
00005   label arch0 = "Constant part of conditional volatility";
00006   label arch1 = "Coefficient of lagged squared residuals";
00007   label garch1 = "Coefficient of lagged conditional volatility";
00008
00009
  
```

上記のように、Model Program Editor ウィンドウにて、SASコードによりモデル式を指定する。このコードはSAS/ETSのMODELプロシジャにより実行される。つまり、ここではMODELプロシジャを使用したPROCステップの一部を指定していることになる。

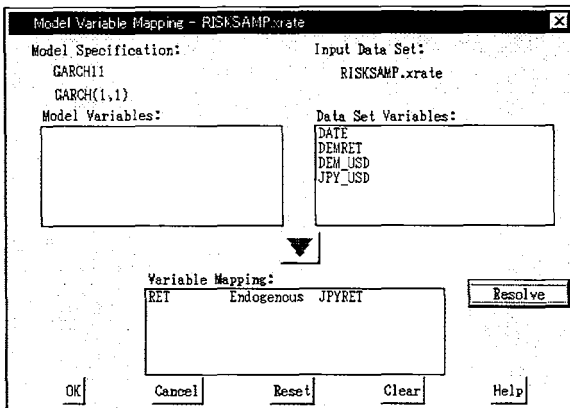
従って Model プロシジャで記述できるモデル式は、全て取扱いが可能といえる。

各モデル式で使用されている変数の役割の設定は、幾何ブラウン運動の例の①のように SASプログラミングで記述するか、[Model]→[Variable Definition...]で行える。



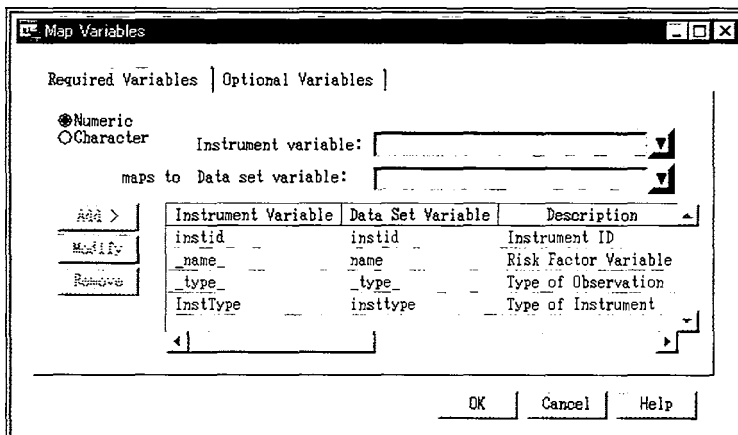
OLS法などの推定方法や欠損値の扱い、その他の設定はオプションとして指定できる。

以上の設定を終えたあとに、実際にデータにモデルをあてはめ推定を行なう。データの使用変数とモデル式で定義されている変数とを紐付けをすることでモデル式を適応できる。従って形式が一致していれば、データの使用変数名の制約はないということになる。



2-4. ポジションデータの取扱い

データの構造が決まり、プライシングアルゴリズムの登録後に、分析対象となるポートフォリオデータを[Portfolios]タブにてリスク分析環境に登録する。データ登録では、データの使用変数と[Configuration]タブで既に登録済みの変数との紐付けを行なう。リスクファクターモデルと同様にデータの使用変数名に制約はない。



3. Risk Dimensions によるリスク計測

前章までに、リスク計測に必要な構成要素の Risk Dimensions 環境への登録を説明した。本章では、前章までに登録した構成要素を組合わせた様々なリスク計測をプロジェクトとして管理する方法について説明する。

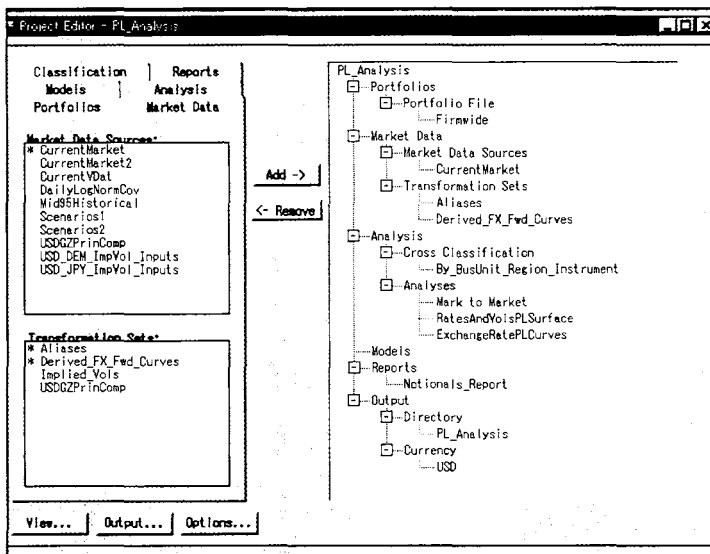
3-1. リスク計測のプロジェクト管理

必要な構成要素を全て登録した後、[Analysis] タブにおいて、これらを組み合わせて目的に応じた様々な分析を行うことができる。Risk Dimensions では、構成要素の組合せを「分析プロジェクト」として扱う。分析プロジェクトは、分析作業を実行させるためのスクリプトで、名前を付けて保存することができる。分析プロジェクトを実行すると、保存された設定項目に従って Risk Dimensions が自動的に作業を開始する。設定項目は以下の通りである。

- ① 処理の対象となるポートフォリオファイル
- ② 処理に使用するマーケットデータ
- ③ 使用するリスクファクターの変換設定
- ④ 実行する分析のタイプ(3-2 参照)
- ⑤ 結果を分類する Cross-Classification の設定
- ⑥ レポーティングの設定(4章参照)

リスク分析を行うときには、ポートフォリオ全体だけでなく、ポートフォリオの一部だけを対象にしたいというケースもある。分析の切り口を変えたいときに⑤の Cross-Classification の設定において、グループ変数を指定して実行結果を分類することで、ポートフォリオの細部をチェックすることや、結果を細かく切り貼りすることが可能になる。

以下に、完成したプロジェクト設定の例を示す。



3-2. Risk Dimensions の分析手法

Risk Dimensions で実行できる分析のタイプは以下の通りである。

センシティブティ分析[Sensitivity Analysis]

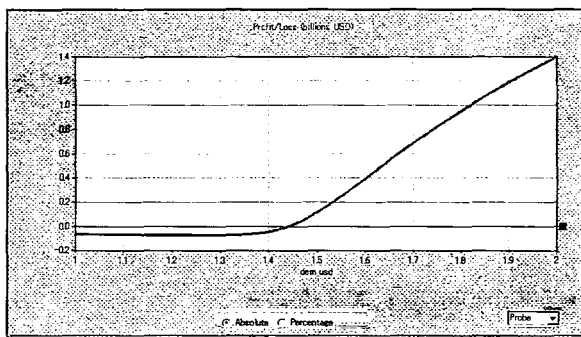
リスクを計測する指標として一般的に利用される、デルタ、ガンマ、シータを算出し(ローとベガは暗黙的に使用可能)、ポートフォリオのリスクをモニタリングする。

シナリオ分析[Scenario Analysis]

ユーザーが指定するシナリオ通りにリスクファクターが動いた場合の、特定の期間にわたるポートフォリオの損益を算出する。

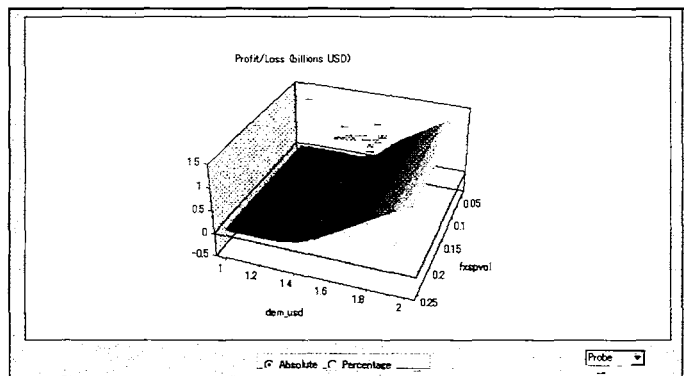
損益曲線分析[Profit/Loss Curve Analysis]

ひとつのリスクファクターだけが変動し、他のファクターが動かない場合に、変動するファクターの関数としてポートフォリオ損益の変化を算出する。



損益 2次元分析[Profit/Loss Surface Analysis]

2つのリスクファクターだけが変動し、他のファクターが動かない場合に、変動するファクターの関数としてポートフォリオの損益の変化を算出する。



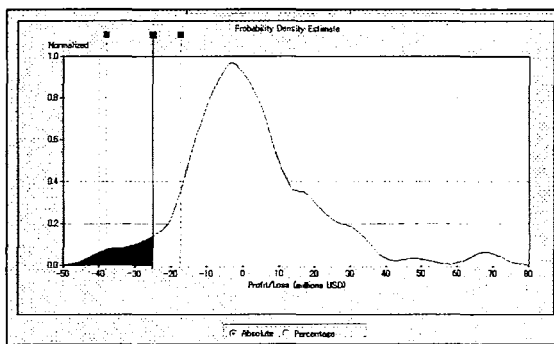
ヒストリカルシミュレーション[Historical Simulation]

過去に生じたマーケット変動が将来もそのまま起こるとする考えのもと、リスクファクターおよびポートフォリオ価値の過去の値を単純集計することによってポートフォリオの損益の変化を算出

する。

モンテカルロシミュレーション[Monte Carlo Simulation]

過去のマーケットデータからリスクファクターのモデルを作成し、各モデルに初期値を、誤差項に乱数より得られた数値をそれぞれ代入すると、モデルのアウトプットとしてリスクファクターの将来の予測値が得られ、それを用いて個別銘柄のプライシングを行う。これに現在のポートフォリオの保有比率をかけ合わせると、将来のポートフォリオの価値やポートフォリオ全体の収益率が算出される。この初期値の代入から収益率算出までのステップを反復することによって、ポートフォリオの将来の収益を分布として得ることができる。



シナリオシミュレーション[Scenario Simulation]

過去に起きた特定期間のマーケット変動データ(あるいはリスクファクターを大きく変化させるシナリオデータ)をユーザーが指定して、そのシナリオのもとでのポートフォリオの収益の分布を生成する。

デルタ・ノーマル[Delta-Normal]

リスクファクターの変動に対するポートフォリオの感応度と、リスクファクターの分散共分散行列よりポートフォリオの変動の分散を計算し、これに信頼水準と保有期間の条件を与え、VaR を算出する。デルタ・ノーマルでは、リスクファクターの変動が多変量正規分布または対数正規分布となることを前提としている。

カレント・エクスポージャー分析[Current Exposure Analysis]

リスクファクターモデル、ポジションデータ、プライシング・プログラムを用いて、ポートフォリオの各デリバティブ取引の時価評価(Mark to Market)を個別に算出する。

ポテンシャル・エクスポージャー分析[Potential Exposure Analysis]

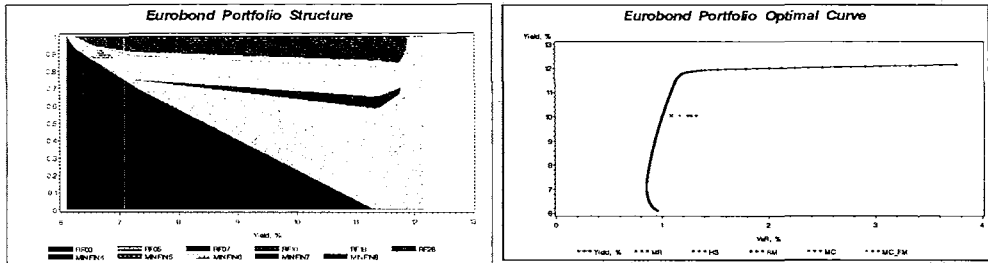
モンテカルロシミュレーションにより、デリバティブ取引の契約期間中に生じるカレント・エクスポ

ージャーの増加見込み額(ポテンシャル・エクスポージャー)を予測する。

3-3. リスクレポート

SAS/EIS ソフトウェアや REPORT プロシジャなど、SAS のレポート作成ツールを利用してカスタム・レポートを作成することができる。

Risk Dimensions では、レポート作成用バッチプログラムや EIS アプリケーションのレポート・テンプレートを分析環境に登録することができる。レポート書式を登録しておけば、[Analysis] タブで分析の結果を保存するファイルに自動的にテンプレートの書式を設定することが可能になる。



口頭論文発表
調査・マーケティング

日本S A Sユーザー会 (SUGI-J)

建築生産における建築物の耐久性確保に関する 実務者の意識と実態

○小島 隆矢* 小野 久美子** 植木 暁司***

* 独立行政法人建築研究所 住宅・都市研究グループ 主任研究員

** 国土交通省国土技術政策総合研究所 住宅研究部 研究官

*** 国土交通省国土技術政策総合研究所 総合技術政策研究センター 主任研究官

The questionnaire to construction working members
about the durability of a building

○Takaya KOJIMA* Kumiko ONO** Kyouji UEKI**

* Building Research Institute

** National Institute for Land & Infrastructure Management

要 旨

本報では、建設プロジェクトに関わる実務者を対象としたアンケート調査のデータに基づき、建築物の耐久性確保についての取り組みの実態および実務者の意識に関する分析結果を報告する。質的変数のグラフィカルモデリング(グラフィカル対数線形モデリング)により定性的な因果構造を同定し、順序ロジスティック回帰分析により定量モデル化および内容の解釈を行ったところ、現状の問題点が浮き彫りとなる結果となった。なお、グラフィカルモデリングには廣野元久氏作成のソフト L-GM、順序ロジスティック回帰分析等には JMP 5.0.1-J を用いている。

キーワード： グラフィカルモデリング、順序ロジスティック回帰、JMP、建築実務

1. はじめに

近年、住宅の品質確保の促進等に関する法律など、建築物の施主・ユーザー保護の視点に立った政策が展開されているが、建築物の耐久性については、ユーザーニーズや立地条件等の状況に応じて仕様を確定する技術が確立されていないのが現状である。そのため、独立行政法人建築研究所・国土交通省国土技術政策総合研究所では、ユーザーからの要求レベルに応じて適切な建築材料・部材・構法の選定する目的志向型耐久設計を実現するための支援ツールの研究開発に取り組んでいる¹⁾。

本稿では、この研究の一環として2002年2～3月に実施した、建設プロジェクトに関わる実務者を対象としたアンケート調査について報告する。調査目的は、現状の建築生産プロセスにおける取り組みの実態、実務者の意識などを把握することにある。業界団体などの紹介によりリストアップした調査対象者(設計事務所、総合請負業、専門工事業、材料メーカー、住宅メーカー、官公庁等の業種にて、企画・計画、意匠設計、構造設計、工事監理、施工管理、専門工事、材料、研究その他の業務に従事する実務者)約490名に配布し、有効回答数188名(有効回収率38%)を得た。

2. 分析方針

本稿で主に取り上げる設問は、建築生産プロセスを「企画・計画」から「竣工・維持管理」まで 6 段階のフェイズとして、各フェイズについて、下記 3 項目を評価させたものである。

・ニーズ提示頻度：

施主・発注者から耐久性に関わる要望が提示されることは、1.ほとんどない～4.よくある(4段階評価)

・ニーズ確定必要性：

耐久性に関わる要望が確定していることが、1.必要でない～4.必要である(4段階評価)

・重要度：

耐久性確保のための取り組みが、1.重要でない～4.重要である(4段階評価)

これらの設問のデータは、回答者×フェイズ×上記 3 項目 という、3 相 3 元データの形式をなすことになる。そこで、回答者×フェイズを観測個体として、因果関係の分析を実施した。より具体的には、上記 3 項目に「フェイズ」および「回答者の主な業務内容」を加えた 5 変数を分析対象として、質的変数のグラフィカルモデリング(グラフィカル対数線形モデリング)により定性的な因果構造を同定し、順序ロジスティック分析により定量モデル化および内容の解釈を行う。「業務」「フェイズ」の水準の内容は以下の通り。

・業務： 1.企画設計 2.意匠設計 3.構造設計 4.工事監理 5.工事計画管理

6.専門工事 7.材料製造販売 8.研究開発 9.その他

・フェイズ： 1.企画・計画 2.基本設計 3.実施設計 4.施工計画 5.施工実施 6.竣工・維持管理

なお、グラフィカルモデリング(以下、GM)には廣野元久氏作成のソフト L-GM、順序ロジスティック回帰分析等には JMP 5.0.1-J を用いている。

3. 予備的検討・事前処理

3.1 分析対象サンプル

有効回答者数は 188 名であるので、本来のサンプルサイズは有効回答者数 188 名×フェイズ 6 水準＝1128 となる。しかし、分析においては、以下に示すようにいくつかの事前処理が必要であった。

まず、「業務」は SA のはずであったが、複数の業務を選択した回答者が 19 名ほどいた。JMP を用いた分析では観測個体に重みをつけることができるので、これらの回答者のデータはダブルカウント(一部、トリプル)して、通常の回答者の 2 倍(3 倍)の行数を割り当てる代わりに、その行のウェイトを 1/2(1/3)とすることにより対応した。L-GM にはこの機能はないので、他の回答者に比べて 2 倍 or 3 倍のウェイトを与えることになるが、やむを得ない。

また、「ニーズ提示頻度」は、前の設問の回答により該当者のみ回答する形式であったため 140～150 名しか回答していない。これらの回答者のデータは、この設問を用いた分析においては欠測値として分析から除外している。

結局、分析対象サンプル数は、分析の種類により、980～1080 程度の範囲にて変動することとなる。

3.2 データのモニタリング

「ニーズ提示頻度」「ニーズ確定必要性」「重要度」の度数分布を図 1 に示す。「確定必要性」「重要度」に

については、「1:殆ど重要(必要)でない」の度数が非常に少ない。そこで、「2:あまり重要(必要)でない」と統合し、3段階評価として、以後の分析を進める。なお、「提示頻度(4段階)」も含め、いずれも数値が大きい水準ほど、頻度・重要度・必要性が高くなることを表している。

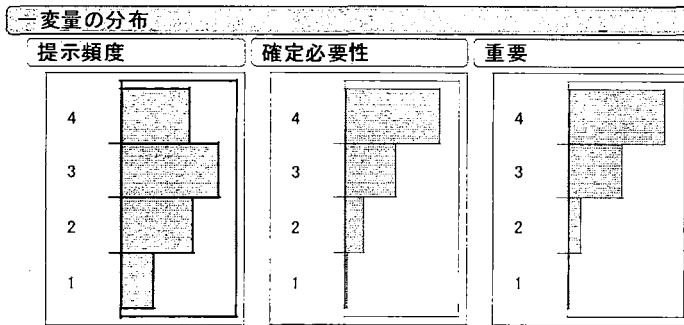


図1 「ニーズ提示頻度」「ニーズ確定必要性」「重要度」の度数分布

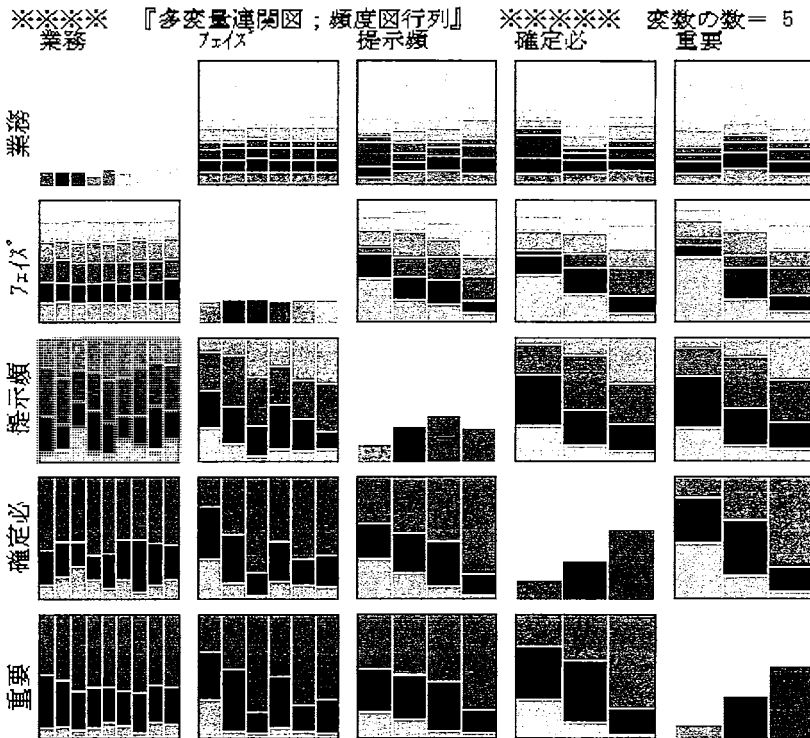


図2 多変量連関図(L-GMによる)

図2は、分析対象5変数の多変量連関図である(L-GMによる。水準の順序は、左 or 下から右 or 上に向かって、1,2,3…。)この図より、次のことが分かる。

- ・提示頻度・必要性・重要度については互いに正の相関が高い。
- ・各フェイズと、提示頻度・必要性・重要度の関係はどれも似ている。(重要度が高いフェイズは提示頻度が高い、等)

なお、2元分割表の独立性の検定結果は、もともと直交している業務-フェイズ(どの回答者も6つのフェイズについて答えているので)、および業務-重要度の2元分割表だけが有意ではなく(p=0.495)、他の変数の組に関してはいずれもp<0.005の水準で有意な関連を示している。

4. 質的変数のグラフィカルモデリング

4.1 独立グラフのモデリング

まず、この5変数間の条件付き独立関係を分析する。グラフィカル対数線形モデリング(詳細は文献²⁾などを参照されたい)により、下図のような独立グラフが得られた($\chi^2=1270.68(df=1804) p=1.00$)。

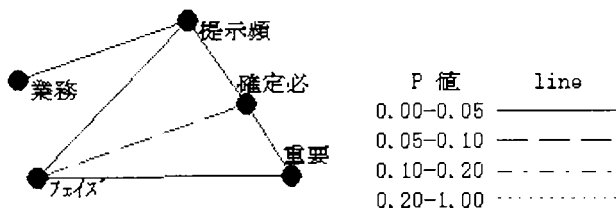


図3 独立グラフ

<補足>独立グラフの見方

独立グラフで線のない変数間は、他の変数を条件付きにすれば(ある水準ごとに層別して関連を調べる、ということ)独立になることを意味している。さらに、他の変数を介して間接的には線がつながる変数の場合、全サンプルでクロス集計を行えば関連があるように見えても、間を取り持つ変数で層別して(間接的な経路を切る)クロス集計をすれば、関連がなくなることを意味している。

独立グラフから因果関係を推論する場合、線のない変数間に直接的な(他の変数を介さない)因果関係を考える必要はほとんどない。線のある変数は直接的な因果関係の候補である。

ただし、線があるからといって必ず直接的因果関係がある、ということにはならない。「 $x \rightarrow z \leftarrow y$ 」(何ら関連のない2変数xyがzに影響する)という因果関係の場合に、結果系変数zで層別するとxyは独立ではなくなる。簡単な例としては、xyを2つの学科の試験の得点、zを総合得点と考えればよい。総合得点が同程度の人ばかり集めれば、両学科の得点は負の相関を示すであろう。

逆に、条件付き独立関係が「 $x - z - y$ 」という独立グラフで表される場合、「 $x \rightarrow z \leftarrow y$ 」という因果関係(因果合流)を考えることは、一般には否定される。xyに何らかの関連があった場合、その関連がzという結果系変数で説明されてしまったことを意味するからである。結果によって原因が説明されるというのは不自然である。

以上より、独立グラフは、因果グラフ(直接の因果関係を矢線で表した図)をもとに、因果合流する変数間に線を追加し、矢線を線でおきかえた図(「モラルグラフ」と呼ぶ)になっていることが期待される。これらの知識を使えば、独立グラフから因果グラフが推定できる場合がある。(補足終わり)

以下、図3の独立グラフに基づき、変数間の因果関係を推論する。「業務」「フェイズ」が原因となって、「提示頻度」「確定必要」「重要度」に違いが生じる、という因果の順序は明らかであるから、問題は結果系

3変数の順序である。

まず、「提示頻度→必要性←重要度」という因果合流は、重要－提示頻度の間が切れていることから否定される。次に、「業務→提示頻度←必要性」という因果合流も、業務－必要性の間が切れていることから否定される。結局、可能な因果の順序としては、「提示頻度→必要性→重要度」ということになる。この因果関係においては「業務→提示頻度←フェイズ」という因果合流で「業務－フェイズ」の間が切れているが、これは提示頻度によって関連が説明されたのではなく、もとより関連がなかったのである(どの業務の人も、全てのフェイズについて回答している)。

4.2 因果グラフのモデリング

これまでの考察により、因果の順序は以下のように決まったことになる。

{業務, フェイズ} → {提示頻度} → {必要性} → {重要度}

この順序情報を取り入れたグラフィカル対数線型モデリングの結果、得られた因果グラフを図4に示す。

n = 980 逸脱度 = 1114.743 (df = % 1684) p = 1.0000

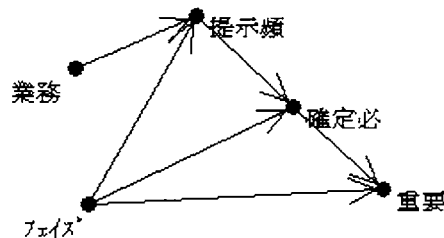


図4 因果グラフ

結果的に独立グラフに矢線をつけただけになったが、この手順を踏まないと、例えば「フェイズ－必要性」間の線が、直接の因果を表すのか、重要度への因果合流によるものかが分からない。

因果グラフは、以下のような因果関係を表している。

1) 「業務」「フェイズ」が「提示頻度」に影響する。

独立グラフで「業務－フェイズ」間が切れていることから、業務とフェイズの交互作用はなく、業務による違いとフェイズによる違いを単純に加算して提示頻度の分布が決まる。

2) 「フェイズ」「提示頻度」が「必要性」に影響する。交互作用の有無は不明。

3) 「フェイズ」「必要性」が「重要度」に影響する。交互作用の有無は不明。

「業務」が直接影響するのは「提示頻度」だけ、というのはやや意外な結果である。

以下、順序ロジスティック回帰分析により、上記 1)～3)の関係を具体的に調べていく。

5. 順序ロジスティック回帰分析

5.1 「ニーズ提示頻度」を目的変数とした分析

まず、「ニーズ提示頻度」を目的変数とした順序ロジスティック回帰を実施する。「業務」「フェイズ」の主効果および交互作用項(念のため)を説明変数としたモデルの要因効果(検定結果およびパラメータ推定値)は以下の通りである。なお、順序ロジスティック回帰の要因効果は、パラメータ推定値の数値が大きく

なるほど、目的変数が小さい水準の度数が多くなることを示している。

要因	自由度	Waldカイ2乗	p値(Prob>ChiSq)
フェイズ	5	80.4676874	0.0000
業務	8	16.6587336	0.0339
業務*フェイズ	40	50.5389351	0.1227

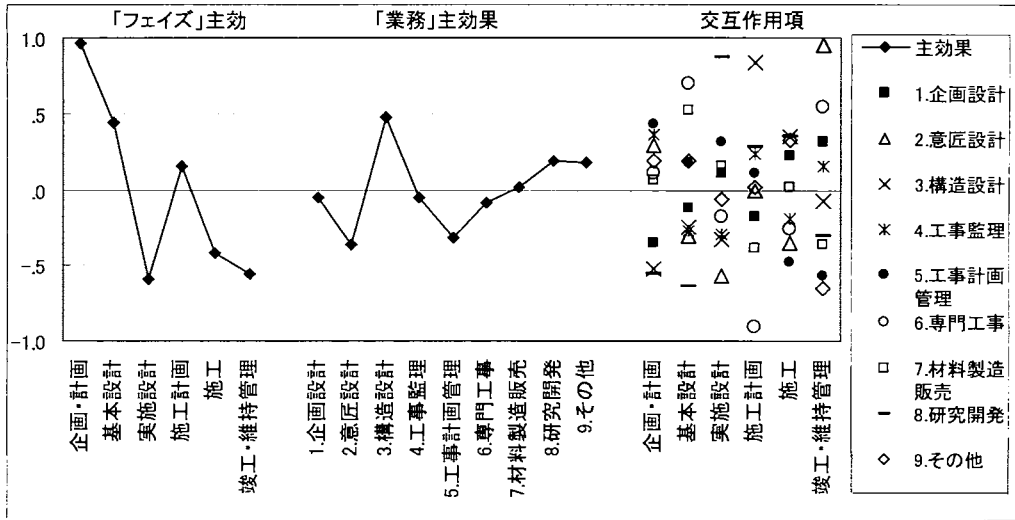


図5 ニーズ提示頻度についての順序ロジスティック回帰(グラフは要因効果)

グラフィカルモデリングの結果を信用すれば、「業務」「フェイズ」の主効果だけを解釈し、交互作用は無視してよいことになる。しかし、検定結果は有意ではないものの、回帰分析における一般的な変数選択の基準(p 値<0.2~0.25)は満たしている。また、図5によれば、中にはかなり大きい効果を持つ交互作用項もあるので、それらは解釈に加えた方がよいかもかもしれない。

しかし、ひとまず主効果を中心に解釈してみよう。この結果は、施主にとってニーズの提示が可能となるのが、設計図が具体的に固まる「実施設計」フェイズおよび、具体的に施設ができあがる「施工」「竣工・維持管理」のフェイズのように、施主が具体的な図面や実物を確認できるフェイズであることを示している。すなわち現状では施主にとって実物が提示されるまで、ニーズの伝達が困難であるということを表していると考えられる。また、ニーズの提示頻度が高い業務は意匠設計と工事施工計画管理の担当者であり、施主と接する機会が他の業務担当者に比較して多いことから納得できる結果である。

5.2 「ニーズ確定必要性」を目的変数とした分析

GMの結果によれば、「フェイズ」「提示頻度」の主効果および交互作用が説明変数の候補であるが、「業務」を取り入れたモデルも可能性があった(図6の検定結果参照)。特に、「業務」「フェイズ」と「提示頻度」の交互作用は取り入れるべきか微妙なところである。要因効果の採否によりパラメータ推定値に大きな変化はなかったので、一応、どちらの交互作用項も取り入れたモデルを採用して考察を進める。

「フェイズ」と「提示頻度」、「業務」と「提示頻度」の、主効果と交互作用を含めた要因効果を表す図を図6に示す(「業務」は9水準もあるので、交互作用のパターン別に図を2つに分けた)。

まず、主効果だけを解釈すれば、「提示頻度が多いほど必要性が高い」「企画・計画フェイズの必要性

は提示頻度の割には低く、実施設計フェイズの必要性は提示頻度の割には高く回答される」ということになろう。しかし、図6をみると、交互作用は無視できない。提示頻度が「1」の場合には、逆に確定必要性が少し高まる場合があるようである。フェイズでいえば「施工計画」と「施工実施」、業務でいえば3番目のグラフの5つの業務が該当する(設計者は全てこちらのパターンに分類されている)。この結果から、以下のような解釈が考えられる。

1つめのグラフ:フェイズ毎の分析

企画フェイズでは提示頻度に関係なく、他のフェイズと比較して施主のニーズが確定していなくてもよいと考えている傾向が見られる。企画フェイズにおけるニーズの確定に対する期待が低いことにより、プロジェクトの当初段階で要求を確定し、次フェイズ以降に伝達することを重視しないことによる問題が生じている可能性があるかと推察される。

また、ニーズの提示頻度とニーズの確定度の必要性との関連の傾向を見ると、企画・計画、実施設計、竣工・維持管理フェイズには提示頻度に応じて確定必要性が高まる右下がりの傾向が、施工計画、施工フェイズには山形(提示頻度1の確定必要性が高い)となる傾向が、基本設計フェイズにはその中間の傾向が見られる。これは施工計画、施工フェイズが、既に生産に入っているため先送りが出来ない性格を持つこと、また専門性が高く、施主からのニーズの提示頻度が少なくても業務を推進するための確定が必要な項目を担当者が把握しているのではないかと推察される。

2つめのグラフ:業務毎の分析(右下がり)

工事計画管理、材料製造販売、研究開発担当においては、ニーズの提示頻度が高いフェイズにおいてニーズが確定する必要があると考えられている傾向が見られる。この傾向は「施主の要望頻度=確定必要度」と認識していることを表していると推察される。

3つめのグラフ:業務毎の分析(山形)

企画・計画、意匠設計、構造設計、工事監理、専門工事担当者においては、ニーズの提示頻度に関わらず、確定する必要があると考えている項目が存在すること、また、提示頻度1の確定必要性が特に高いことは、「施主が認識していない確定が必要なもの」が数多くあると考えていることが見て取れる。これは建設プロジェクトの推進にあたり、施主より高い専門性を有してい

要因	df	Waldカイ2乗	p値
業務	8	10.3253984	0.2429
フェイズ	5	79.7031087	0.0000
提示頻度	3	53.0424441	0.0000
業務*提示頻度	24	29.5285108	0.2009
フェイズ*提示頻度	15	17.9030697	0.2678

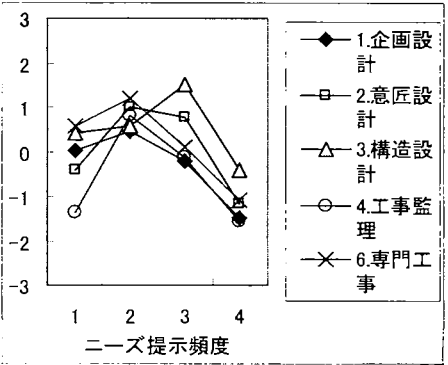
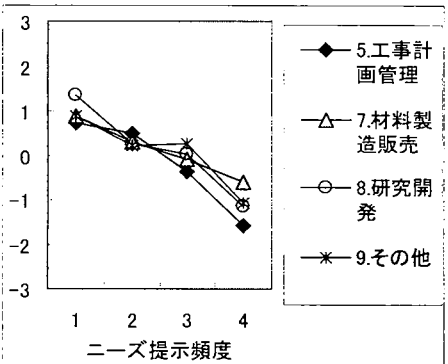
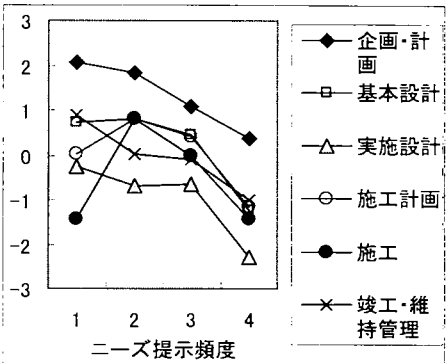


図6 ニーズ確定必要性についての順序ロジスティック回帰(グラフは要因効果)

ることの意識の現れとも考えられる。

5.3 「重要度」を目的変数とした分析

GMの結果によれば、「フェイズ」「確定必要性」の主効果および交互作用が説明変数の候補である。念のため「業務」なども取り入れたモデルも検討したが、変数選択の結果、「フェイズ」「確定必要性」の主効果のみのモデルを採用した。交互作用の効果はごく小さなものであったので、モデルから除外している。要因効果および検定結果を図7に示す。

なお、「提示頻度」は分析に用いないので、これが欠測値となるサンプルも用いている(その他の欠測値もあるので、N=1080となっている)。

この結果から以下のような解釈ができる。

全体としては、ニーズ確定が必要なフェイズほど重要度が高いとされている。このことから、耐久性確保のためにニーズが確定していることは重要な要因の1つであると考えられていることがわかる。

一方、フェイズの主効果より、耐久性確保のために重要なフェイズは、実施設計、施工、維持管理フェイズであり、企画・計画、基本設計、施工計画フェイズはあまり重要視されていないようである。一般的なプロジェクト管理においては「目的の明確化」は重視されるべきであるから、企画・計画フェイズの重要度が最も低いのは意外な結果(あまり芳しくない結果)である。

要因	df	Waldカイ2乗	p値
フェイズ	5	61.4696572	0.0000
確定必要	2	161.049156	0.0000

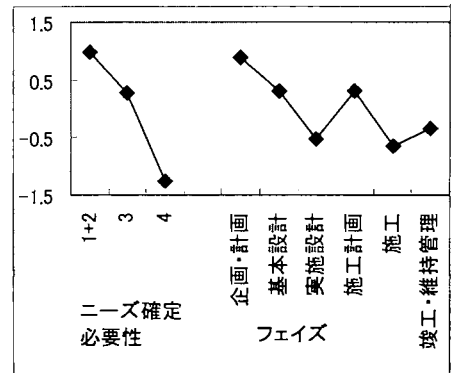


図7 重要度についての順序ロジスティック回帰(グラフは要因効果)

6. まとめと考察

ここまでの分析の主要な結果を図8にまとめる。

各段階の分析結果からは、次のような「現状の問題」が読みとれる。

「ニーズ提示頻度」を目的変数とした順序ロジスティック回帰分析より

施主・発注者がニーズを具体的に提示できるのは実施設計ならび施工、竣工・維持管理段階であることが明らかとなった。初期段階においては「プロジェクトの目的」が明確にされていない可能性がある。

「ニーズ確定必要性」を目的変数とした順序ロジスティック回帰分析より

実務担当者は企画フェイズにおけるニーズの確定を求めている傾向が明らかになった。これは上記提示頻度の考察とあわせると、「ニーズを提示できない施主」と「ニーズを重視しない実務担当者」が少なからず存在するようである。実務担当者は施主よりも確定することが必要な内容を把握しており、(相談や提示の有無は不明だが)施主になり代わり確定していることも多いのではないかと推察される。

「重要度」を目的変数とした順序ロジスティック回帰分析より

実務担当者はニーズの確定必要性が高いほど、耐久性確保のための取り組みの重要性も高くなると考えている一方で、企画・計画フェイズにおける取り組みをさほど重要と見なしていない。これはプロジェクトの方向性を決定する主導権の所在が明確でないと考えられ、この部分に問題がある。

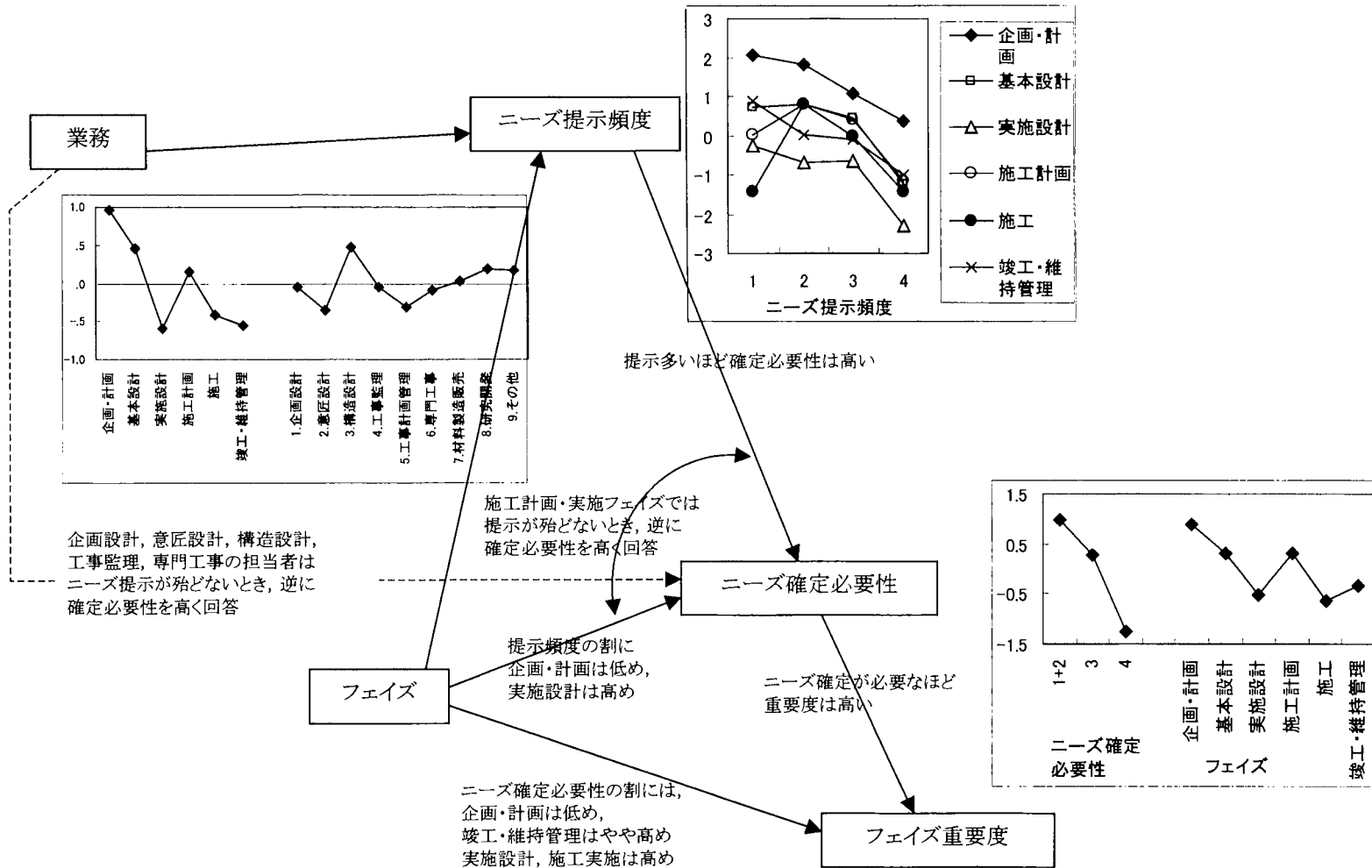


図8 主要な結果のまとめ

上記の問題点をまとめると、次のようになる。

- ・ 施主のニーズがPJの初期段階(企画・計画～基本設計)においてはあまり表現されない。
- ・ 一方、ニーズを確定することは「実務担当者」にとっては重要であるため、施主からのニーズ提示がない状態で、実務担当者の裁量により意思決定を行うことが少なくない。

これは専門家への「お任せ」型の業務形態といえ、ニーズを反映した設計目標を示すことが現状では困難なことを示している。本調査の分析結果から、今後の研究開発における課題として、

1. 施主によるニーズの確定への支援
2. 実務担当者による施主のニーズの確定への支援

という2点が設定され、解決に向けての検討を開始することとなった。

7. 分析手法に関する検討事項

最後に、本報の分析について再考したい。主要な論点は以下の通り。

- ・ GM では「業務－確定必要」間が切れたモデルを採用したが、後のロジスティック回帰では復活させている。また、そのロジスティック回帰においても、要因効果のパターンをみると交互作用の存在は明らかであるのに、交互作用の要因効果が有意でない。これは、実質的には交互作用パターンは2群に分ければ十分であるのに、業務のカテゴリ数が9水準もあることに起因している。実質的な自由度に比べて数倍の自由度が課せられ、p 値が大きめの値になってしまうのである。水準数が多い場合は、p 値や F 値だけを見て(要因効果を見ずに)変数選択するのはまずいということになる。また、質的変数の GM においては事態はさらに深刻である。GM は、その結果から「どんな要因効果を解釈すべきか」を読みとるものであるから、要因効果を見ないと GM が機能しないようでは本末転倒なのである。
- ・ GM で「業務－確定必要」間を切らないとすれば、「提示頻度←確定必要」の因果順序も可能となる。このモデルにも魅力があり、「一部の業務・フェイズにて、ニーズ確定必要性が高い場合、施主にニーズ提示を求める群と、施主には求めず自分で決めてしまう群の2極分解が激しくなる」と解釈できる結果となる。ただしこの関係は目的変数の水準を全体的に上げる(下げる)という「平均」に対する効果ではなく、中間の水準より両端の水準が多く(少なく)なるという「分散」に対する効果なので、順序ロジスティック回帰では表現できず、名義ロジスティック回帰を行う必要がある。名義ロジスティックは順序ロジスティックに比べてパラメータ数が多くなる分、推定値が不安定になりやすい。要因効果の出力も複雑で、見やすくしにくい。かといって、単なるクロス集計からでは本報の分析結果ほど明確には要因効果を読みとれない。そこで結局、このモデルは採用しなかったのである。

注釈・参考文献

- 1) 本報の調査は、独立行政法人建築研究所の研究課題「耐久性能評価に基づく建築部材仕様選定システムのプロトタイプ開発」および国土交通省国土技術政策総合研究所の研究課題「耐久性能に関する要求レベル対応型の建築部材仕様選定システムの開発」の一環として行われたものである。
- 2) 日本品質管理学会テクノメトリクス研究会編:グラフィカルモデリングの実際, 日科技連, 1999

JMPによるワインの顧客価値分析

林俊克¹⁾ 平野広隆²⁾

¹⁾(株)資生堂製品開発本部 ²⁾(株)アーキテクト

Analysis of Wine's Customer Value by JMP

Toshikatsu Hayashi¹⁾ and Hirotaka Hirano²⁾

¹⁾Shiseido Product Development Div. and ²⁾Architect

要 旨

2003年3月、首都圏に居住するワインを飲む女性200名(20歳～59歳)を対象に、ワインに関する2種類の平易なフリーワードによるアンケート調査(非定形自由文形式群100名、定形自由文形式群100名の計200名)を行い、結果をJMPを用いて価値ポートフォリオ、CSポートフォリオ、価値認識構造図に可視化し、価値意識と価値認識の構造を分析した。その結果、女性にとってのワインの顧客価値とその認識構造が把握でき、国産ワインのあるべき方向を提案することが出来た。JMPのテキストマイニングへの応用の可能性が示唆されたものと考えられる。

キーワード： テキストマイニング、JMP、茶釜、顧客価値

はじめに

JMPは、諸データから「素早く、労少なく、役に立つ情報を抽出し可視化する」というデータマイニングの目的を高いレベルで達成した、非常に便利なツールである。本報では、ワインの顧客価値の分析をテーマに、JMPをテキストマイニングの手法を用いたマーケティングリサーチ(ニーズの抽出・価値意識の解明)に応用した事例を報告する。諸解析は主としてJMP(V5)を用い、テキストマイニングに際しての形態素解析は奈良先端科学技術大学院大学 情報科学研究科が提供する「茶釜」(公式ホームページ <http://chasen.aist-nara.ac.jp> 参照)を使用した。茶釜の出力データをJMPに取り込む際の前処理(分析に供するワードの絞り込み等)はExcelを使用した。もちろん、JMPのみでも茶釜出力データの前処理は可能であり、小島らが「JMPによる統計解析入門(2002/12、オーム社)」で詳述している。

方法

2003年03月09日～10日、東京30Km圏に居住する自宅で週1回以上ワインを飲む女性200名(20歳～59歳)を対象に、ワインに関する2種類の平易なフリーワードによるアンケート調査(非定形自由文形式群100名、定形自由文形式群100名の計200名)を行った。

1. 非定形自由文形式群のアンケート

非定形自由文形式群では、

- (1) 今市場で売っているワインとは？
- (2) 理想のワインとは？
- (3) 国産ワインとは？

との設問に対して、思いつくままにその定義をフリーワードのショートセンテンスで記入してもらった。一種の投影法で演者らはこのアンケート手法を「定義法」と呼ぶ。次に、定義したそれぞれの内容に関して、「少し詳しく文章にしてください」と非定形の自由文で説明をもらった。

2. 定形自由文形式群のアンケート

一方、定形自由文形式群では、非定形自由文形式の時と同様、まず

- (1) 今市場で売っているワインとは？
- (2) 理想のワインとは？
- (3) 国産ワインとは？

との問いに対して、思いつくままに定義をフリーワードで記入してもらい、次に定義したそれぞれの内容に関して、「少し詳しく文章にしてください」とお願いする際「○○なので□□だから△△だ」のフォーマットに従った定形の自由文で説明をもらった。演者らはこのアンケート手法を文章完成法と呼ぶ。

3. アンケート結果の処理

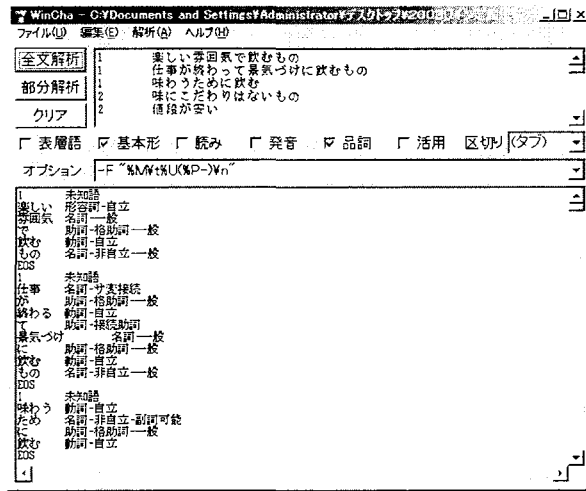
得られたフリーワード(今市場で売っているワイン、理想のワイン、国産ワインそれぞれの定義と定形、非定形のそれらを詳しく記述した文の計12コーパス)は茶筌(茶筌 version 2.1 for Windows)により、形態素解析した。

その際、定形の詳述文は、「なので」「だから」の前のワードを原因、後のワードを結果として、因果を維持しながら形態素解析を行った。

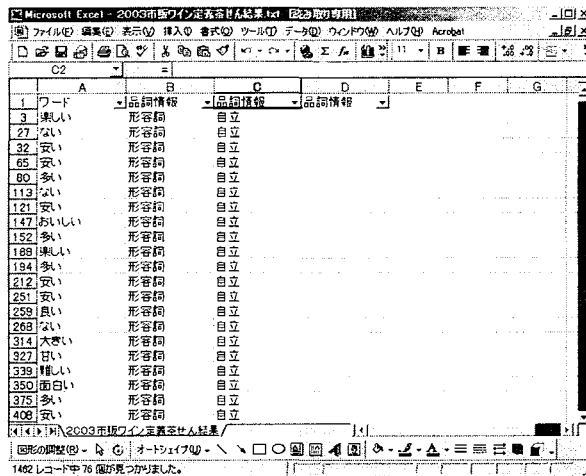
因果を維持しながら茶筌で形態素解析を行う方法については、演者が「Excel で学ぶテキストマイニング入門(2002/10、オーム社)」で詳述している。

非定形詳述文は、文中の接続助詞「ので」「から」「ば」の前のワードを原因、後のワードを結果として、因果を維持しながら形態素解析を行った。

辞書は標準のまま用いユーザー定義語は設けず、出力オプションは基本形と品詞とした。



形態素解析の後には、Excel のフィルター機能を用いて品詞情報を基に分析に必要なワード(基本形)を選別し、分析用データを作成した。本報では、非自立語、フィラー、記号、助詞、助動詞、接頭詞、副詞、連体詞を除外し、主として自立の形容詞、動詞および名詞、未知語(ex. ロゼ)で構成されるワードを分析用データとして採用した。



分析用データは、JMPを用いて価値意識分析、価値認識構造分析を行った。

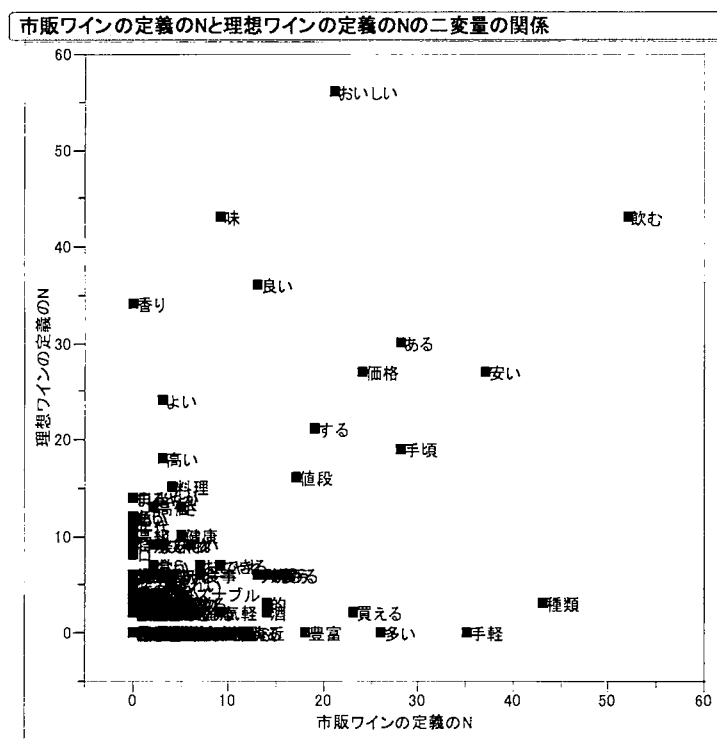
価値意識の分析は、[分析]>[2変量の関係]を用いて、市販のワインの定義に出現するワードの出現度数をX、説明変数、理想のワインの定義に出現するワードの出現度数をY、目的変数として散布図を表示することで行った。

価値認識構造分析は、[グラフ]>[特性要因図]を用いて、原因ワードを子、X、結果ワードを親、Yとして特性要因図を表示することで行った。

結果

1. ワインの価値意識の分析

価値意識の分析に供するアンケートは非定形自由文形式群100名と定形自由文形式群100名の両群全く同一であるので、両群のデータをプールし、ワインの価値意識を分析した結果が次の図である。演者らはこの散布図を価値ポートフォリオと呼ぶ。



価値ポートフォリオの対角線上に位置するワードは、市販のワインに既にある属性であり、かつ同じ程度に理想でも望まれている属性であるから、今市場にあるワインが顧客(調査対象)に対してちょうどピッタリの満足を与えている価値であると考えられるが、ここでは、

飲む、安い、手頃、価格、値段

といったワードが該当しており、市販のワインが価格において適正な価値を提供できていることがわかる。

対角線の下側(価値ポートフォリオの右下)は、「理想<市販」の領域であり、市販のワインでその属性がよく認識されているが、理想ではさほど求められない、即ち顧客(調査対象)にとって「既に満たされている」価値であると分析できるが、ここでは、

種類、手軽、買える、多い、豊富

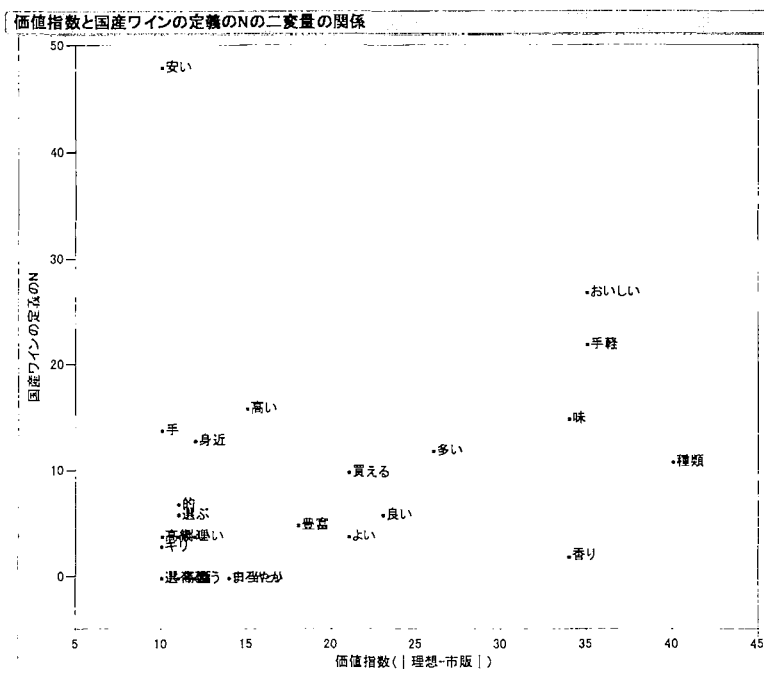
といったワードが該当している。つまり、ワインは手軽に買えること、種類が多いことは顧客にとっては、既に当たり前価値で、更にワインの品種を増やしてもあまり有り難がってくれない可能性が高いと考えられる。

逆に対角線の上側(価値ポートフォリオの左上)は、「理想>市販」の領域であり、市販のワインではその属性はあまり認識されていないが、理想では多く求められる、即ち顧客にとって「未だ満たされていない」価値即ち顧客の潜在ニーズであると分析できるが、

おいしい、味、良い、香り、よい、高い

といったワードがそれに該当している。つまり、味は当然として、香りの良いワインや高いワインが市場で求められていると考えられる。

また、|理想-市販| (理想のワインで定義されたワードの度数と市販のワインで定義されたワードの度数の差の絶対値)を「価値指数」として横軸にとり、縦軸に国産ワインの定義の度数をプロットしてCSポートフォリオを描くと、次図のようになる。

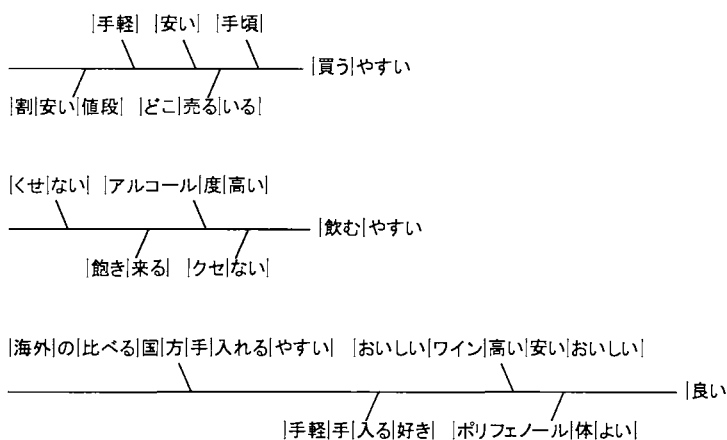


理想との乖離が大きく、市場のワインがまだ十分に顧客に満足を与えていない価値属性(種類が多すぎる、おいしい、手軽、味、香り等)のうち、香りの評価が非常にいいことから、国産ワインが緊急に改善しなければならない課題が香りの良さにあることがわかる。

このように、JMPの散布図描画機能を用いて価値ポートフォリオを作図することで、視覚的に、顧客の潜在的な価値意識を理解することができる。Excel を用いても同様な分析は可能であるが、散布図に布置されたポイントのラベル(ワード)を表示する機能が提供されておらず、手作業のラベル付けが必要であり、作業効率が格段居に違う。

2. ワインの価値認識構造分析

ワインの価値認識構造の分析は、非定形自由文形式と定形自由文形式のそれぞれが抽出する構造の比較を行うため、独立に解析した。非定形自由文形式群の解析結果を次図に示す。演者らはこの特性要因図を価値認識構造図と呼ぶ。価値認識構造図は、結果ワードの上位3ワード(買いやすい、飲みやすい、良い)について示す。



顧客が買いやすいと考えるワインは、

価格が安い或いは手頃で手軽に入手できる

ものであり、飲みやすいと考えるワインは、

味にクセがなく、アルコール度が高くなく、飽きがこない

ものであり、良いワインは、

手に入りやすいという意味で国産、(おいしいワインは高いから)安くておいしいもの、ポリフェノール等体によさそう

なものであると推察出来る。

次に、定形自由文形式群の解析結果を次図に示す。価値認識構造図は、結果ワードの上位3ワード(買う、飲む、おいしい)について示す。



顧客は、ワインを買うに際しては、

価格(安い、手頃、高価)と味(自分の好み、おいしい)と買い易さ(どこでも売っている、手軽)

を考慮しており、飲むに際しては、

状況(普段、特別)と味(甘口、まろやか、飲みやすい、おいしい)

を考慮しており、おいしさは、

(料理に)合わせて飲む

ことを考慮している様子が伺える。

ここに示した価値認識構造図は、JMPが出力する特性要因図のごく一部であるが、全てを詳細に検討することにより、さらに多様な知見を発見することが出来る。

Excelのピボットテーブルを用いることで、原因と結果の対応を分析することは可能であるが、クロス集計表を視覚化する機能が提供されていないので、解釈が非常に困難であるが、JMPの特性要因図は、因果関係を直感的に理解できる点で優れていると考える。

考察

平易なフリーワードのアンケートを実施し、JMPによって価値ポートフォリオ、価値認識構造図に可視化したワインの顧客価値を総合的に咀嚼すると、女性にとってのワインの価値は、

香りの良さ

まろやかな口当たりの良さ
料理に合うことによるおいしさ
価格の安さ、入手の手軽さ

が重要なものであるが、一方で

価格の高いワイン
への欲求も見られることが理解された。その他、
コクがある、後味が良い、色がきれい、ボトルのデザインが良い
等も重要であった。

国産ワインの評価に関しては、味や安心感には全く問題がないが、高価で人に出せるような外国産のワインを理想とする考えが強いため、どうしても一般的で安いイメージを払拭し切れていないのが実状で、喫緊の課題は香りの改善であった。

また、今回非定形自由文形式と定形自由文形式の2つのアンケート手法を比較検討したが、非定形自由文形式は思ったこと、感じたことがさほどストレスなく表現できるので、アンケート記入時の負担が低い反面、論理的な因果分析が困難で、接続助詞「ので」「から」「ば」の前のワードを原因、後のワードを結果とみなすといった、荒っぽい割り切りを必要とした。一方、定形自由文形式は、アンケート記入の際の負担感が大きい反面、論理的な因果関係を正しく分析できる利点をもっていた。

通常、両手法によって得られる知見には若干の差が生じることが多いが、今回は比較的良好一致を見たが、前段で両手法に共通して設けた定義形式のアンケートで、予め着眼点を固定し、書きやすい、書きにくいに関わらず必ずその内容をやや詳細に記述させる工夫が功を奏した為と考えられる。

今回の調査結果から、女性のワイン価値意識に上手く合致するワインとして、以下のようなことが提案できると考えられ、JMPのテキストマイニングへの応用の可能性が示唆されたものと考ええる。

- ① 女性にとってのワインの価値を真っ向から攻めるとすると、香りが良くまろやかな口当たりのワインを開発すべきである。
- ② 「ザ・国産ワイン」と言えるようなシンボリックなブランドの創出も手。
- ③ 価格を日常用ギリギリのやや高めの価格に設定し、人に出して恥ずかしくないものにすべきである。

参考文献

林 俊克. :Excel で学ぶテキストマイニング入門(オーム社) 2002

- 田久 浩志, 林 俊克, 小島 隆矢. :JMPによる統計解析入門(オーム社) 2002
- 朝野 熙彦. :魅力工学の実践(海文堂) 2001
- 林 俊克, 平野 広隆. :VACAS&DIONISOS が解明する女性にとってのワインの感性価値. 日本感性工学会第 10 回あいまいと感性研究部会研究発表会講演論文集, 11~15 頁, 2003
- 林 俊克. :テキストマイニングの現在. マーケティング・リサーチャー94 号, 16~25 頁, 2003
- 林 俊克. :VACAS による感性商品開発(ファンデーションの開発事例). 感性工学 第 2 巻 1 号・通巻 006 号, 25~27 頁, 2002
- 林 俊克. :感性工学的手法によるファンデーションの商品開発. 日本感性工学会感性商品部会報 第 1 号, 23~33 頁, 2002
- 町田 明子, 林 俊克. :ネット上書き込み情報のテキストマイニング. 第 4 回日本感性工学会大会予稿集 2002, 249 頁, 2002
- 町田 明子, 林 俊克. :ネット上書き込み情報のテキストマイニング. 日本行動計量学会第 30 会大会発表論文抄録集, 98~99 頁, 2002
- 林 俊克. :感性工学と化粧品開発. FRAGRANCE JOURNAL, 29 巻第 4 号通巻 246 号, 46~51 頁, 2001
- 林 俊克, 道官 克一郎, 平野 宏隆:ワインの顧客価値に関する研究. 第 3 回日本感性工学会大会予稿集 2001, 149 頁, 2001
- 林 俊克, 真柳 真譽美,, 平野 宏隆:女子大生の魅力的牛乳像の解明. 日本行動計量学会第 29 会大会発表論文抄録集, 104~105 頁, 2001
- 林 俊克, 田久 浩志, 道官 克一郎, 平野 宏隆:デマテルによる看護婦の化粧意識の解析. 第 2 回日本感性工学会大会予稿集 2000, 58 頁, 2000
- 林 俊克, 田久 浩志, 道官 克一郎, 平野 宏隆:デマテルによる看護婦の化粧意識の解析. 日本行動計量学会第 28 会大会発表論文抄録集, 113~114 頁, 2000
- Dohkan,K., Hayashi,T., Masuda,M. and Fukuchi,Y. :An Application of Kansei Engineering to a Cosmetic Product- Application of Dohkan Method-, Kansei Engineering II ?Human sensibility ergonomics-, Edited by Soon Yo Lee, p.13-22, Ingankyngyungsa, 1999

看護師のセクシャルハラスメントに対する意識について

田久浩志¹⁾ 岩本晋²⁾

1) 中部学院大学 人間福祉学部 健康福祉学科

2) NPO 福祉法人 OIEMASE

The nurse's opinion research about sexual harassment

Takyu Hiroshi Chubu Gakuin University takyu@chubu-g.ac.jp

Iwamoto Susumu Non Profit Organization OIEMASE

要旨

JMP5.01 で看護師が持つセクハラに対する認識の定量解析を行った。各種の質問、および不快な思いをしたときの看護師がとる具体的な行動と年齢の関係をロジスティック回帰で示し、医療現場の新人教育の参考にすることを提案した。

キーワード: セクシャルハラスメント、看護師、JMP ソフトウェア

【はじめに】

最近、TVや新聞などでセクシャルハラスメント(以下セクハラと略)の話題をよく耳にするようになった。中には、故意にセクハラ行為しているケースもあるだろうが、自分自身にとっての普通の行動が知らない内に相手にとって不愉快な行為をしているケースも考えられる。従来、筆者らは学生におけるセクハラを意識調査^{1,2)}をしてきたが、女性の年齢が増加するにつれてどのように意見が変化するかは定かではない。今回、学生とほぼ同じ居住地の看護師を対象に意識調査を行った。看護師独自の意見の差が存在するか否かは不明ではあるが、看護師を社会人女性の一例と考え解析を行ったので報告する。

【対象と方法】

対象は岐阜県S市の女性看護師191名、コントロールは同じくS市のC学院大学女子学生457名である。対象の居住地域がほぼ同一なので地域による意識差はないと仮定した。セクハラを想定する場面として露骨なお誘いの場面でなく、1:通常の学校や職場での生活で少し離れた間柄の人との対応、もしくは、2:食事の席などで同席した初対面の人と対応する場面と考えた。

フェース項目として性別、独身既婚、年齢、ファッション、化粧に凝るか、異性の目を気にするか、男性に厳しいか、男女の兄弟の状況、喫煙の有無、過去一年のイッキ飲みの有無などを質問した。

セクハラに関する質問項目として11項目の質問(表1)を「別に感じない(1点)」「あまり不快でない(2点)」「やや不快(3点)」「極めて不快(4点)」の4段階で評価した。得られた点数は合計をして「拒絶度」と定義し、セクハラ行為に対して寛容か厳しいかの指標とした。また、女性が不快な思いをしたとき、どのような具体的な行動をとるか(1:忘れる、2:自分の中にとどめる、3:仲間を巻き込み悪い風評をながす)を調べた。統計解析にはSAS社のJMPVer5.01を使用した。

【結果と考察】

解析では年代:1 18-22の女子学生と、23歳以上の看護師の意見を比較した。以下の記述で「看護師」と示す場合は23歳以上の看護師を意味している。

1.フェース項目について

看護師の全年代をまとめ、化粧・髪型にこる、ファッション、男性・女性の兄弟の有無、などの区分で拒絶度の平均値の差をt検定で検討したところ、化粧・髪型のみ 5%の危険率で有意差の低下が見られた。また女子学生に比較して、看護師では拒絶度の上昇傾向が見られた(図1)。

2.セクシャルハラスメントの質問項目について

女子大生と看護師で拒絶度の変化を求めたところ、全般的に看護師の方の拒絶度が増加していた。少数ではあるが、女子学生の拒絶度に 10~15 といった低値が存在したが、看護師ではほとんど存在せず、これが年代による拒絶度上昇の一因になっていることが示唆された。これより女子学生と看護師でセクハラに対する意識が変化することが考えられた。

個別の質問をみると、年齢を聞く、飲み会でのお酌を強要する、などの項目で、年代の上昇につれて「極めて不快」と回答する割合が増加していた(図2、3)。女子学生では、学年から年齢が明確になるため、年齢に対する質問にはあまり敏感に反応しないと考えられた。

女性の形容詞として「色っぽい、セクシー」「美人、きれい」「かわいい」の3種類を取り上げ、看護師でどのような意見を持っているかを求めた(図4)。その結果、「色っぽい、セクシー」では 60%近くが「やや不快」「極めて不快」と回答し、「美人、きれい」「かわいい」ではその逆の傾向を示した。また、「美人、きれい」「かわいい」でも 15-16%は「やや不快」「極めて不快」と回答していた。これは、褒め言葉と考えられる形容詞も使い方に留意しなくてはならないことを意味している。しかし、今回の調査は少し離れた間柄との人、もしくは初対面の人との関係を想定しているので、相手とのコミュニケーションのとり方によってはこの反応が改善されることも考えられる。

3.不快な思いをした男性への対処について

不快な思いをした男性への女性の対処に関して検討したところ、年代により対処に有意な差は見られないが、30代の看護師で「忘れる」と答えるものが増加している傾向がみられた。そこで、年代ではなく拒絶度と女性の対処法について、ロジスティック回帰で分析を行った(図3)。その結果、看護師の拒絶度の平均値である 31 点でも、不快な目にあうと、自分の中にとどめる、もしくは、悪い風評をながす人が 80%近く存在することが明らかになった。男性は女性が不快な思いをした場合に被害が自分にも及ぶ、あるいは、職場での人間関係に多大な影響が生じることを認識すべきであろう。

【まとめ】

今回の調査は少し離れた間柄との人、もしくは初対面の人との関係を想定しているという制約があり、どのような場面にも今回の結果があてはまるわけではない。また、一般の社会人女性と看護師との間で意見の相違があるかは定かではない。しかしそのような条件でも、女性の化粧や服装によってセクハラに関連する状況に対して寛容か否か(拒絶度が低いか高いか)を判断しがたいこと、女性を褒めても不快に感じる人が存在すること、何に対しても「別に感じない」という女性が存在することが明らかになった。結局のところ、男性は妙な先入観はもたずに相手に対応し、かつ、不用意な発言が場合によっては自分に災難が及ぶことを認識し行動することが、セクハラに関するトラブルを少なくするポイントと言えよう。

最後に、これらの解析結果は、保健・医療・福祉現場で働く職員にとって、よりよい職場環境をつくるための基礎資料となることを指摘したい。

【参考文献】

大学生のセクシャルハラスメントに関する意識調査: 入山貴弘, 渡辺朋恵, 田久浩志, 第3回 中部学院大学国際シンポジウム, 2002

男女学生間のセクハラ感の定量調査: 田久 浩志, 第 56 回 関西 SAS ユーザー一会, 2003

表1 質問項目

- 1 スリーサイズ、体型などを訊かれるのはどうか
- 2 恋人はいるのかと訊かれるのはどうか
- 3 年齢を訊かれるのはどうか
- 4 飲み会で「つげ」とお酌を強要されるのはどうか
- 5 女のクセに…という発言はどうか
- 6 髪の毛、肩、腰など体を触られるのはどうか
- 7 色っぽい、セクシーと言われるのはどうか
- 8 美人、きれいと言われるのはどうか
- 9 かわいいと言われるのはどうか
- 10 必要もないのに個人的な性体験を尋ねられ
- 11 男性が他の女性の身体、服装や性的な関係などを他の人がいるところで話題にする

図1 年代と拒絶度

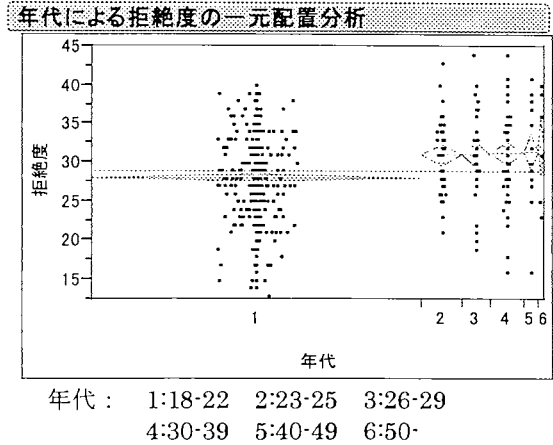


図2 年代と年齢を聞く

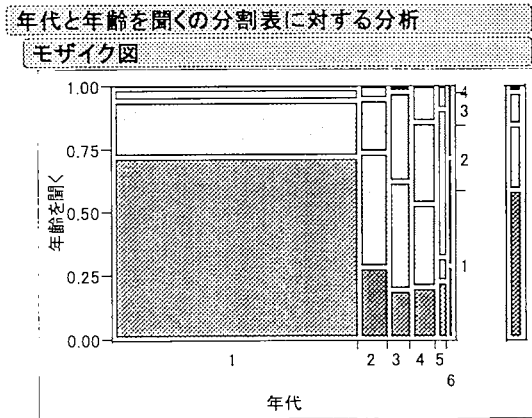


図3 年代とお酌の強要

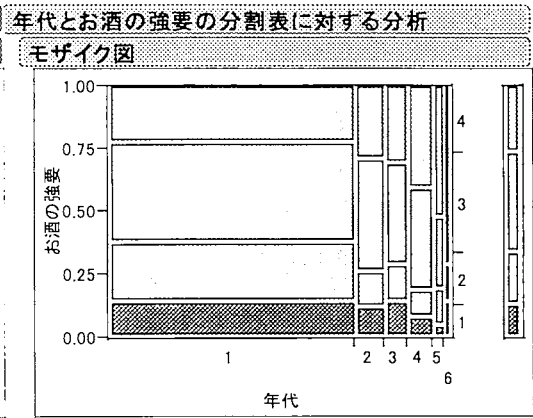


図4 褒め言葉に対する反応

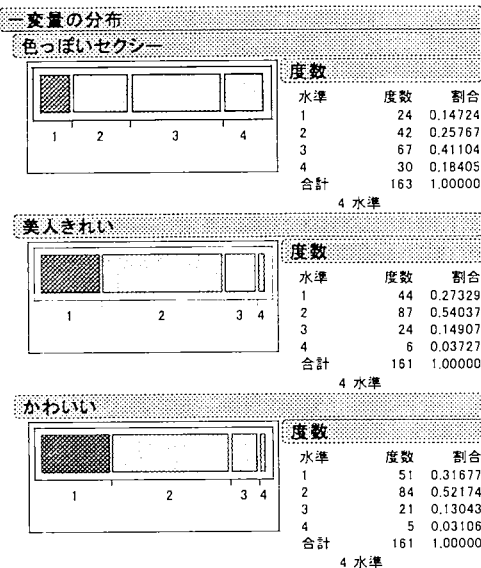
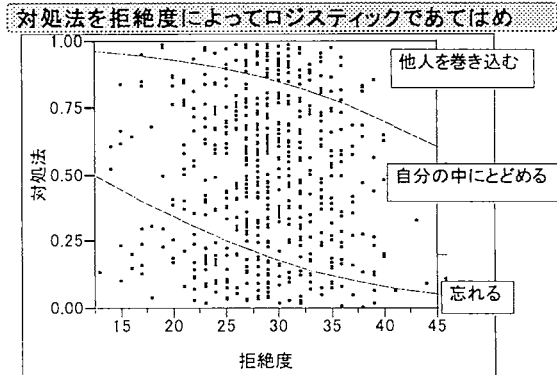


図5 拒絶度と男性への対処法



日本SASユーザー会 (SUGI-J)

Life Time Value を基準とした施策の最適化方法 — 遺伝的アルゴリズムによる解析事例 —

○小谷田 知行, 堀 彰男
株式会社 浜銀総合研究所
戦略研究部

Optimization methods for plan using Lifetime Value
An application of Genetic Algorithms to marketing

Tomoyuki Koyata, Akio Hori
Hamagin Research Institute, Ltd.
Dept. of Strategic Management & Consulting

要 旨

銀行は、顧客を維持・獲得・成長させるため、数々の施策を行っている。施策を行うと顧客生涯価値(以下、LTV)は変化するため、この変化分、あるいは変化後の値によって施策を評価することができる。この際、経営者が知りたいことは、個々の施策の評価だけではなく、銀行全体としての最適な施策の組み合わせである。しかしながら、最適な施策を求めるには、施策の膨大な組み合わせを試す必要がある。そこでまず、LTVの精度を把握するためにその分布を算出した。次に、遺伝的アルゴリズム(以下、GA)を使用して最適な施策の発見を試みた。

本報告では、計算方法を概説し、適用した事例について紹介する。

キーワード： LTV, 最適化, 遺伝的アルゴリズム, DATA ステップ

1.はじめに

銀行が現在得ている収益は、過去の戦略・施策の結果蓄積されたストックから生まれているものが主である。預貸のスプレッドから得られる資金収益の多くは、過去の営業活動の結果得られていることが代表的な例であろう。つまり、銀行にとって、現在の顧客は、今期の収益だけでなく、将来の収益を生み出す源泉となっている。

そこで、顧客を現在の短期間の収益ではなく、将来得られるキャッシュフローの現在価値で評価する指標であるLTVが必要となった。LTVは顧客の価値を評価する一つの指標であるが、顧客に対して施策を行った場合、施策の前後でLTVは変化する可能性がある。つまり、施策によってワレット(財布)・シェアが高まり、その結果LTVも変化するのである。そのため、施策を行ったことによるLTVの変化量を使用して、施策を評価することができる。施策の評価にLTVを使用することによって、将来得られると期待される収益を考慮した施策を検討することが可能となる。

銀行に限らず、経営判断は、限られた資源をどのように配分するかということが問題となる。LTVを経営判断に使用するためには、LTVの総和をコントロールできることが必要であり、そのためには、

LTV の精度を求める必要がある。施策については、限られた資源から複数の施策を行い、その費用対効果の合計が最大となることが望ましい。ただし、最大値を求めるためには、考えられる施策について、その対象や時期などの組み合わせる必要があり、膨大な組み合わせ数となる。そこで、遺伝的アルゴリズムを使用して、最適解を求めることを試みた。

本報告では、まず、従来の LTV 算出の問題点について整理し、つぎに本報告での LTV と GA の計算方法の概説と株式会社横浜銀行殿から提供していただいたデータを適用した事例について紹介する。

2.LTV の算出

2.1.一般的な LTV 算出方法

LTV は、将来得られると期待されるキャッシュフローの現在価値である。そのためには、顧客の継続率や収益額が必要であり、それらの将来の推定値として過去の収益額や継続率の実績値を使用する。しかしながら、銀行は、長期間の顧客データを分析に使用できる形で保存していない。そのため、年齢の異なる複数の顧客の値をつなぎ合わせて、長期間の LTV を推定する方法が用いられることが多い。

式 1 は、よく使用される LTV の算出方法である。LTV 算出の際は、状態を複数定義する。

例えば、口座無、給振無(口座有)、給振有、の 3 状態を定義する。すると、「年齢 30 歳で給振有顧客のその後 10 年間の LTV」や「年齢 30 歳で給振無(口座有)顧客のその後 10 年間の LTV」を求める事ができる。

この際、状態遷移確率行列 $Q(l, l+1)$ と収益行列 $P(l, l+1)$ は、年齢 l 歳から $l+1$ 歳の実際の平均値(あるいはそれから推定した値)を使用することとなる。

$$LTV(h, i) = \sum_{l=h}^{i-1} [\{d(l-h)Q(h, l)(Q(l, l+1)\#P(l, l+1))\} \mathbf{1}] \quad < \text{式1} >$$

$LTV(h, i)$: 年齢 h 歳から i 歳までの LTV ベクトル

$$LTV(h, i) = \begin{cases} LTV(h, i, 1) \\ M \\ LTV(h, i, n) \end{cases}$$

$LTV(h, i, a)$: 年齢 h 歳状態 a の顧客の年齢 i 歳までの LTV

$d(j)$: j 年後までの割引率

$\#$: 行列の要素毎の積

$\mathbf{1}$: 要素がすべて 1 の列ベクトル

$P(h,i)$: 年齢 h 歳から年齢 i 歳の収益行列

$$P(h,i) = \begin{pmatrix} P(h,i;1,1) & \Lambda & P(h,i;1,n) \\ M & O & M \\ P(h,i;n,1) & \Lambda & P(h,i;n,n) \end{pmatrix}$$

$P(h,i;a,b)$: 年齢 h 歳状態 a の顧客が年齢 i 歳に状態 b となる時の収益額($i > h$)

$$Q(h,i) = \begin{cases} \prod_{l=h}^{i-1} Q(l,l+1) & (h > i) \\ I & (h = i) \end{cases} \quad (I: \text{単位行列})$$

$Q(h,i)$: 年齢 h 歳から年齢 i 歳の遷移確率行列

$$Q(h,i) = \begin{pmatrix} Q(h,i;1,1) & \Lambda & Q(h,i;1,n) \\ M & O & M \\ Q(h,i;n,1) & \Lambda & Q(h,i;n,n) \end{pmatrix}$$

$Q(h,i;a,b)$: 年齢 h 歳状態 a の顧客が年齢 i 歳に状態 b となる遷移確率($i > h$)

$$\left(\sum_{k=1}^n Q(h,i;a,k) = 1 \right)$$

2.2.LTV 算出方法に関する問題

LTV の算出に関する問題は、3 つに整理できる。

(1)算出のためのデータについて

式 1 で問題となるのは、遷移確率行列 $Q(l,l+1)$ と収益行列 $P(l,l+1)$ を、年齢 l 歳から $l+1$ 歳の実際の値を使用することである。例えば、この場合、今 20 歳の顧客が 30 歳になったときに、今の 30 歳と同じ取引行動を行うとは限らないため、遷移確率と収益額が同じとは限らないという問題がある。

(2)顧客満足とLTV の関係について

LTV を算出するためには、過去のデータを使用する。すると使用した時点で収益額と継続率の高い顧客の LTV が高くなる。例えば、現在の銀行でこのことを行うと、住宅ローンのある顧客の LTV が高いという結果になることが多い。しかし、その顧客は、自ら進んでその銀行と住宅ローン取引を行ったとは限らない。また、現在もその銀行に満足しているかわからない。このような場合でも、住宅ローンのある顧客の LTV は高いという結果となる。つまり、顧客の満足と LTV の値は、必ずしも関係はない。

(3)LTV の精度について

式 1 で説明したように継続率と収益額を計算し、それを現在価値に割り引くことで LTV を求めることができる。これは、上記 2 つの問題を抱えているとはいえ、LTV を算出するための現実的な方法の

ひとつであろう。しかしながら、LTVを使用する目的は、顧客や施策を評価することが多い。そのため、LTVの精度が分からないと、評価が難しい。

実際、銀行における特定顧客群の年間収益額の分布は、図1の様になっている。よく言われるように一部の顧客が収益の多くをもたらしていることがわかる。さらに条件を追加して顧客を限定しても、程度の差はあるものの、傾向は同じである。このように分布が偏っている変数の平均値を使用してLTVを算出しても、LTVの分布がわからないため、その結果の評価が難しい。

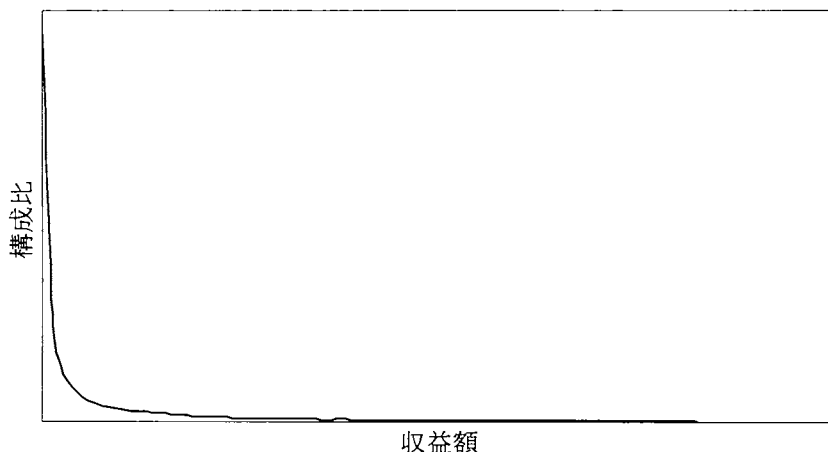


図1:収益額の分布

本報告では、LTV算出方法における問題点「(3)LTVの精度について」を、モンテカルロ法を使用したシミュレーションによってLTVの分布を求めることで確認する。

2.3.LTV計算方法

本報告において、LTVのモンテカルロ・シミュレーションは、式1における状態遷移確率行列 $Q(l, l+1)$ と収益行列 $P(l, l+1)$ を平均値ではなく、ある分布に従う乱数によって決定した。顧客の状態は、17(内1つは口座無)に分類した。

計算は、(1)状態遷移の決定、(2)収益額の決定、の手順を10,000回行った。つぎに状態遷移の決定方法と収益額の決定方法について述べる。ただし、いずれの場合も乱数は、別途用意したFORTRANプログラムによって生成し、それ以外はDATAステップで処理している。

(1)状態遷移

状態遷移は、パラメータに実際の状態遷移人数を使用して、多項分布に従う乱数によって決定した(図 2)。

$N(h)$: 年齢 h 歳から年齢 $h+1$ 歳の状態遷移人数行列

$$N(h) = \begin{pmatrix} N(h;1,1) & \Lambda & N(h;1,17) \\ M & O & M \\ N(h;17,1) & \Lambda & N(h;17,17) \end{pmatrix}$$

$N(h;a,b)$: 年齢 h 歳状態 a の顧客が、年齢 $h+1$ 歳に状態 b となる実際の人数

$$N(h;a) = \sum_{k=1}^{17} N(h;a,k)$$

このとき、多項分布 $M\left(N(h;a); \frac{N(h;a,1)}{N(h;a)}, \Lambda, \frac{N(h;a,17)}{N(h;a)}\right)$ に従う乱数を発生させて、年齢 h 歳の状態 a から次の状態を決定する。

図 2: 状態遷移確率のパラメータ推定方法と乱数発生方法

(2)収益額

収益額の分布は、状態によって形状が大きく異なっている。状態によって特に差があるのは、収益額マイナスの存在の有無と収益額ゼロの顧客割合であった。つまり、取引がほとんどない状態の場合、収益ゼロや収益マイナスの顧客が多く存在する。そこで、まず、状態遷移の決定と同様に多項分布に従う乱数を発生させて、三角分布(収益<0)、収益ゼロ、対数正規分布又はパレート分布(収益>0)を決定した(図 3)。

三角分布が選択された場合は、さらに三角分布に従う乱数を発生させ、対数正規又はパレート分布が選択された場合は、各々の分布に従う乱数によって、最終的な収益額を決定した。対数正規分布とパレート分布は、事前に最尤法によってパラメータを推定し、適合度の高い分布を採用した。

$N(h;a,b;-1)$: 年齢 h 歳状態 a から年齢 $h+1$ 歳状態 b に遷移する 収益マイナスの顧客数
 $N(h;a,b;0)$ // 収益ゼロの顧客数
 $N(h;a,b;1)$ // 収益プラスの顧客数
 $N(h;a,b) = N(h;a,b;-1) + N(h;a,b;0) + N(h;a,b;1)$

このとき、多項分布 $M\left(N(h;a,b); \frac{N(h;a,b;-1)}{N(h;a,b)}, \frac{N(h;a,b;0)}{N(h;a,b)}, \frac{N(h;a,b;1)}{N(h;a,b)}\right)$ に従う乱数を発生させて、収益額がマイナス、ゼロ、プラスを決定する。

図 3: 収益額のパラメータ推定と乱数の発生方法—収益マイナス、ゼロ、プラスの決定—

(3)結果

モンテカルロ・シミュレーションを実行し、状態別に示したものが図4である。ここでは、二つの状態について表示している。

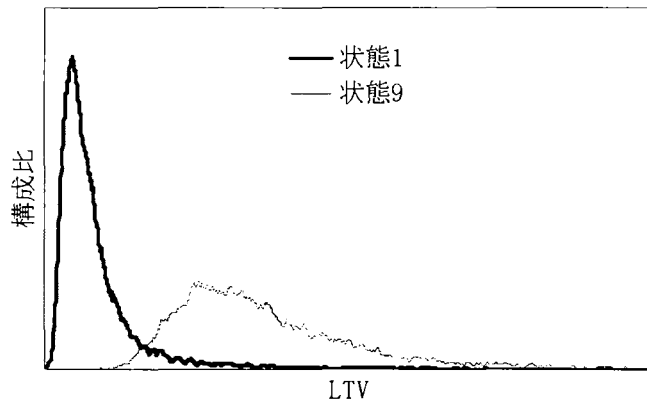


図4:状態別 LTV の分布

この例では、状態9のLTVの平均値は状態1の約4倍あり、分布を見てもその重なりは小さい。このことから、状態9のLTVは、状態1よりも高い場合が多いといえる。

もちろん、逆転する可能性もありえるし、その確率を求めることも可能である。さらに、状態9の分布は裾が広いので、平均から乖離することも多いだろう。これらの評価方法は、LTVの使用目的による。本稿では、LTVの分布を確認することにとどめておく。

3.GAによる施策の最適化

3.1.LTVによる施策の効果の考え方

あらためてLTVに影響をあたえる要因を整理すると、次の4つに分類できる。

(1)新規顧客の開拓コストと既存顧客の維持コストの関係

サービス業においては、一般的に新規顧客の開拓コストは、既存顧客の維持コストと比較して高いと言われることが多い。これについては、銀行や商品によって事なり、かつ銀行の戦略によって変化する可能性がある。

(2)ライフステージによる収益変化

顧客は、就職、結婚、出産などのイベントによって、ライフステージが変化する。これによって銀行から見た顧客の行動も変化する。この変化の仕方は顧客によって異なっており、そのことがLTVに影響をあたえる。顧客のライフステージが銀行の働きかけによって変化する可能性は低く、この要因は、銀行がコントロールすることは難しい。

(3)メイン化による収益・継続率の変化

従来から銀行では「メイン化」と称して、給与振込や公共料金の自動振替等の獲得を推進して来た。これにより、顧客における自行のワレット・シェアを高め、収益額の増加と継続率の向上を図っていた。顧客の「メイン化」は、銀行の施策によって変化する可能性が高く、それにより LTV も変化する。したがって、この要因は、施策によってコントロールできる可能性が高い。

(4)その他

その他の要因として、ロコミ効果や営業コスト等があるが、計測が難しい。

本報告では、「(3)メイン化による収益・継続率の変化」の要因に着目し、施策前後の LTV の変化分を施策の効果と考える。つまり、施策によってメイン化が進み、その後の収益・継続率が変化する。収益・継続率が変化することで LTV が変化する。

以下では、GA による最適な施策の組み合わせを求める方法と、事例として株式会社横浜銀行から提供頂いたデータとダミーデータを適用した結果を紹介する。なお、施策実施後のデータは、研究用のダミーデータであり、株式会社横浜銀行のデータではないことを断っておく。

3.2. 施策の評価と GA

個別の施策の評価方法は、前述した通り、施策前後の LTV を比較することで行う。

図 5 は、実際に計算した例である。この場合、施策後の LTV の平均値は、施策前の 50% 増となったが、分布は重なっているところが多い。これと、施策に要する費用を勘案して、施策を評価することができる。

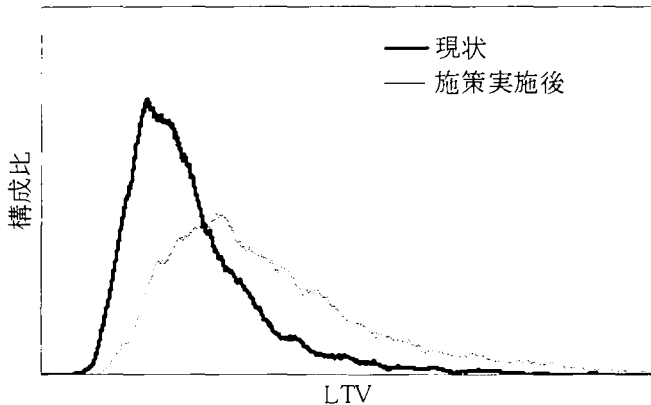


図 5: 施策前後の LTV の比較

施策の候補について図5を算出したとしても、実際にすべての施策を行うわけではない。最終的に知りたいのは、銀行全体の顧客について、LTVの合計値を最大にする施策の組み合わせである。

ただし、一つの施策について考えても、その対象は顧客の状態と性年齢によって効果が異なる。そのため、状態と性年齢別に効果を推定しなければならない。さらに、施策が複数あり、その組み合わせを考えると、施策の組み合わせ候補別に状態と性年齢の効果を推定しなければならず、組み合わせ数は膨大になる。したがって、銀行全体のLTVを最大にする施策の組み合わせを求めするためには、膨大な計算量が必要となる。そこで、この組み合わせ最適化問題を解くためにGAを使用することとした。

3.3.計算方法

(1)施策の組み合わせの表現方法

GAを使用するため、図6の様に施策ならびに施策の組み合わせを表現した。すなわち、施策の対象を性年齢と状態で特定し、施策の種類をあわせて一つの施策をバイナリ表現している。ただし、年齢については、順序に意味があるため、実行前にグレー表現に変換している。

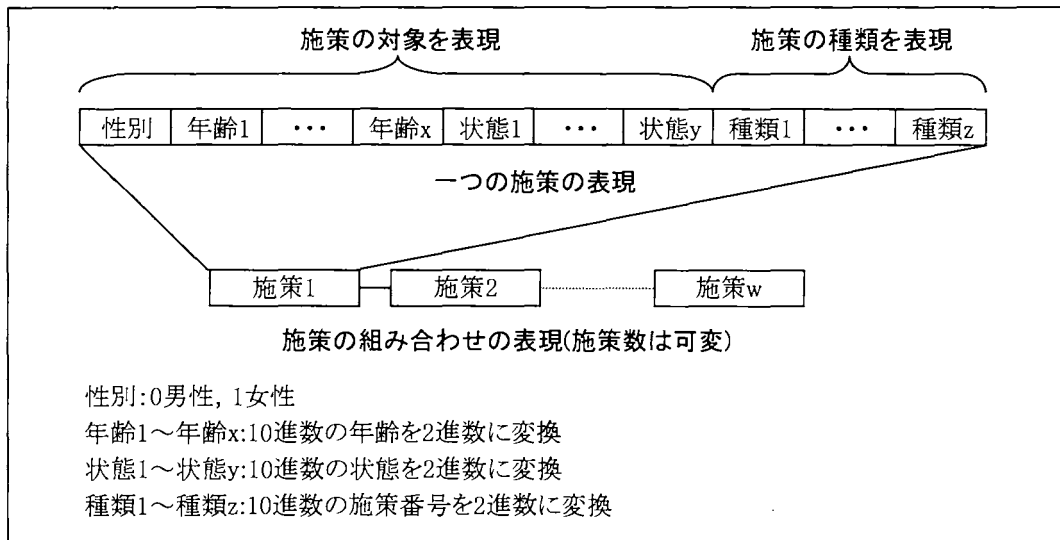


図6:施策の組み合わせの表現方法

(2)施策の組み合わせの評価方法

施策の組み合わせの評価方法(適合度)は、全顧客のLTV合計の平均を採用している。

また、一つの顧客に同じ対象者に対して複数の施策があった場合、その施策が異なる場合は、施策単独の効果を加算している。同じ施策の場合は、効果は施策一つ分、費用は施策数分だけ評価している。これにより、同じ対象者に対する複数の施策が増加することを防いでいる。

(3)GA の実行

GA は、初期集団を生成し、世代交代(進化)を繰り返し、終了条件に合致したら進化を終了する。この3つのステップについて説明する。

初期集団は一様乱数を使用して生成した。ただし、施策数を決定する必要があるため、まず一様乱数によって施策数を決定し、その施策数に基づいて施策の組み合わせを生成した。さらに、実際にありえない施策を含む組み合わせが生成された場合(致死遺伝子)は削除し、再度生成しなおした。

世代交替時の親の撰択(複製撰択)は、ルーレット撰択を使用した。子の生成は、ルーレット撰択によって選ばれた親の施策の組み合わせを交叉させることと、突然変異させることで行った。交叉が起きる場所は、バイナリ表現の施策と施策の間で起きることとした。これは、施策の組み合わせを変更させるためである。この際、交叉が起きる施策と施策の境目の位置に制限を設けないことで、遺伝子に含まれる施策数を変化させることと、末端で交叉が発生することで親と同じ個体が残る可能性を残した。突然変異については、発生場所に制限を設けなかった。これにより、初期集団に発生しなかった施策についても、発生するようにした。

終了判定は、評価が20世代改善しなかった場合とした。

(4)結果

実際に計算したLTVの合計値の分布を図7に示す。結果は、LTVの合計を現状よりも平均値は30%向上させることとなった。撰択された施策を見ると、LTVの低い状態の顧客に対する施策によってLTVの低い顧客が少なくなっていることと、LTVが高い顧客に対する施策によって、継続率が上昇しLTVが増加したことが要因と考えられる。

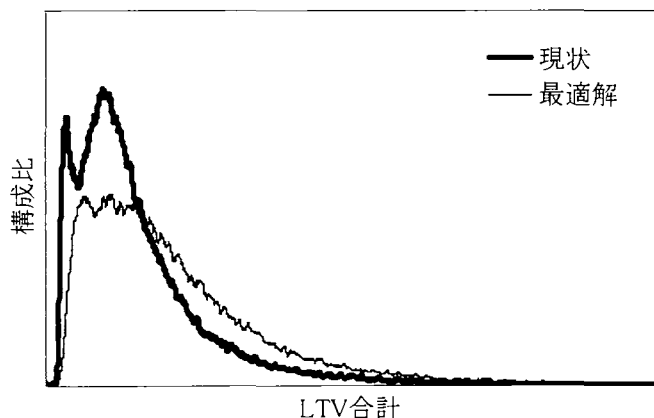


図 7:GA の結果(LTV の合計)

4.まとめと今後の展開

本報告では、モンテカルロ・シミュレーションを使用した LTV の分布の算出と、GA による最適な施策の組み合わせを求めることを試みた。その結果、LTV は分布を算出することで、その精度や施策の効果を評価する一つの指標として使用できる可能性を示すことができた。

LTV の計算方法については、パラメータ推定時に、現在は収益額の分布として三角分布やパレート分布等を使用している。しかし、この分布の適合度が低い場合が存在する。そのため、経験分布に変更して、適合度の向上を図る必要がある。また、LTV の評価方法は使用する目的によって異なるため、今後も検討が必要である。さらに、GA については、解空間が大きいいため、より効率的な探索が可能といわれている並列分散 GA における島モデルを SAS/CONNECT を使用して実施することを検討している。探索効率が向上することによって、より複雑なモデルを適用できる。例えば、本稿で想定している対象顧客が明確な施策だけではなく、支店の統廃合や新規出店といった銀行全体に影響する施策について、GA を使用して最適解を求められる可能性がある。

LTV の活用方法として、顧客の評価や今回紹介した施策の評価に加えて、支店あるいは行員の業績評価に使用することが考えられる。つまり、LTV の使用により、将来のキャッシュフローにつながる評価体系を構築することができる。また、本稿における施策の決定だけではなく、新規出店などの投資が必要な案件の意思決定のための一つの指標として活用できる可能性がある。

参考文献

- [1]石塚直樹著(2001),「SASによるモンテカルロ・シミュレーション」,第20回日本SASユーザー会総会および研究発表会論文集
- [2]岸本義之著(2001),「銀行業における顧客生涯価値」,慶應経営論集,18(2),1-21
- [3]戸谷圭子ほか著(2002),「カスタマー・セントリックをビジネスに結びつけるには」,金融財政事情,2002年1月14日号,47-50,金融財政事情研究会
- [4]坂和正敏ほか著(1995),「遺伝的アルゴリズム」,朝倉書店

日本SASユーザー会 (SUGI-J)

Bioinformatics の手法を活用したクレジットカード取引履歴データの途上審査モデルへの適用事例

○堀 彰男, 小谷田 知行
株式会社 浜銀総合研究所
戦略研究部

An Application of Transaction Records to Credit Risk Model
Using Methods of Bioinformatics

Akio Hori, Tomoyuki Koyata
Hamagin Research Institute, Ltd.
Dept.of Strategic Management & Consulting

要 旨

クレジットカードやカードローンの途上審査のモデル構築は、現在、決定木やロジスティック回帰等の手法を用いて行うことが多い。その際、分析に使用するデータの中で重要なものの一つに利用返済履歴があるが、履歴情報は、1 顧客当たりのレコード数が決まらず、かつデータ量が多く扱いにくいので、1 顧客1レコードとなる様に何らかの集約を行って使用することが多い。しかしながら、利用返済履歴を集約することは、利用返済のパターンの構造をモデル化しにくいという課題があった。

そこで、Bioinformatics で用いられている隠れマルコフモデルと動的計画法を用いて、利用返済パターンの構造をモデル化することを試みた。

キーワード： Bioinformatics, 隠れマルコフモデル, 動的計画法, DATA ステップ

1.はじめに

モデルに投入するデータは、そのモデルに適した形式(データ単位, 変数など)にするために、データの最小単位である履歴情報を加工する必要があり、その際に「情報の減少」が生じてしまう。データの「扱い易さ」と「情報の多寡」はトレードオフの関係にあり、現在広く用いられている審査モデルでは、時間単位(年月日)でデータを集約した集約情報が一般的に用いられている。

しかし、時間単位でデータを集約することで、時系列の「利用パターン」の情報を喪失し、それをモデルに反映できないという問題がある。

そこで本報告では、Bioinformatics の分野で DNA を構成するアミノ酸や塩基の配列パターンの構造をモデル化することで、DNA の機能予測に用いられている「隠れマルコフモデル(Hidden Markov Model;以下適宜, HMM と称す)」と、複数の DNA の機能的あるいは進化的関連性を計量する際に用いられている「動的計画法(Dynamic Programming;以下適宜, DP と称す)」に着目し、クレジットカードの利用返済履歴情報を用いて、利用返済パターンの構造をモデル化することを試みた。事例として、株式会社横浜銀行殿から提供頂いたクレジットカード情報を用いた途上審査モデルへの適用結果を併せて報告する。

2.隠れマルコフモデルによる利用返済パターンのモデル化

2.1.モデルの概要

隠れマルコフモデルは、観測不可能な状態からなるマルコフ過程と、その状態に依存するシンボル出力器の組合せによって、シンボル系列に対応する状態系列を表現するモデルである。

カジノにおけるサイコロゲームの例²⁾が理解し易いので、以下ではそれを紹介する。

あるカジノでは、ほとんどの場合に公正なサイコロを使用しているが、時々不正なサイコロを使用しゲームをコントロールしているとする。公正なサイコロでは全ての目が確率 1/6 が出るが、不正なサイコロでは 6 の目が確率 1/2、その他の目が確率 1/10 が出る。サイコロを振るたびにカジノは公正なサイコロから不正なサイコロに確率 0.05 で切り替え、不正なサイコロから公正なサイコロには確率 0.1 で切り替えると仮定する。この様な場合、全体の過程は HMM の一例となり、図 1 の様に表現でき、これを HMM のトポロジー(位相)と呼ぶ。また、サイコロの出目の確率を「シンボル出力確率」、サイコロの状態(公正・不正)間の切り替えの確率を「状態遷移確率」と呼ぶ。

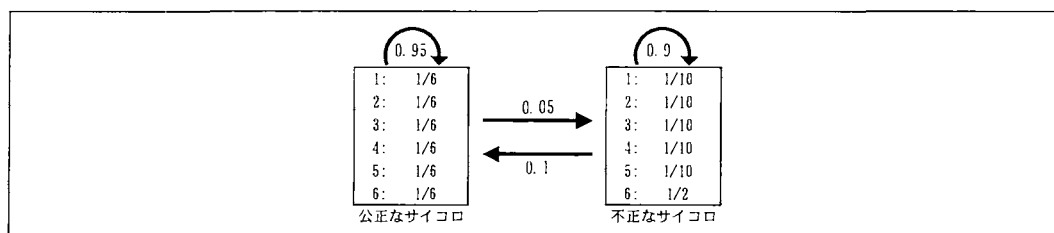


図 1: 隠れマルコフモデルのトポロジー

このカジノの例では、サイコロを振って出た目の列は観察することができる。しかしながら、公正なサイコロが使用されたのか不正なサイコロが使用されたのかについては、カジノが秘密にしているため、知ることができない。つまり、使用されたサイコロが公正なのか不正なのかが、「隠れた」状態ということになる。

以上の様に HMM のトポロジーを設計し、シンボル出力確率と状態遷移確率を求めることができれば、観察されたシンボル系列 x とその背後に隠れた状態系列 π の同時確率 $P(x, \pi)$ は式 1 で求めることが可能となる(式 1 は、配列の長さが L の場合)。

$$P(x, \pi) = a_{0, \pi_1} \prod_{i=1}^L e_{\pi_i}(x_i) \cdot a_{\pi_i, \pi_{i+1}} \quad \text{〈式1〉}$$

$a_{\pi_i, \pi_{i+1}}$: 状態 π_i から状態 π_{i+1} への遷移確率, $e_{\pi_i}(x_i)$: 状態 π_i でシンボル x_i を出力する確率

観察されたシンボル系列 x 各々に対応する状態系列 π を考える際に、全ての組合せの同時確率を算出することは、式 1 を用いれば理論的には可能である。全ての同時確率を算出すれば、その中で最も確率の高い(最も尤もらしい) x と π の組合せ(以下、パス π^* と称す)が、予測すべき状態系列となる。しかし、組合せ総数は π^L であり、 x と π の増加に伴い、指数関数のオーダーで増加する。この組合せ爆発問題に対し、これを多項式のオーダーで解くことができる手法として「ビタビ・アルゴリズム」が知られており、次節でその概要を説明する。

2.2.モデルの解法

前述の最も尤もらしいパス π^* は、ビタビ・アルゴリズムを用いて求めることができる。シンボル系列の $i-1$ 番目について、状態 π_{i-1} で終わる最も尤もらしいパスの確率 $v_{\pi-1}(i-1)$ が全ての状態 π_{i-1} についてわかっていると仮定すると、観察系列の i 番目について、それらの確率 $v_{\pi}(i)$ は式 2 で求めることができる。

$$v_{\pi}(i) = e_{\pi}(x_i) \cdot \max_{\pi-1} (v_{\pi-1}(i-1) \cdot a_{\pi-1,\pi}) \quad \text{〈式 2〉}$$

$a_{\pi-1,\pi}$: 状態 π_{i-1} から状態 π_i への遷移確率, $e_{\pi}(x_i)$: 状態 π_i でシンボル x_i を出力する確率

つまり、 i 番目までの同時確率を、 $i-1$ 番目までの同時確率から導くことができるので、小さな部分問題の解を記録しておき、その解を利用しながら徐々に大きなサイズの部分問題を解いていくという、再帰的な問題解決手法であるといえる。

図2は前述のカジノのサイコロゲームではモデルが複雑なので、さらに単純なモデルに対し、ビタビ・アルゴリズムを適用し、最も尤もらしい状態系列を求める手順を示している^[1]。

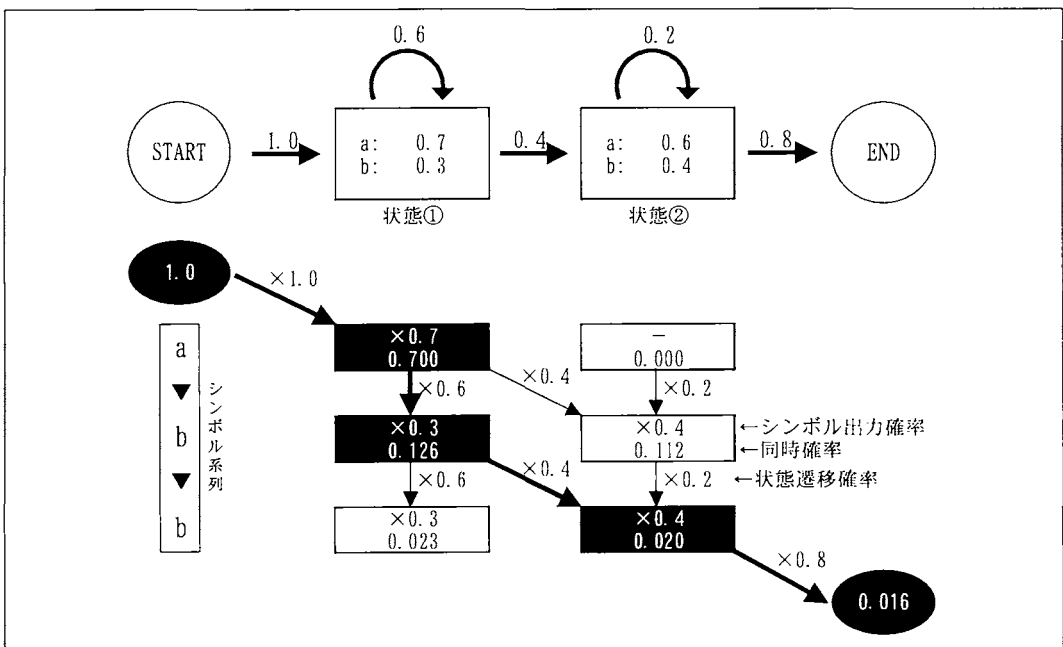


図 2:ビタビ・アルゴリズムの例^[1]

左上の箱から右下の箱まで順番に計算を行い、箱の下段に書かれている数値がその段階で最も確率の高い同時確率である。白黒反転した箱と太い矢印が、求めたい最も尤もらしいパス π^* で、その同時確率は 0.016 となる。

図 2 の例では、ビタビ・アルゴリズムは 8 通り (2 の 3 乗) の組合せ最適化問題を、6 (2 × 3) 個の箱に数値を埋めていくことで解いている。一般化すると、 π^x 通りの組合せ最適化問題を、 $\pi \times x$ 回のステップで解いているということである。

2.3.パラメータの推定

全てのシンボルおよび状態があらかじめ既知の学習用データ(以下、既知配列と称す)がある場合、HMM のパラメータである状態遷移確率とシンボル出力確率は、それら既知配列において出力および遷移した回数を数え上げることで、最尤推定することが可能である。シンボル出力確率を求める一般式を式 3、状態遷移確率を求める一般式を式 4 に示す。

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \quad \langle \text{式 3} \rangle, \quad a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad \langle \text{式 4} \rangle$$

$e_k(b)$: 状態 k でシンボル b を出力する確率, $E_k(b)$: 状態 k でシンボル b を出力した回数
 a_{kl} : 状態 k から状態 l へ遷移する確率, A_{kl} : 状態 k から状態 l へ遷移した回数

2.4.途上審査モデルへの適用

(1)考え方

今、HMM とビタビ・アルゴリズムを用いることで、観察されるシンボル系列の背後に隠れている状態系列を予測することが可能になった。ここでは、これをクレジットカード情報を用いた途上審査モデルに適用する際の基本的な考え方を述べる。

HMM におけるシンボル系列とは、会員ごとに作成した一定期間内における配列情報のことであり、会員のカード利用履歴をそのまま再現していると言える。状態系列は、会員の状態と解釈し、デフォルト(以下適宜、BLACK もしくは B と称す)と正常(以下適宜、WHITE もしくは W と称す)の 2 種類からなる会員属性であるとする。つまり、会員の状態 (BLACK or WHITE) によって、カードの利用パターンが異なるのではないかという仮説に基づく。

具体的には、図 3 に示す様に、状態系列が既知の配列を用いて HMM のパラメータを推定し、得られたパラメータから HMM のトポロジーを設計する。設計した HMM のトポロジーをもとにビタビ・アルゴリズムを適用することで、状態が未知の配列(以下、未知配列と称す)のシンボル系列に対して最も尤もらしい状態系列のパスを推定する。

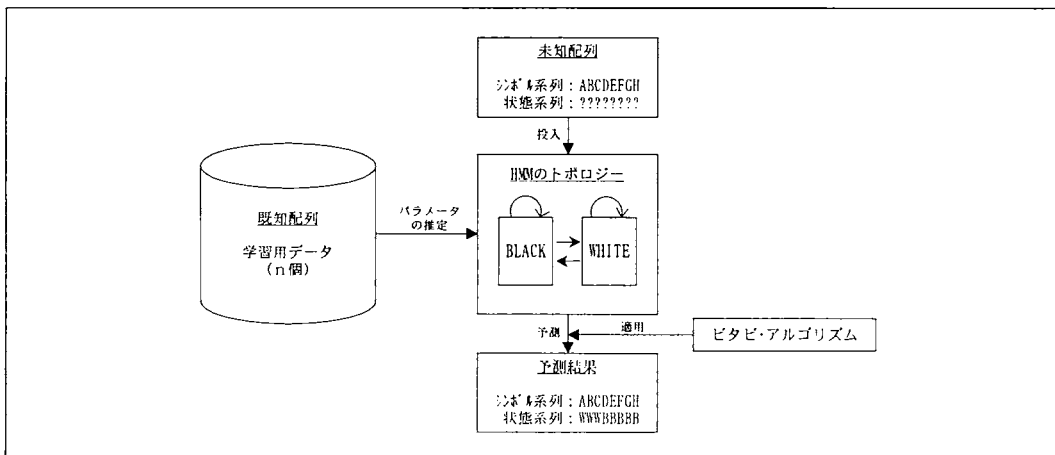


図 3: 途上審査モデルへの適用

(2)前提条件

取引観測期間は12ヶ月間とし、BLACKはデフォルトした月から遡って観測し、WHITEはランダムに決定した月から観測した。

シンボル出力確率と状態遷移確率を最尤推定するためには、あらかじめシンボルと状態が既知でなくてはならない。従って、取引履歴そのものであるシンボルに対して状態を定義する必要がある。今回は、BLACKは前半の6ヶ月をWHITE、後半の6ヶ月をBLACKとし、WHITEは全てWHITEと一意に定義した。

また、取引回数による結果への影響を把握する目的で、取引観測期間中に1回以上取引している会員を対象とした実験(以下適宜、実験①と称す)と、12回以上取引している会員を対象とした実験(以下適宜、実験②と称す)を行った。使用したデータの件数を示したものが表1である。

表1:使用したデータ件数

	実験①		実験②	
	BLACK	WHITE	BLACK	WHITE
既知配列(学習用)	500	9,500	250	9,750
未知配列	100	100	100	100

(3)結果

予測された未知配列の状態系列に関し、その予測精度を実験別に示したものが表2である。ここでは、各配列の最後の状態のみの一致と不一致に着目し、予測精度を算出している。

表2:モデルの予測精度(上:実験①,下:実験②)

実験①		予測		一致割合	χ^2 検定
		一致	不一致		
観測	BLACK	53	47	53.00%	$p \cong 0.549 > 0.05$
	WHITE	99	1	99.00%	$p < 0.05$
全体		152	48	76.00%	$p < 0.05$

実験②		予測		一致割合	χ^2 検定
		一致	不一致		
観測	BLACK	68	32	68.00%	$p < 0.05$
	WHITE	89	11	89.00%	$p < 0.05$
全体		157	43	78.50%	$p < 0.05$

表2の結果を以下に挙げる2つの側面から考察する。

- a)実験①, ②とも、WHITEの予測精度がBLACKの予測精度よりも高い。
- b)実験②の方が実験①よりも、全体としての予測精度が高い。

a)は、パラメータを最尤推定する際に使用した既知配列に占めるBLACK配列の構成比が低いいため、WHITEからBLACKに遷移する状態遷移確率が小さく推計される傾向にあるためと考えられる。また、前提条件における状態の定義で、BLACKからWHITEへの状態遷移確率を実質的にゼロにしていることも原因であると考えられる。

b)は、本報告の目的であるパターンのモデル化において、一定取引回数未達の短い配列では、パターンを形成すること自体に無理があるということを示しているものと考えられる。

3.動的計画法による利用返済パターンのモデル化

3.1.モデルの概要

「アラインメント」とは、'align'で「一列に並べる,揃える」, 'alignment'で「整列」という意味を持ち、Bioinformaticsの分野で、塩基配列やアミノ酸配列の機能的あるいは進化的な関連性(類似性)を計量する際に用いられる一連の作業のことを言う。

例えば、図4上に示す2つの配列をアラインメントしたいとする。ここで、同じ配列位置に異なる塩基やアミノ酸が格納されることよりも、その配列位置にギャップを挿入しnull状態にした方が評価が高い(つまり、AとBの機能がまったく異なる)とすると、図4下に示す4通りのギャップを挿入するアラインメントが最も効率の良いアラインメントとなることは、直感的に理解できる。

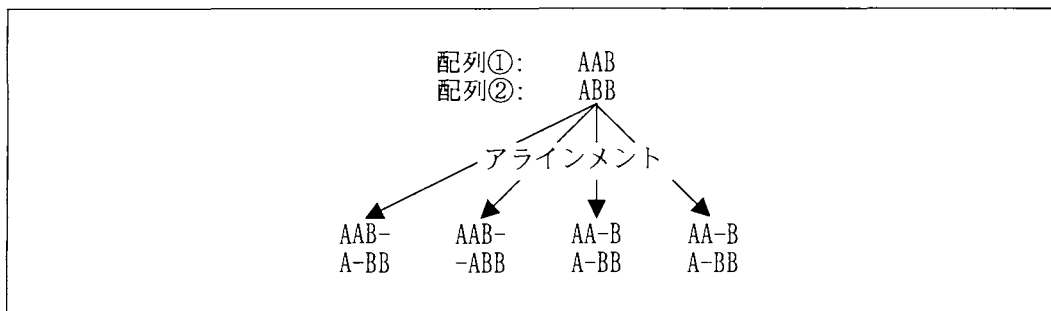


図4:アラインメントの概念

図4の様な単純な例であれば簡単に解を求めることができる(ただし、図4が最適解である保証はない)が、配列長が長くなりかつ塩基やアミノ酸の種類が増えることで、直感で解を導くことは困難になる。因みに、配列長が n の2本の配列間には、全ての位置にギャップが入った場合の最大配列長 $2n$ から n 個を選択する組合せ問題になるので、式5に示した組合せ数が存在し、その組合せ総数は n が増加するに従い、階乗のオーダーで増加する(表3)。この組合せ爆発問題を問題の規模 n の2乗のオーダーで解くことができる手法として「動的計画法」が知られており、次節でその概要を述べる。

$$\binom{2n}{n} = {}_{2n}C_n = \frac{(2n)!}{n!(2n-n)!} = \frac{(2n)!}{(n!)^2} \quad \langle \text{式5} \rangle$$

表3:組合せ総数

n	組合せ数	n	組合せ数
1	2	11	705,432
2	6	12	2,704,156
3	20	13	10,400,600
4	70	14	40,116,600
5	252	15	155,117,520
6	924	16	601,080,390
7	3,432	17	2,333,606,220
8	12,870	18	9,075,135,300
9	48,620	19	35,345,263,800
10	184,756	20	137,846,528,820

3.2.モデルの解法

動的計画法とは、サイズの小さな部分問題の解を記録しておき、その解を利用しながら徐々に大きなサイズの部分問題を解いてゆき、最終的に解きたい問題の解を得るという手法である。言い換えると、 n 時点の解を $n-1$ 時点の解を用いて解く、再帰的な問題解決手法である。この考え方を配列間のアラインメントに応用する。

例えば、以下の2つの配列をアラインメントする問題を考える。

配列①: HEAGAWGHEE
配列②: PAWHEAE

図 5:アラインメントしたい2つの配列

ここで、配列を構成する塩基やアミノ酸同士の(機能的な)類似度を表したスコアテーブルと、ギャップを挿入した場合のスコア(以下、ギャップコストと称す)をあらかじめパラメータとして与え、DP 行列(図 6)の各段階で式 6 を解いてゆくことで、アラインメントの最適解を求めることができる(図 7)。この2つの配列の類似度を表すスコアはDP行列の右下に記録されたスコアである1となる。言い換えると、配列①と配列②の遺伝学的距離が1ということになる。

$$F(i, j) = \max \left\{ \begin{array}{l} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{array} \right\} \quad \text{〈式 6〉}$$

$F(i, j)$: DP 行列, $s(x_i, y_j)$: スコアテーブル, d : ギャップコスト

$F(i, j)$		配列①										
		H	E	A	G	A	W	G	H	E	E	
配列②	P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
	A	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
	W	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
	H	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
	E	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
	A	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
	E	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
	E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

図 6: DP 行列とアラインメントの最適な経路^[2]

配列①: HEAGAWGHE-E
配列②: --P-AW-HEAE

図 7:アラインメントの最適解

3.3. 途上審査モデルへの適用

(1) 考え方

今、DPを用いることで2つの配列の類似度を計量することが可能になった。ここでは、これを途上審査モデルに適用する際の基本的な考え方を述べる。

会員ごとに一定期間内における配列情報を作成する。この配列情報は、会員のカード利用履歴をそのまま再現していると言える。従って、カード利用の傾向が類似しているということは、会員のクレジットカードに対する価値観も類似しており、さらには、会員の属性も類似しているのではないかと考えられる。言い換えると、デフォルトする会員はカード利用の傾向が類似しており、ある特定の使い方をしている会員は最終的にデフォルトするといった、カード利用のパターンがあるのではないかと仮説に基づく。

従って、会員の状態(BLACK or WHITE)があらかじめ既知の配列と、状態の予測を行いたい未知の配列とで類似度を計量することで、未知配列がどの状態の既知配列に類似している確率が高いかを求めることが可能となり、その結果を用いてBLACKとWHITEの判別を行う。

具体的には、図8に示す様に、未知配列に対する既知配列の距離行列を算出し、未知配列に対して距離の近い上位m個の既知配列(以下、近隣配列と称す)を抽出し、その中に占めるBLACKの配列の構成比を算出することで、未知配列の状態を確率的に判別する。

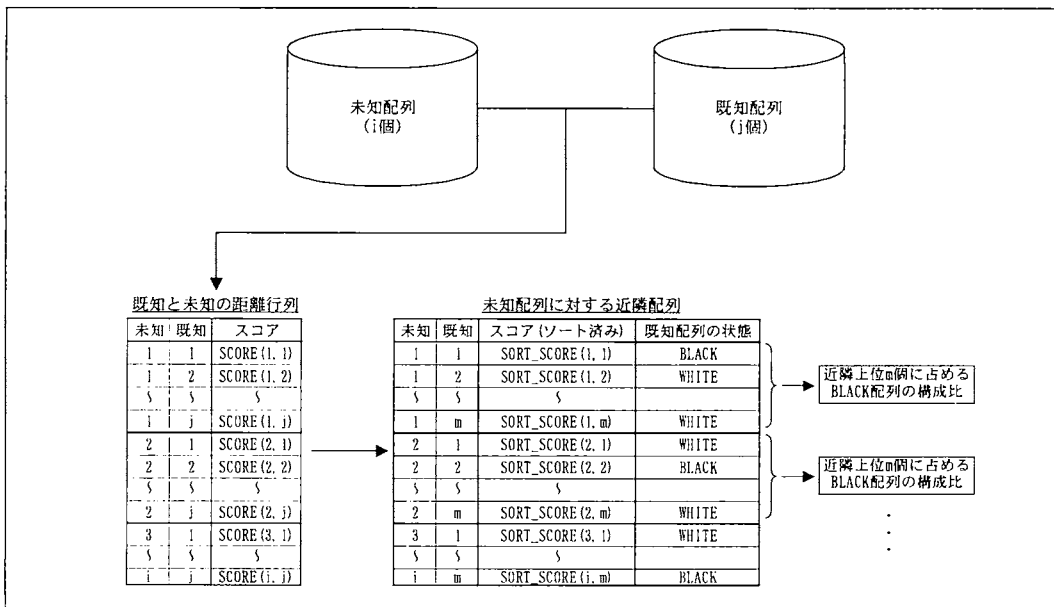


図8: 途上審査モデルへの適用

(2) 前提条件

取引観測期間は6ヶ月間とし、BLACKはデフォルトした月から遡って観測し、WHITEはランダムに決定した月から観測した。

スコアは、シンボルが一致していればスコア0、一致していなければスコア2とし、ギャップコストはスコア1とした。また、取引回数による結果への影響を把握する目的で、取引観測期間中に1回以上

取引している会員を対象とした実験(以下適宜, 実験③と称す)と, 6 回以上取引している会員を対象とした実験(以下適宜, 実験④と称す)を行った. 使用したデータの件数を示したものが表 4 である.

表 4: 使用したデータ件数

	実験③		実験④	
	BLACK	WHITE	BLACK	WHITE
既知配列	500	9,500	300	9,700
未知配列	100	100	100	100

(3)結果

図 8 で示した様に, 未知配列に対する既知配列の距離行列を作成し, 未知配列の近隣上位 m 配列に占める BLACK 配列の構成比を算出する(ここでは, m を 100 とした). そして, 既知配列全体に占める BLACK の構成比(3%)を閾値とし, 近隣配列に占める BLACK 配列の構成比が閾値以上の未知配列を BLACK, 閾値未満の未知配列を WHITE とした時の一致と不一致による予測精度を示したものが表 5 である.

表 5: モデルの予測精度(上: 実験③, 下: 実験④)

実験③		予測		一致割合	χ^2 検定
		一致	不一致		
観測	BLACK	88	12	88.00%	$p < 0.05$
	WHITE	46	54	46.00%	$p \equiv 0.424 > 0.05$
全体		134	66	67.00%	$p < 0.05$

実験④		予測		一致割合	χ^2 検定
		一致	不一致		
観測	BLACK	84	16	84.00%	$p < 0.05$
	WHITE	87	13	87.00%	$p < 0.05$
全体		171	29	85.50%	$p < 0.05$

表 5 の結果を以下に挙げる 2 つの側面から考察する.

- a) WHITE の予測精度が, 実験④において, 大幅に向上している.
- b) BLACK の予測精度は, 実験③, ④とも大きな差はない.

a)は, HMM の結果と同様, 一定取引回数未満の短い配列では, パターンを形成すること自体に無理があるということを示しているものと考えられる.

b)は, BLACK の会員あたりの平均取引回数が WHITE よりも全般的に高く, 実験③と④で大きな差がなかったことと, BLACK は WHITE と比べて特徴的なパターンを形成する傾向が強く, デフォルトに至るまでの利用返済のパターンの種類が限定されるためと考えられる.

4.まとめと今後の課題

本報告では, Bioinformatics の分野で用いられている隠れマルコフモデルと動的計画法を用いて, 利用返済パターンの構造をモデル化することを試み, これを途上審査モデルへ適用した事例を紹介した. 結果, 従来扱いが困難であった履歴情報のパターンの構造をモデル化することに成功し, 実際の業務への適用に関しても, ある程度の精度を確保することができた.

HMM の今後の課題は, 状態系列を厳密に定義することが事実上不可能である点に関し, バウム・ウェルチ・アルゴリズムを用いて, シンボル系列のみから確率的に状態系列そのものを予測する方法

を試みたい。また、繰り返し学習のアルゴリズムを用いて、最適な既知配列群のみ抽出し、効率的な
途上審査システムを実現したい。

DPの今後の課題は、Smith-Waterman-Gotoh アルゴリズムを用いた計算の効率化と、シンボル間
のスコアテーブルやギャップコストを意図的に設定することによる、結果への影響を把握したい。また、
予測結果を常に既知配列群にフィードバックすることで、システムの運用とモデルのチューニングを
同時並行で行える途上審査システムを実現したい。

最後に、本手法はあらゆる履歴情報のパターンをモデル化することが可能であり、応用範囲も広
いと考えられる。今後は、クレジットカードの不正利用の検知や、POS 情報や WEB アクセスログ情報
などを活用し、マーケティング分野などに応用していきたい。

5.本手法の SAS による実現

未知配列と既知配列の情報を格納した 2 つのデータセットから、該当する任意のオブザベーション
を抽出し、同じ DATA ステップ内で計算を実行するために、POINT=オプションを活用した。以下に
DP におけるコードの例を示す。

```
001 data dp;
002   set unknown;
003   array known {10}; /* 既知配列のシンボル系列 */
004   array state {10}; /* 既知配列の状態系列 */
005   array score {10}; /* スコア */
006   retain known: state: score;;
007   /** 既知配列の情報の読込 **/
008   if _n_ eq 1 then do;
009     do i = 1 to 10;
010       set known point = i;
011       known {i} = known;
012       state {i} = state;
013     end;
014   end;
015   /** 計算の開始 **/
016   do i = 1 to 10;
017     %DP (known {i}, score {i}); /* スコアの算出 */
018   end;
019   %BBLSORT; /* スコア順に配列をソート (バブルソート) */
020   %JUDGMENT; /* 判定 */
021 run;
```

参考文献

- [1] 鹿野清宏他編著(2001),「音声認識システム」, オーム社
- [2] Richard Durbin ほか著, 阿久津達也ほか訳(2000),「バイオインフォマティクス—確率モデルによる遺伝子配列解析—」, 医学出版

口頭論文発表
SASソリューション

日本SASユーザー会 (SUGI-J)

ゲノム創薬向け統合ソリューション SAS Scientific Discovery Solutionsの紹介

段谷 高章

SAS Institute Japan 株式会社
カスタマーサービス本部

An Introduction to Genomics and SAS Scientific Discovery Solutions

Takaaki Dantani

Customer Service Department, SAS Institute Japan Ltd.

要 旨

近年、ゲノム関連機器の発展に伴い、複雑で大容量のデータの分析と管理が必要となってきた。SASはゲノムデータの管理と解析のために SAS Scientific Discovery Solutions (SDS)という新たな製品を発表した。本論文では、新製品の SAS Scientific Discovery Solutions に含まれる2つの Solution である SAS Research Data Management (RDM)と SAS Microarray Solution (MAS)を簡単に紹介する。

キーワード： SAS SDS、 SAS RDM、 SAS MAS

1. はじめに

デオキシリボ核酸(DNA)はアデニン、シトシン、グアニン、チミンの4つの塩基から構成されており、二重螺旋構造を有している。遺伝子とは各タンパク質のアミノ酸配列を指定するもので、生物には数千～数十万種類のタンパク質が存在し、塩基3つの順序で一つのアミノ酸を指定している。この遺伝子の全貌を明らかにすることで、従来のアプローチでは困難だった医学的問題の解決に繋がることが期待されている。

近年、DNAシーケンサやMicroarray、質量分析計などの進歩にはめざましいものがある。これらの機器からのデータは膨大な量であり、分析には優れたコンピューターサイエンスと解析手法が必要となる。本論文は、SAS がゲノムデータの管理と分析のために発表した SAS Scientific Discovery Solutions (SDS)に含まれる2つの Solution である SAS Research Data Management (RDM)と SAS Microarray Solution (MAS)を紹介するものである。

2. SAS Scientific Discovery Solutions

Scientific Discovery Solutions (SDS)は複数のSolution群から構成されており、2つのSolutionが公開されている。SAS Research Data Management (RDM)とSAS Microarray Solution (MAS)である。まず、RDMはJavaベースのグラフィカル・ユーザー・インターフェース(GUI)にてデータの読み込みや管理を行なうことが可能なSolutionである。次に、MASはRDMを基盤としたアプリケーションで、Microarrayデータの分析方法と入力エンジンを搭載したものである。各々の機能に関する詳細は後述する。

3. SAS Research Data Management

ゲノム実験によって発生する膨大なデータは組織で管理し、利用するという観点が必要となる。当然ながら、研究者は分析に必要なデータをプログラミングすることなく直接取得することを望むと考えられる。さらに、数値のようなデータのみが唯一の情報ではなく、データと関連する分析結果を見る必要もある。RDMはSDSにおけるデータ管理のコアであり、ゲノムデータと関連する補助情報を集約するためのJavaクライアント-サーバーアプリケーションである。図1はRDMの画面である。左に階層構造で記載されている個所が、データや文書などの一覧となっており、右の画面は、必要な情報の検索を行う画面である。

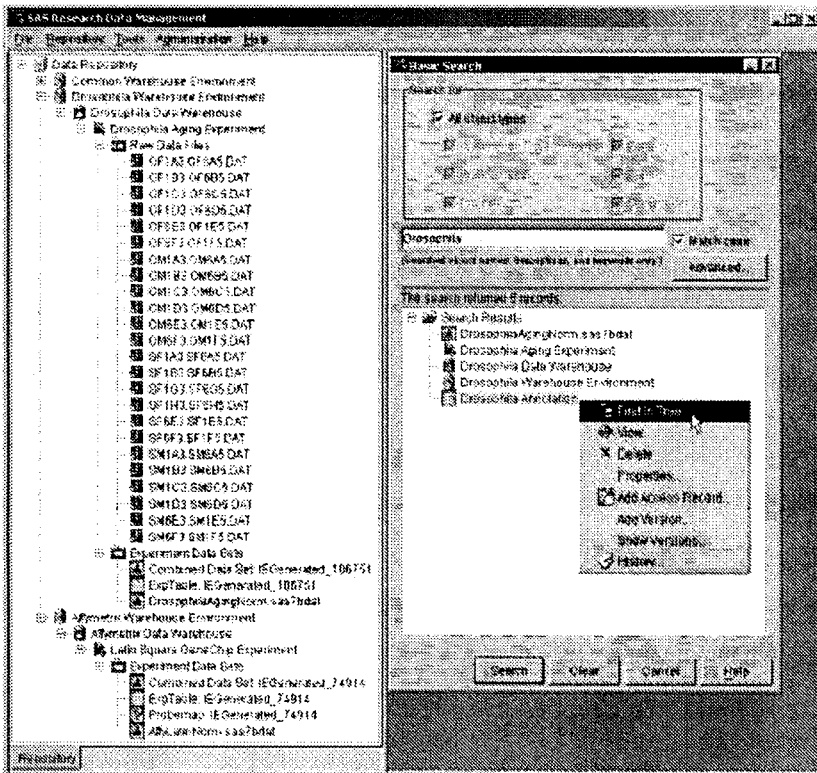


図1 RDMの画面

3.1 Pooled Metadata Repository

データウェアハウジングの概念には、データを抽出し、分析や報告するのに適した形式でデータを再構成する流れを含んでいる。データウェアハウジングの鍵となるものはメタデータである。データ名や保存場所・フォーマットや構造等の前後関係の情報がメタデータである。また、メタデータにはデータ変換のプロセスも含まれている。

RDM はデータを集中的に管理するためのプラットフォームを提供するものであり、メタデータを統合整理する方法として採用されているのが、Pooled Metadata Repository (PMR) である。データを物理的に動かすのではなく、作成されたメタデータを読み込む事で、データや補助情報の閲覧や利用が可能となるのが PMR の考え方である。ユーザーは PMR を通じてデータや文書を含む全ての情報を検索、ダウンロードする事が可能となる。検索とダウンロード機能に加えて、PMR に新しいデータや文書を登録するアップロード機能も存在する。アップロードされた情報は、アクセス可能なユーザーであれば誰でも閲覧可能となる。

3.2 Security

セキュリティモデルとしては、ユーザー名とパスワードを必要とする方法を選択している。このアクセス権限に関しては、ユーザーレベル・ユーザーグループレベルで制御することが可能であり、閲覧専用のアクセス権限と編集が可能な権限が存在する。RDM で採用されているセキュリティにはシステム内での行動に関する Audit Trail と、データのバージョン管理も含まれている。Audit Trail とデータのバージョン管理はデータソースと修正方法を確認することが可能なものである。

4. SAS Microarray Solution

RDM を基盤とした Solution としてまず開発されたのが、Microarray データを管理・分析・視覚化するための製品の MAS である。MAS で追加される機能は、入力エンジンと分析プロセスの二つである。

4.1 Input Engines

入力エンジンは測定機器からの出力されたデータを MAS のウェアハウス内に取り込む機能である。入力エンジンは入力するデータの構造によって特有のものであり、各々の実験に合わせてカスタマイズすることが可能である。入力エンジンを利用する事で、生データを容易に読み込むことが可能となる。生データだけではなく、実験の要因の構造を示すファイルが必要となる。

次ページの図 2 は Input Engine のメニュー画面を表示したものである。ここでは、どのようなデータであるかを指定し、読み込みに使用するエンジン(SAS マクロプログラムと Java Class)を指定することができる。エンジンを新たに作成することで、デフォルトでは読み込むことができないようなデータでも読み込むことが可能となる。

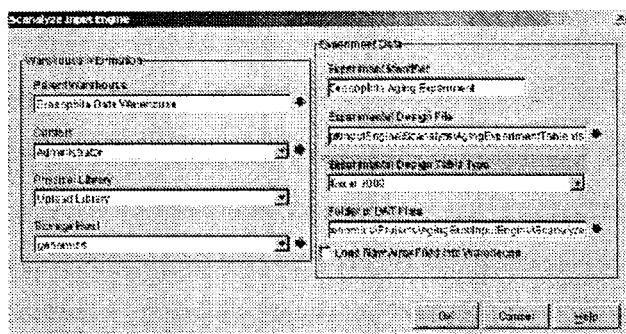


図 2 Input Engine のメニュー

4.2 Analytical Process

分析プロセスは、RDM 上でデータの操作と統計的な計算を行うマクロプログラムである。このプロセスは分析データ・統計結果・グラフ等の作成を行うもので、柔軟に利用することができる。例えば、単純なデータ表示から複雑な統計モデリング機能まで持ち合わせている。ユーザーは入力パラメータに適切な値を入力する事で、分析プロセスを実行することが可能である。パラメータの値はデータによって特有であるが、MAS に登録した分析プロセスを読み込む際は、パラメータやコードの編集を行う必要はない。

MAS には以下の 4 つの分析プロセスが搭載されている。

- DataContents : SASデータセットの内容をHTML形式で表示する
- ArrayGroupCorrelation : ユーザーによって選択されたグループに分割し、多変量相関分析を行う
- MixedModelNormalization : 配列全てを横断した線形混合モデルをあてはめることにより、Microarray データを標準化する
- MixedModelAnalysis : Gene-by-Gene に基づく混合モデルをあてはめる

分析プロセスに関しても、入力エンジンと同様にユーザー特有のものを追加することが可能であり、SAS マクロ言語を用いてプログラムを記述する必要がある。また、入力パラメータの指定等も考慮する必要もある。さらに、出力に際して JMP を使用するのであれば、JSL のコーディングが必要となる。

4.3 Mixed Model Analysis

MixedModelAnalysis は非常に複雑な分析プロセスであり、あらかじめ正規化されたデータを用いて、高水準の混合モデル分散分析を行うものである。重要な入力パラメータは%str()で囲われた PROC MIXED ステートメントである。このプログラムは主効果によって調整された三元配置型の階乗モデルである。モデルを各々の遺伝子毎に当てはめるには BY ステートメントを使用する。個々の遺伝子水

準に等しい配列は、変量効果であると考えられ、同じ点から観測された 2 つの測定値の間に強い交互作用が説明されることが多々ある。次ページの図 3 は MixedModelAnalysis のパラメータ設定画面である。

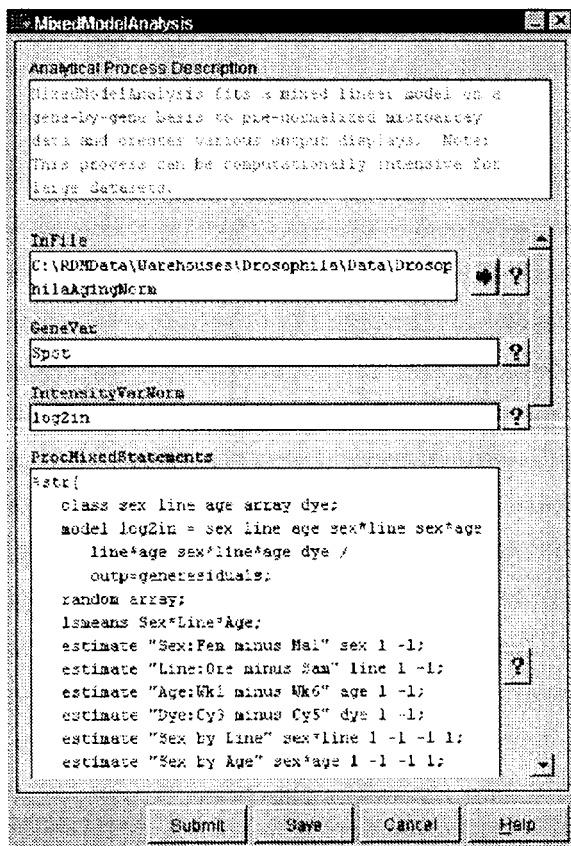


図 3 MixedModelAnalysis のパラメータ設定画面

MixedModelAnalysis は ArrayGroupCorrelation と同様に、SAS サーバー上で処理を実行し、JSL ファイルを作成する。その後、JMP にて JSL ファイルを実行し、分析結果を図示する。次ページの図 4 は JMP を用いて結果を表示したものである。左上のグラフはボルカノプロットであり、指定した ESTIMATE ステートメント毎に作成される。左下のグラフは、少なくとも 1 つのボルカノマップにおいて Bonferroni 流で切り分けられた遺伝子の最小自乗平均の平行座標プロットである。中央下のグラフは最小自乗平均のグラフであるが、平均 0、分散 1 に標準化されたグラフである。中央上のグラフは、有意な遺伝子において標準化された最小自乗平均の構成要素をプロットしたものである。右の図は、有意な遺伝子において標準化された最小自乗平均の階層的クラスター分析の結果である。横軸は遺伝子を表しており、縦軸はカテゴリーを表現している。このプロットの左側には、各遺伝子に関する情報が表示されている。表示されている画面で利用されている色分けは、クラスター分析に由来するもので、このような JMP のダイナミックリンクと対話性は、統計結果の解釈に適している。

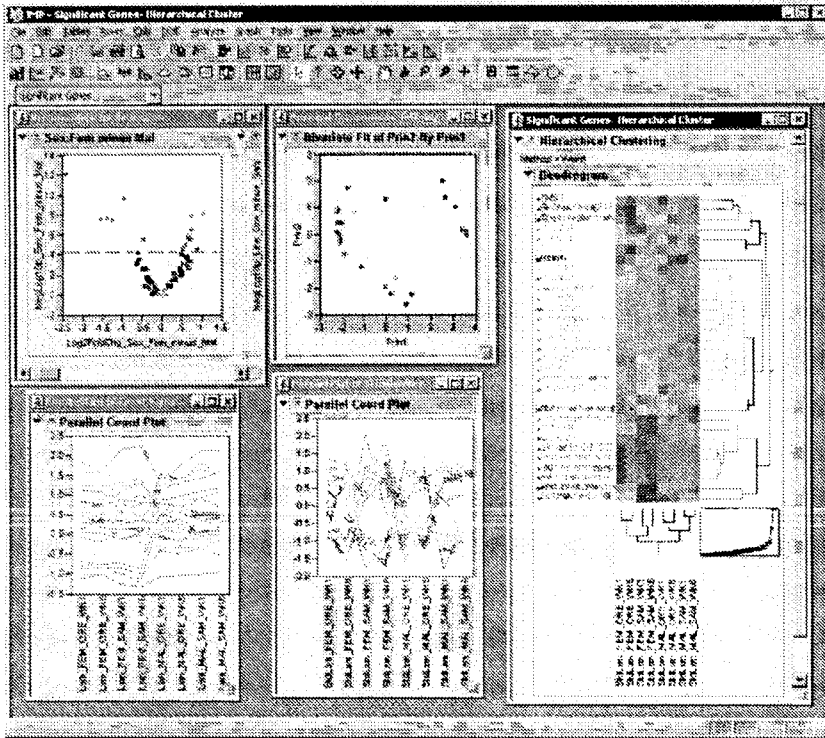


図4 JMPによる分析結果

DataContents と MixedModelNormalization は、JMP による出力を生成することではなく、Output Delivery Systemを通じて、HTML形式で結果を出力する。

5. まとめ

RDMは、使いやすいインターフェイスで情報の検索と利用が可能となるデータウェアハウスである。その管理技術を用いて、様々なフォーマットやタイプの異なるデータを管理することが可能となる。編集可能な権限を持つユーザーが関連するデータ(文書等も含む)を登録する事で、アプリケーションを使用しているユーザーも情報の閲覧が可能となる。RDMの使用によるデータ管理の合理化とデータアクセスの簡便化は、組織の生産性を向上することになると考えられる。また、入力エンジンや分析プロセスを新たに作成することも可能であり、Microarray や今後追加されるであろう他の Solution 以外の使用法も考えられる。ゲノム創薬の分野だけでなく、使用法しだいでは他の分野での利用も可能となってくる。

MASはRDMを基盤とし、統計解析者と科学者が共に協力するという形態の元にデザインされている。統計解析者は実験計画に沿って適切な分析プロセスを作成することができるようになる。一方、科学者にとっては、SASプログラミングなしに利用する事ができ、統計解析者が作成したプロセスを再利用する事ができる。このように、科学者や統計解析者が専門

知識を必要とする業務に、より多くの時間をかけることができるようになると思う。当然ながら、ソフトウェアは、調査・考察・個々が協力することで得られる効果の代わりにはならない。しかし、上記のような業務活動を円滑に進める一端を担うことは可能である。ゲノム関連情報の量と密度が増大した近年、この流れはより顕著であると思う。

引用文献

- Deng, S., Chu, T.-M., and Wolfinger, R.D. (2002), Transcriptome variability in the normal mouse, manuscript to be published in the CAMDA proceedings, Duke University.
- Gibson, G. (2002), MMANMADA Tutorial, <http://statgen.ncsu.edu/ggibson/Pubs.htm>
- Jin, W., Riley, R., Wolfinger, R.D., White, K.P, Passador-Gurgel, G. and Gibson G. (2001), Contributions of sex, genotype and age to transcriptional variance in *Drosophilamelanogaster*, *Nature Genetics*, 29:389-395.
- Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001), Assessing gene significance from cDNA microarray data via mixed models, *Journal of Computational Biology*, 8, 625-637.
- Wolfinger, R.D. et al. (2003), An Introduction to Genomics and SAS® Scientific Discovery Solutions, *SUGI 28 Proceedings*, SAS Institute, Inc., Cary, NC.

ポスターセッション 統計解析

日本SASユーザー会 (SUGI-J)

一般化推定方程式および SAS の解析ツール

○王露萍* 高田 康行** 野口 知雄*

* アベンティス ファーマ株式会社 生物統計・データマネジメント部

** 持田製薬株式会社 医薬開発部

Generalized Estimating Equations (GEE) and Analysis Tools of SAS

○Luping Wang* Yasuyuki Takata** Tomoo Noguchi*

*Biostatistics/Data management Department, Aventis Pharm.

**Clinical Development Department, Mochida Pharmaceutical CO.LTD.

要 旨

臨床試験において、応答変数が二値や計数のデータ、説明変数が経時データの場合には、一般化推定方程式を用いる解析が増えてきた。本論文では一般化推定方程式 (GEE) を応用するための理論を示し、SAS の3つの解析ツール (GENMOD プロシジャ、「SASによる貧乏人のGEE」¹⁾、および「GEE マクロ」²⁾) を解析した結果を比較した。3つの解析ツールにおける精度が同じぐらいであったことがわかった。

キーワード： 一般化推定方程式、経時データ、二値データ、SAS の解析ツール

1. はじめに

臨床試験や疫学研究において治療効果や曝露効果を調べるための統計手法として回帰モデルがよく用いられる。例えば、応答変数が連続量の場合の回帰分析、二値変数の場合のロジスティック回帰、生存時間の場合の比例ハザードモデルなどである。これらの回帰モデルで解析するデータが「互いに独立である、あるいは無相関である」という仮定が必要である。しかしながら、臨床試験の場合、同一対象者に対して経時的に得られるデータを通して解析し、治療効果を評価することがよくある。このような場合には無相関の仮定は正しいとは言えない。そのまま相関を無視した解析を行うと、推定されたパラメータの効率の損失・解析パラメータの分散の一致性が保証されないなどの問題が存在する。したがって、このようなデータにモデルを当てはめ、治療およびいくつかの説明変数と応答変数との間の関係を定量的に評価する際にはデータ間の相関を何らかの形で考慮すべきである。応答変数が連続量で、説明変数の中に経時データがあり、誤差に正規性が仮定できるような場合には、線形モデルが解析に利用可能である。SAS では MIXED プロシジャを用いることにより実行可能である。応答変数が離散型の場合、GEE が解析に利用され、SAS にもいくつかの解析ツールがある。

2. 一般化推定方程式

一般化推定方程式は経時または相関があるデータ、特に応答変数が二値の解析によく使用される。1986年にLiang and Zegerが周辺モデルの1つとして一般化推定方程式を提案してから、応答変数が二値や順序分類といった計数データ、説明変数が経時データの場合には、時点間の相関を考慮し、GEEにより解析を行うことで、ロバスト(robust)な結果を得られると考えられる。

GEEではN人の対象者がいて各対象者に対して経時的に応答変数を測定する状況を考える。i番目の対象者の n_i 個の応答変数を $Y_i=(y_{i1}, \dots, y_{in_i})^T$ ($i=1, \dots, N$)、各時点($j=1, \dots, n_i$)での p 個の説明変数を x_{ij} する。ここでの目的は各対象者内の応答変数間の相関を考慮した上で、平均的な応答 $\mu_i = E(Y_i)$ と説明変数 $X_i = (x_{i1}, \dots, x_{in_i})^T$ との間の関係を定量的に評価することである。

1) 一般化推定方程式(パラメータ β を求める推定方程式):

Liang and Zeger³⁾⁴⁾から提案した一般化推定方程式が下記にある。

$$\sum_{i=1}^k D_i^T V_i^{-1} S_i = 0$$

$D_i = \frac{\partial \mu_i}{\partial \beta}$ という部分はロジスティックモデルに基づいた反応変数の期待値である。

V_i^{-1} は応答変数 Y_i の作業共分散行列の逆行列で、 Y_i の真の相関構造がわからないので、作業相関構造を指定する。SASを用いて解析するとき作業共分散行列を指定する必要がある。

$S_i = y_i - \mu_i$ は応答変数の観察値と期待値との差である。

従って、以下の一般化推定方程式⁵⁾を解くことによって未知パラメータの推定を行う。

$$U(\beta) = \sum_{i=1}^k \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} (y_i - \mu_i) = 0$$

2) 作業相関構造:

推定の効率を高めるために作業相関構造行列を指定することができる。作業相関構造の仮定はいろいろあるが、以下の4種類の相関構造をよく用いる。

- 自己回帰(Autoregressive correlation)

$$\text{Corr}(y_{is}, y_{it}) = \alpha^{|s-t|} \quad (s \neq t)$$

$$\begin{pmatrix} 1 & r & r^2 & \dots & r^{n-1} \\ r & 1 & r & & r^{n-2} \\ r^2 & r & 1 & & r^{n-3} \\ \vdots & & & \ddots & \vdots \\ r^{n-1} & r^{n-2} & r^{n-3} & \dots & 1 \end{pmatrix}$$

時系列解析の時にこういう相関構造をよく使う。例えば、ガン患者の発熱をエンドポイントとし、自己回帰は時点間の相関がある発熱データなどに適切かもしれない。

- 交換可能(Exchangeable):

$$\text{Corr}(y_{is}, y_{it}) = \alpha \quad (s \neq t)$$

$$\begin{pmatrix} 1 & r & r & \cdots & r \\ r & 1 & r & & r \\ r & r & 1 & & r \\ \vdots & & & \ddots & \vdots \\ r & r & r & \cdots & 1 \end{pmatrix}$$

すべての時点間の相関は一定である。全部の組み合わせが同様な相関と考えられる。例えば、時点間の相関ではなく家族内の相関を問題にしているときに、兄弟間の相関が一番目の子供と二番目の子供の相関が r とし、一番目と三番目の子供の相関も r とする場合である。

- 独立(Independent):

$$\text{Corr}(r) = I$$

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

全ての時点間の相関はゼロであり、「独立である」とする相関構造である。独立な相関構造を仮定して、ロバスト分散を使って、検定と信頼区間を計算する場合がある。

- 制約なし(Unstructured correlation):

$$\text{Corr}(y_{is}, y_{it}) = \alpha_{st}$$

$$\begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & & r_{2n} \\ r_{31} & r_{32} & 1 & & r_{3n} \\ \vdots & & & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{pmatrix}$$

時点間の相関は何も制約をいれず、相関係数は全部違う場合である。

3) GEE の特徴⁶⁾

- GEEは一般化線形モデルを相関のあるデータに拡張したものと考えられる。
- 応答変数 $Y_i=(y_{i1}, \dots, y_{in})^T$ の同時分布を仮定せず、その周辺分布のみをモデル化している。最尤法ではなくモーメント法の考え方でパラメータ推定を行う。
- 応答変数の周辺平均に対するモデルのみが正しければ、仮定した相関構造が真の相関構造でなくても $N \rightarrow \infty$ の時、回帰係数の推定値にバイアスが入らない、つまり、 $\hat{\beta}$ の一致性は保たれる。
- 仮定した相関構造が正しくない場合でも、 $\hat{\beta}$ のロバスト分散の一致性は保たれる。

3. SAS の解析ツール

SAS のサンプルライブラリーのデータ(GEE Model for Binary Data)を少し加工し、GEE の解析を行った。SAS の解析ツールは GENMOD のプロシジャ、「SAS による貧乏人の GEE」および「GEE マクロ」がある。Windows 版 SAS システムリリース 8.2 の稼動環境で、GEE に解析用の SAS ツールを利用し、回帰係数およびロバスト分散からオッズ比およびその 95%信頼区間を算出した。

データ構造: データステップ

```
data resp;
  input idx outcome center2 active female age baseline visit int;
cards;
  101      0          0          0          0          46          0          1          1
  101      0          0          0          0          46          0          2          1
  101      0          0          0          0          46          0          3          1
  101      0          0          0          0          46          0          4          1
  102      0          0          0          0          28          0          1          1
  102      0          0          0          0          28          0          2          1
  102      0          0          0          0          28          0          3          1
  102      0          0          0          0          28          0          4          1
                                     中略
  254      0          1          0          1          39          0          1          1
  254      0          1          0          1          39          0          2          1
  254      0          1          0          1          39          0          3          1
  254      0          1          0          1          39          0          4          1
  255      0          1          1          0          25          0          1          1
  255      1          1          1          0          25          0          2          1
  255      1          1          1          0          25          0          3          1
  255      1          1          1          0          25          0          4          1
;
run;
```

- 1組の繰り返し測定が1オブザーベーションとなるようなデータを作る。この例では患者さん1人(変数 idx)につき4オブザーベーション(変数 visit (1,2,3,4))がある。
- outcome は結果変数であり、center2, active, female, age, baseline は説明変数である。
- Int: 回帰モデルの切片。全オブザーベーションで値「1」とする。GENMOD プロシジャおよび「SAS による貧乏人の GEE」では必要ないが、「GEE マクロ」では必要がある。

1) GENMOD プロシジャを用いた解析

(1) オプションの指定

GENMOD プロシジャを実行するには、オプション Dist, Type 等を指定する必要がある。上記のデータセットの解析では、Dist=Bin で、二項分布、ロジットリンク関数を指定し、Type=ind で相関構造を独立と仮定した。

- モデルで使用する組込み確率分布を指定するオプション

表1 確率分布およびリンク関数

DIST=	分布	デフォルトのリンク関数
BINOMIAL BIN B	二項分布	ロジット
GAMMA GAM G	ガンマ分布	逆数 (-1 乗)
IGAUSSIAN IG	逆ガウス分布	逆数の2乗 (-2 乗)
MULTINOMIAL MULT	多項分布	累積ロジット
NEGBIN NB	負の二項分布	対数
NORMAL NOR N	正規分布	恒等
POISSON POI P	Poisson 分布	対数

- 相関構造のタイプを指定するオプション

表2 相関構造のタイプ

キーワード	相関行列のタイプ
AR AR(1)	1次自己回帰
EXCH CS	交換可能
IND	独立
MDEP (数値)	m -従属 (m =数値)
UNSTR UN	構造化されていない (制約なし)
USER FIXED (matrix)	固定、ユーザー指定相関行列

(2) マクロプログラム

```
/******  
/* プログラム名 : [G_GEE. SAS] */  
/* 作成者 : L. Wang */  
/* 作成日 : 2003/05/05 */  
/* */  
/* 機能 : GEEによるオッズ比、95%信頼区間およびP値等の算出 */  
/* */  
/* idsn..... 入力データセット */  
/* odsn..... 出力データセット */  
/* class..... 個人を示す変数 (患者のID) */  
/* y..... 応答変数 */  
/* x..... 説明変数 (複数可) */  
/* dist..... 確率分布を指定するオプション */  
/* type..... 作業相関構造を指定するオプション */  
/* keta..... 少数点以降の桁数 */  
/******  
  
%macro G_GEE(idsn=, odsn=, class=, y=, x=, dist=, type=, keta=);  
  
/*-----回帰係数および分散を算出-----*/  
proc genmod data=&idsn descending;  
  class &class ;  
  model &y=&x / dist=&dist converge=1e-12 ;  
  repeated subject=&class/ type=&type ;  
  make 'GEEEMPPEST' out=m_wrk;  
run;  
quit;  
/*-----オッズ比および95%信頼区間を算出-----*/  
data &odsn;  
  length parm odds low_CL up_CL p_value $100;  
  label odds='オッズ比' low_CL='95%信頼区間下限' up_CL='95%信頼区間上限';  
  set m_wrk;  
  if parm="Intercept" then delete;  
  odds =put(round(exp(estimate), 0.1**&keta), 12. &keta);  
  low_CL =put(round(exp(lowerCL ), 0.1**&keta), 12. &keta);  
  up_CL =put(round(exp(upperCL ), 0.1**&keta), 12. &keta);  
  p_value=put(round(probZ, 0.1**&keta), 12. &keta);  
  keep parm odds low_CL up_CL p_value;  
run;  
  
proc print label;run;  
  
%mend G_GEE;
```

```
%G_GEE(idsn=resp, odsn=odds, class=idx, y=outcome, x=center2 active female age
baseline, dist=bin, type=ind, keta=10);
```

(3) 解析結果

Parameter	オッズ比	95%信頼区間下限	95%信頼区間上限	p_value
center2	1.9145650743	0.9580887375	3.8259080607	0.0659516524
active	3.5443526779	1.7965673145	6.9924660233	0.0002623041
female	1.1465763294	0.4837996848	2.7173173537	0.7560352346
age	0.9814184464	0.9567944118	1.0066762045	0.1479743818
baseline	6.3326556343	3.2143153611	12.476226779	0.0000000957

2) 「SASによる貧乏人のGEE」を用いた解析

第16回薬効評価研究会で紹介された「SASによる貧乏人のGEE」にあるプログラムに対して下線部分を変更した後、プログラムを実行した。その結果得られたオッズ比、95%信頼区間およびP値はGENMODプロシジャを用いた解析結果と少数第7位以上で一致した。

(1) 従来プログラム

```
①Lower=t(beta0)-rse#1.95996;
Upper=t(beta0)+rse#1.95996;
②proc logistic data=resp descending covout outest=mcov;
model outcome=center2 active female age baseline;
output out=pgee p=ey;
run;
```

(2) 変更したプログラム

```
①Lower=t(beta0)-rse#probit(0.975);
Upper=t(beta0)+rse#probit(0.975);
②proc logistic data=resp descending covout outest=mcov;
model outcome=center2 active female age baseline/converge=1e-12;
output out=pgee p=ey;
run;
```

3) 「GEE マクロ」

「GEE マクロ」では、Liang と Zeger の GEE のアプローチを使用し、同じ個人の観察時点間の相関関係を扱う回帰係数を推定するためのプログラムである。1989年にOriginal versionを作成され、1994年にUpdateし、Version2.03にされた。<http://www.statlab.uni-heidelberg.de/statlib/GEE/GEE1>から自由にGEE1_203.SASをダウンロードし、「GEE マクロ」が実行できた。「GEE マクロ」からの最終的な出力は回帰係数、回帰係数の分散(robust varianceも含む)、p-値及びオッズ比、95%信頼区間などがあつた。

(1)「GEE マクロ」を実行するために使うオプション

項目	オプション		
		内容	「GEE マクロ」
Link 関数	Identity	$g(\mu) = \mu$	Link=1
	Logarithm	$g(\mu) = \ln[\mu]$	Link=2
	Logit	$g(\mu) = \ln[\mu/(1-\mu)]$	Link=3
	Reciprocal		Link=4
平均一分散の関係	Gaussian	$\text{Var}(Y)=1 \quad \text{Var}(Y)=\sigma_y^2$	Vari=1
	Poisson	$\text{Var}(Y)=\mu$	Vari=2
	Binary	$\text{Var}(Y)=\mu(1-\mu)$	Vari=3
	Gamma	$\text{Var}(Y)=\mu^* \mu$	Vari=4
作業相関行列	Independent	$\rho=0$	Corr=1
	Stationary M-dep.		
	Non-Stationary M-dep		
	Exchangeable	$\rho \neq 0$	Corr=4
	AR-M		

(2)「GEE マクロ」

上記のデータセットを用いて、「GEE マクロ」を起動する下記の SAS プログラムにマクロのパラメータを指定し、実行した。詳しくは参考文献 2 に譲る。

```
%include 'GEE のマクロライブラリ';
GEE (data=resp,
     yvar=outcome,
     xvar=int center2 active female age baseline,
     id =idx,
     tvar=visit,
     t_values=1 2 3 4,
     link=3,
     vari=3,
     corr=1,
     iter=10,
     syms=150,
     work=330,
     missdel=yes);
```

(3) 解析結果

「GEE マクロ」中にt分布における95%点では「1.96」という値を用いられているが、それをProbit関数Probit(0.975)を用いた上で、桁数も修正した場合、オッズ比、95%信頼区間が GENMOD プロシジャを用いた解析結果と少数点7位まで一致した。

4. 終わりに

SAS のサンプルライブラリーのデータ (GEE model for binary data) を利用し、3つの解析ツールを解析した結果を比較した。3つの解析ツールを用いた解析結果は少数点7桁まで一致でしたことから、それらの精度は同じぐらいであると考えられた。なお、「SAS による貧乏人の GEE」では欠測値があっても欠測のデータを含んで解析できることに対して、「GEE マクロ」および GENMOD プロシジャでは欠測値があった場合は欠測のデータを除いてから解析している。「GEE マクロ」では SAS 社のサポート対象になっていないので、GENMOD プロシジャまたは「SAS による貧乏人の GEE」を使うことを薦める。

参考文献

1. 佐藤俊哉 酒井弘憲 酒井弘憲でもわかる GEE 第 16 回薬効評価研究会 1995.
2. 松岡 淨 GEE 第 16 回薬効評価研究会 1995.
3. Scott L. Zeger and Kung-Yee Liang. Longitudinal Data Analysis for Discrete and continuous outcomes. *Biometrics* 42,121-130 March 1986.
4. Kung-Yee Liang and Scott L. Zeger Longitudinal Data Analysis using generalized linear models *Biometrika* 73,13-22 1986.
5. Scott L. Zeger, Kung-Yee Liang and Paul S. Albert. Models for Longitudinal Data: A generalized Estimating Equation Approach. *Biometrics* 44,1049-1060 December 1988.
6. 松山 裕, 林 邦彦 佐藤俊哉 山本精一郎 大橋靖雄 Generalized Estimating Equations の理論と応用 薬理と治療 24(12):2531-2542,1996

日本SASユーザー会 (SUGI-J)

NLMIXEDプロシジャーを用いた Item Response Model のシミュレーション

○ 板東 説也* 宮岡 悦良** 緑川 修一** 高原 佳奈**

*有限会社 電助システムズ 電腦事業部

**東京理科大学 理学部 東京理科大学 大学院

Simulation Studies of Estimation in Item Response Models using NLMIXED
procedure

Etsuya Bandoh* Etsuo Miyaoka** Shuuiti Midorikawa** Kana Takahara**

*DENSUKE SYSTEMS Co.,Ltd. **Tokyo University of Science

要 旨 近年、統計モデルとして線形混合モデルが注目されてきている。SAS 8.2ではMIXEDプロシジャーを用いることで様々な事柄に対して複雑かつ詳細な解析が可能となった。しかし、我々の身の周りの出来事に関して、全てこの線形な混合モデルで処理出来る訳ではないことも事実である。即ち、非線形な場合の混合モデルをどのように処理して行くかということが問題となる。今まで非線形混合モデルを用いる場合には、IMLを用いプログラムを作成する必要があり、多大な手間と時間を費やしてきた。そんな折に、SAS System V8 から追加されたNLMIXEDプロシジャーでは、容易に非線形な混合モデルを解析することが可能となった。そこで本稿においては、このNLMIXEDプロシジャーでの推定をItem Response Modelを例に用いてシミュレーションによる検証を行ってみた。

キーワード： Item Response Model、NLMIXEDプロシジャー

1. はじめに

これから取扱うItem Response Modelであるが、われわれの身の回りには、試験・心理検査・アンケートなど様々な“テスト”が存在している。例えば、入試選抜や採用試験等である。ではその採点成績は根拠の明確な数字として考えて良いのだろうか。

そこで、これらを数値的に解析しようと開発されたのが1940～1950年代以降発展してきた項目反応理論(Item Response Theory; IRT)を中心とする現代テスト理論である。そこで被験者の能力とテストの正答率の関係についてItem Response Modelという統計的モデルをたてて考える。Item Response Modelは能力を表す変数 θ の関係として定義される。変数 θ はこのモデル上で、実数値をとる連続変数である。また θ は被験者の能力を表す指標であり、被験者個々により異なる値をとるものとする。

能力変数 θ を用いて、被験者の各項目に対する正答率を $P(\theta)$ と定めると、正答率 $P(\theta)$ は変数 θ の関数であり、0から1の間の値をとる。一般に正答率とは低い能力では低い正答率、高

い能力では高い正答率となる。したがって $P(\theta)$ は被験者個人の能力 θ に依存する単調増加関数を仮定する。よってここではロジスティック関数 (logistic function) を用いて正答率を表すことにし、正答率 $P(\theta)$ は 1 つまたは 2 つの母数を与えて次のような式で表される。

$$P(\theta) = \frac{1}{1 + \exp[-(\theta - b)]} \quad (1)$$

または

$$P(\theta) = \frac{1}{1 + \exp[-a(\theta - b)]} \quad (2).$$

ここでの a, b をそれぞれ項目の“識別力(discriminate)”と“困難度(difficulty)”と呼ぶ。 a, b はともに実数値をとる母数であり、総して項目母数(item parameter)と呼ばれる。(1)式を 1 母数ロジスティックモデル (1 parameter logistic model)、(2)式を 2 母数ロジスティックモデル (2 parameter logistic model) といい、以降それぞれを 1 母数モデル、2 母数モデルと呼ぶことにする。以後この 2 つの Item Response Model を用いて話を進めて行く。

尚、次節では、“正答”または“誤答”の 2 値変数を表す結合確率関数を定め、さらに能力母数 θ についての周辺尤度関数を求める。そして得られたモデルについて、被験者の反応データが与えられたときの項目母数の推定について述べておく。しかし項目母数、能力母数を同時に推定することは母数が増えてしまい困難であるため、項目母数にのみに限定し、最尤法により対数周辺尤度方程式を解くことで最尤推定値を得る。また項目母数についての尤度方程式は計算が困難な場合となる積分を含むことから、Gauss 求積法 (Gaussian Quadrature) により近似的な方程式を求めて行く。さらにこの方程式は非線形でありことから Dual-Quasi-Newton Method を用い、数値解として項目母数の最尤推定値を得る。

2. 推定方法

以降の節では 2 母数モデルについて述べる。1 母数モデルについては 2 母数モデルにおいて $a = 1$ とし同様に扱うことができる。

今、 N 人の被験者はそれぞれ互いに独立であり、被験者 $i (i = 1, 2, \dots, N)$ について、それぞれの項目 $j (j = 1, 2, \dots, n)$ の反応もまた互いに独立であるとする。そのとき、 N 人の被験者が n 個の項目について、反応データ $u_{i1}, u_{i2}, \dots, u_{in}, u_{21}, \dots, u_{Nn}$ を得る確率は、次のように表される。

$$P(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \underline{\theta}, \mathbf{a}, \mathbf{b})$$

尚、 $\theta = (\theta_1, \theta_2, \dots, \theta_N), \mathbf{a} = (a_1, a_2, \dots, a_N), \mathbf{b} = (b_1, b_2, \dots, b_N)$ とする。

$$= \prod_{i=1}^N P(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \theta_i, \mathbf{a}, \mathbf{b})$$

$$\begin{aligned}
&= \prod_{i=1}^N \prod_{j=1}^n P(U_{ij} = u_{ij} | \theta_i, a_j, b_j) = \prod_{i=1}^N \prod_{j=1}^n [P_j(\theta_i)]^{u_{ij}} [Q_j(\theta_i)]^{1-u_{ij}} \\
&= \prod_{i=1}^N \prod_{j=1}^n \left[\frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \right]^{u_{ij}} \left[\frac{\exp[-a_j(\theta_i - b_j)]}{1 + \exp[-a_j(\theta_i - b_j)]} \right]^{1-u_{ij}} \quad (3)
\end{aligned}$$

ただし、 u_{ij} は 1 または 0 のいずれかの値をとる。

次に、周辺尤度関数を求める。 $\theta_i (i = 1, 2, \dots, N)$ は固定された被験者 i の能力を表す母数とする。また θ は実数値をとる母数とし、 θ の確率密度関数を $\phi(\theta)$ とする。

そこで、それぞれの被験者 $i (i = 1, 2, \dots, N)$ について、それぞれの項目 $j (j = 1, 2, \dots, n)$ の反応は互いに独立であるとする、 n 個の項目に対する N 人の被験者の反応データ $u_{ij} (i = 1, 2, \dots, N; j = 1, 2, \dots, n)$ を得る確率は、(3)式より次のように定める。

$$\begin{aligned}
&P(U_{11} = u_{11}, U_{12} = u_{12}, \dots, U_{Nn} = u_{Nn} | \mathbf{a}, \mathbf{b}) \\
&= \prod_{i=1}^N P(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \mathbf{a}, \mathbf{b}) \\
&= \prod_{i=1}^N \int \phi(\theta) P(U_{i1} = u_{i1}, U_{i2} = u_{i2}, \dots, U_{in} = u_{in} | \theta, \mathbf{a}, \mathbf{b}) d\theta \\
&= \prod_{i=1}^N \int \phi(\theta) \prod_{j=1}^n P(U_{ij} = u_{ij} | \theta, a_j, b_j) d\theta \quad (4)
\end{aligned}$$

ただし、 u_{ij} は 1 または 0 のいずれかの値をとる。

さらに、周辺対数尤度関数は(4)式より次の式で定められる。

$$\begin{aligned}
&\log L(\mathbf{a}, \mathbf{b}; u_{11}, u_{12}, \dots, u_{Nn}) \\
&= \sum_{i=1}^N \log \int \phi(\theta) \prod_{j=1}^n P(U_{ij} = u_{ij} | \theta, a_j, b_j) d\theta \quad (5)
\end{aligned}$$

ただし、 u_{ij} は 1 または 0 のいずれかの値をとる。

いま θ の分布を標準正規分布と仮定する。つまり $\phi(\theta) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{\theta^2}{2}]$ と仮定すれば、

b_j についての対数尤度方程式は次の式で表される。

$$\frac{\partial}{\partial b_j} \log L(a_j, b_j; u_{1j}, u_{2j}, \dots, u_{Nj}) = 0$$

$$\sum_{i=1}^N \frac{\int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\theta^2}{2}\right] P(U_{ij} = u_{ij} | \theta, a_j, b_j) \frac{\partial}{\partial b_j} P(U_{ij} = u_{ij} | \theta, a_j, b_j) d\theta}{\int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\theta^2}{2}\right] P(U_{ij} = u_{ij} | \theta, a_j, b_j) d\theta} = 0$$

この方程式の左辺は計算が困難な積分を含むこともあることから、SAS ではそれらに対して Gauss 求積法及び適合型 Gauss 求積法を用いた推定を行っている。そこで求積法の違いや標本の大きさの違いによって推定値がどのように変化するかを見るために以下のシミュレーションを行っていく。

3. シミュレーション

ある1つの項目に対し、 N 人の被験者の反応データ u_1, u_2, \dots, u_N より項目母数について以下の Step1~Step3 で 1 母数モデルのシミュレーションを行う。

Step1: 被験者数と項目母数の真の値を設定する。そして、ある1つの項目について、被験者の能力を平均 0、分散 1 の正規分布に従うと仮定し、能力母数の擬似乱数を発生させる。

Step2: 生成した擬似乱数より、1回1回のベルヌーイ試行で各被験者の正答率に基づく擬似データの作成をする。ここでの擬似データは被験者の人数分の 0-1 のデータである。

Step3: 作成した擬似データより項目母数の推定を行う。尚、推定に用いたプログラムは、Program1 が Gauss 求積法、Program2 が適合型 Gauss 求積法として下記に記した。

```

/* Gauss 求積法 */
proc nlmixed data=item noad;
  parms b1=0.7;
  z=(x-b1);
  p=1/(1+exp(-z));
  model u~binomial(1,p);
  random x~normal(0,1) subject=i;
  ods output ParameterEstimates=pe;
run;

```

(program1)


```

/* 適合型 Gauss 求積法 */
proc nlmixed data=item;
  parms b1=0.7;
  z=(x-b1);
  p=1/(1+exp(-z));
  model u~binomial(1,p);
  random x~normal(0,1) subject=i;
  ods output ParameterEstimates=pe;
run;

```

(program2)

4. 結果

上記シミュレーションによる結果を下記、表 1~2 に記載しておく。尚、全ての結果についてはシミュレーションによって得られた推定値の平均値を記載している。

表1. 1母数モデル(推定回数の変化)

Parameter	推定回数	被験者数	真の値	推定値の平均 (Gauss 求積法)	推定値の平均 (適合型 Gauss 求積法)
B	100	500	0.5	0.7258545	0.4875879
	1,000			0.4993304	0.5062841
	5,000			0.4998722	0.5015052
	10,000			0.5007595	0.4988295

表 2. 1母数モデル(被験者数の変化)

Parameter	推定回数	被験者数	真の値	推定値の平均 (Gauss 求積法)	推定値の平均 (適合型 Gauss 求積法)
B	10,000	100	0.5	0.5073437	0.5044951
		500		0.5007595	0.4988295
		1,000		0.4993125	0.4979776

以上の結果より、1母数モデルにおける困難度bの推定値は、かなり真の値に近い結果が得られた。また Item Response Model における求積法による真の値と推定値との差については、1

母数モデルでは被験者数が少ない場合に適合型 Gauss 求積法の方が真の値に近い推定値を導きだしているという結果が得られた。また 2 母数モデルに関しては、現在様々なシミュレーションを行い調査中である。

5. まとめ

最後に、本稿では 1 母数モデルにおける被験者数及推定回数の変化による結果のみの記載に留めた。また NLMIXED プロシジャーによる疑問点及び改良点としては、Gauss 求積法による推定を行った際、推定値の値は表示されるにも関わらず、SE 等他の値が欠損値表示されている点は、実際内部でどのような処理が行われているのかという点が非常に疑問の残る点である。また現在 NLMIXED プロシジャーで用いることのできる能力母数の分布は標準正規分布のみであり、他の分布を用いることができれば、更に使用範囲が広がるのではないかと改良を期待している。

参考文献

- [1]. 赤木愛和・池田央(監訳), 教育・心理検査法のスタンダード, 図書文化社, (1993)
- [2]. Binet,A., & Simon,T., The development of intelligence in young children, The Training School, (1916).
- [3]. Dobson,A.J., 統計モデル入門, 共立出版, (1993)
- [4]. 池田央, 現代テスト理論, 朝倉書店, (1994)
- [5]. 森正武, 数値解析, 共立出版, (1973)
- [6]. Rasch,G., Probabric models for some intelligence and achievement tests
Copenhagen, Nilsen and Lydiche. (1960)
- [7]. 佐藤次男・中村理一郎, よくわかる数値計算, 日刊工業新聞社, (2001)
- [8]. 東京大学教養学部統計学教室編, 人文・社会科学系の統計学, 東京大学出版, (1994)
- [9]. 豊田秀樹, 項目反応理論 [入門編], 朝倉書店, (2002)
- [10]. Tucker,L.R., Maximum validity of a test with equivalent items, Psychometrika 11 (1946),
1-13.

- [11]. Van Der Linden,W.J., & Hambleton,R.K., Handbook of modern Item Response Theory, Springer, (1996).
- [12]. Pinheiro,J.C. and Bates,D.M, Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model, Journal of Computational and Graphical Statistics, 4, 12-35.(1995)
- [13]. 伊藤 陽一, NLMIXED プロシジャを用いた項目反応理論モデルのパラメータ推定, 日本 SAS ユーザ会 (SUGI-J), (2002)
- [14]. 伊藤 陽一・大橋靖雄, QOL 質問票における項目反応理論に対するパラメータ推定, Japanese Journal of Biometrics, Vol.23, No.1.(2002)

日本SASユーザー会 (SUGI-J)

変量効果モデルによるメタ・アナリシス

DerSimonian-Laird 法の SAS マクロの作成

○中西 豊支 浜田知久馬

東京理科大学大学院工学研究科

Developing SAS macro for meta-analysis using DL method
(random effect model)

Yushi Nakanishi and Chikuma Hamada

Graduate School of Engineering Tokyo University of Science
1-3, Kagurazaka, Shinjyuku-ku, Tokyo 162-8601

要旨

メタ・アナリシスで、研究間で効果の均一性の検定が有意な場合は、各研究の効果は変動を伴うと仮定する変量効果モデルを用いるが自然である。研究間の効果の変動の大きさをモーメント法によって推定する変量効果モデルとしては DerSimonian-Laird 法が有名である。SAS では固定効果モデルによるメタ・アナリシスは生存時間をエンドポイントとした場合 PHREG の STRATA 文を用いて可能だが、変量効果モデルを用いたメタ・アナリシスはプロシジャでは可能でない。そこで本稿では、変量効果モデルの代表である DerSimonian-Laird 法を用いて統合ハザード比を計算するための SAS マクロを示す。

キーワード： メタ・アナリシス DerSimonian-Laird 法 PHREG プロシジャ SAS マクロ

1. はじめに

メタ・アナリシスは異なった研究の結果をまとめるための手法である。通常、個々の臨床試験のサンプルサイズは十分でない場合が多く、特にがん研究において大規模臨床試験は困難である。メタ・アナリシスは類似した研究の結果を統合することによりサンプルサイズを増やし検出力を上げる。メタ・アナリシスには二つのアプローチがある。固定効果モデルと変量効果モデルである。前者は、本来効果は研究間で均一という考え方に基づく。現実として研究ごとに効果の推定値はばらつくが、それは偶然変動であると考ええる。一方、変量効果モデルでは試験ごとに効果が異なっていることが前提になる。SAS では固定効果モデルによるメタ・アナリシスは生存時間をエンドポイントとした場合 PHREG プロシジャの STRATA 文を用いて可能だが、変量効果モデルを用いたメタ・アナリシスはプロシジャでは可能でない。そこで本稿では、変量効果モデルの代表である DerSimonian-Laird 法を用いて統合ハザード比を計算するための SAS マクロを示す。

2. メタ・アナリシスのモデル

各研究から計算した effect size の推定値を適当な変換により漸近的正規近似が仮定できる状況を考える。この仮定は最尤法による推測を行っているときは妥当である。

$$\hat{\theta}_i | \theta, s_i^2 \sim N(\theta, s_i^2) \quad (1)$$

ここで $\hat{\theta}_i$ は各研究から計算した effect size の推定値を適当な変換したもの、 s_i^2 は $\hat{\theta}_i$ の推定分散である。 $\hat{\theta}_i$ の例としてハザード比あるいはオッズ比の対数変換を行うことがあげられる。

2-1 固定効果モデル

固定効果モデルは各研究効果が同一の effect size θ を持ち、均一性(homogeneity)を仮定した方法である。

帰無仮説 $H_0: \theta_1 = \dots = \theta_k$ の下では θ の対数尤度 $l(\theta) = l(\theta | \hat{\theta}_i, s_i^2)$ は

$$l(\theta) \propto Q = \sum_{i=1}^k \frac{(\hat{\theta}_i - \theta)^2}{s_i^2} = \sum_{i=1}^k w_i (\hat{\theta}_i - \theta)^2 \quad (2)$$

となる。ただし、 $w_i = \frac{1}{s_i^2}$ とした。

θ の漸近的最尤推定量 $\hat{\theta}_{AMLE}$ は

$$\hat{\theta}_{AMLE} = \frac{\sum w_i \hat{\theta}_i}{\sum w_i} \quad (3)$$

ここで、 $Var[\hat{\theta}_{AMLE}] = \frac{1}{\sum w_i}$

よって、

$$95\%CI: \hat{\theta}_{AMLE} \pm 1.96 \sqrt{\frac{1}{\sum w_i}} \quad (4)$$

また、 Q は

$$Q = \sum \frac{(\hat{\theta}_i - \theta)^2}{s_i^2} = \sum w_i (\hat{\theta}_i - \theta)^2 \sim \chi_k^2 \quad (5)$$

自由度 k の χ^2 分布に従うが Q 統計量は Q_1 と Q_2 とに分解できる。

$$\begin{aligned}
 Q &= \sum_{i=1}^k w_i \{(\hat{\theta}_i - \hat{\theta}_{AMLE}) + (\hat{\theta}_{AMLE} - \theta)\}^2 \\
 &= \underbrace{\sum_{i=1}^k w_i (\hat{\theta}_i - \hat{\theta}_{AMLE})^2}_{Q_1} + \underbrace{\sum_{i=1}^k w_i (\hat{\theta}_{AMLE} - \theta)^2}_{Q_2} \quad (6)
 \end{aligned}$$

Q_1
 χ_{k-1}^2 に従う

Q_2
 χ_{1}^2 に従う

Q_1 、 Q_2 を用いてそれぞれ研究効果の均一性の検定、有意性の検定を行うことができる。

2-2 変量効果モデル

固定効果モデルでは θ_i を共通と考えたが、現実には各研究効果 θ_i がバラツキ τ^2 を伴う（プロトコルの違い、患者の違い、地域の違い、研究者の違いなど）と考えた方が自然である。そこで、この不均一性(heterogeneity)をモデル化した一つの自然なモデルとして

$$\theta_i | \theta, \tau^2 \sim N(\theta, \tau^2) \quad i = 1, 2, \dots, k \quad (7)$$

という変量効果モデルを考える。この仮定の下 (1) 式は

$$\hat{\theta}_i | \theta, s_i^2, \tau^2 \sim N(\theta, s_i^2 + \tau^2) \quad i = 1, 2, \dots, k \quad (8)$$

と置き換えられる。変量効果モデルでは θ 、 τ^2 を推定するために周辺尤度を最大化する制限付き最尤法 (REML 法) を考えるのが自然である。 τ^2 の推定に関する対数尤度

$l(\theta, \tau^2) = l(\theta, \tau^2 | \hat{\theta}_i, s_i^2)$ は

$$l(\theta, \tau^2) \propto - \left[\sum_{i=1}^k \left(\frac{(\hat{\theta}_i - \theta)^2}{s_i^2 + \tau^2} + \log(s_i^2 + \tau^2) \right) + \log \left(\sum_{i=1}^k \frac{1}{s_i^2 + \tau^2} \right) \right] \quad (9)$$

となる。ここで、重み変数 w_i^* を

$$w_i^* = \frac{1}{s_i^2 + \tau^2} \quad \text{とおくと}$$

$$\theta_{REML} = \frac{\sum \hat{\theta}_i w_i^*}{\sum w_i^*} \quad (10)$$

$$95\%CI : \theta_{REML} \pm 1.96 \sqrt{\frac{1}{\sum w_i}} \quad (11)$$

ともとまる。ただし、

$$\hat{\tau}^2 = \frac{\sum w_i^* \left(\frac{k}{k-1} (\hat{\theta}_i - \theta_{REML})^2 - s_i^2 \right)}{\sum w_i^*} \quad (12)$$

である。(12)式の右辺は w_i^* が τ^2 を含むため $\hat{\tau}^2$ を求めるためには反復計算が必要である。

2-3 DerSimonian-Laid 法

REML 法では反復計算が必要である。一方、均一性の検定統計量 Q_1 を利用したモーメント法を適用すると、繰り返し計算の必要がない推定値が得られる。

$$\begin{aligned} Q_1 &= \sum_{i=1}^k w_i (\hat{\theta}_i - \theta_{AMLE})^2 \\ &= Q - Q_1 \\ &= \sum w_i (\hat{\theta}_i - \theta)^2 - (\sum w_i) (\hat{\theta}_{AMLE} - \theta)^2 \end{aligned} \quad (13)$$

となるので、

$$\begin{aligned} E(Q_1) &= \sum w_i \text{Var}[\hat{\theta}_i] - (\sum w_i) \text{Var}[\theta_{AMLE}] \\ &= \sum w_i \text{Var}[\hat{\theta}_i] - (\sum w_i) \left(\frac{1}{\sum w_i} \right)^2 \text{Var}[\sum w_i \hat{\theta}_i] \\ &= \sum w_i \text{Var}[\hat{\theta}_i] - (\sum w_i) \left(\frac{1}{\sum w_i} \right)^2 \{ w_1^2 \text{var}(\hat{\theta}_1) + w_2^2 \text{var}(\hat{\theta}_2) + \dots + w_k^2 \text{var}(\hat{\theta}_k) \} \\ &= \sum w_i \left(\frac{1}{w_i} + \tau^2 \right) - (\sum w_i) \left(\frac{1}{\sum w_i} \right)^2 \left(\sum w_i + \tau^2 \sum w_i^2 \right) \\ &= \sum w_i \left(\frac{1}{w_i} + \tau^2 \right) - \left(\frac{1}{\sum w_i} + \frac{\tau^2 \sum w_i^2}{(\sum w_i)^2} \right) (\sum w_i) \\ &= (k-1) + \tau^2 \left(\sum w_i - \frac{\sum w_i^2}{\sum w_i} \right) \end{aligned} \quad (14)$$

となる。そして、 Q_1 がその期待値である(14)式と等しくなるように $\hat{\tau}^2$ を推定するとモーメント推定量が次のように計算される。

$$\hat{\tau}^2 = \max \left\{ 0, \frac{Q_1 - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}} \right\} \quad (15)$$

よって、(15)式でもとめた $\hat{\tau}^2$ を用いて

$$\theta_{DSL} = \frac{\sum \hat{\theta}_i w_i}{\sum w_i} \quad (16)$$

$$95\%CI : \theta_{DSL} \pm 1.96 \sqrt{\frac{1}{\sum w_i}} \quad (17)$$

DerSimonian-Laird 法による統合した効果の推定値は(16)式、信頼区間は(17)式となる。

3. マクロの計算 (DerSimonian-Laird 法)

本マクロは各研究の個別の生存時間データからハザード比を推定し DerSimonian-Laird 法により統合するものである。

計算手順を以下に示す。

-
1. 各研究の効果を漸近分散法の基づく固定効果モデルで推定する。
 2. 均一性の検定統計量 Q_1 を計算する。
 3. 研究間のバラツキの大きさ τ^2 を推定する。
 4. 各研究の重みを計算する。
 5. 重み付き平均により統合した推定値をもとめる。
 6. 統合ハザード比の 95%信頼区間を計算する。
 7. 統合ハザード比の有意性の検定を行う。
-

4. プログラムの開発

3 節で示した計算手順を行うための SAS マクロを開発した。個別データについて、研究施設、打ち切り情報、生存時間、治療情報をまとめたデータセットを用意する。また、本稿ではハザード比を統合する過程を示す。

1. 各研究の効果の推定値を求める。

コックスの比例ハザードモデルを用いて各研究の対数ハザード比をもとめる。(浜田 1995 参照)

```
*input1=入力データセット1;
%macro isuitei(input1);proc sort data=&input1;by study;
*---- 各研究の効果(logHRi)とその分散(VlogHRi)の推定-----;
proc phreg data=&input1 outest=estby covout ;
model time*censor(1)=treat/rl;by study;
proc transpose data=estby out=estby;var treat:id _type_;by study;
data d1;set estby;
file 'data.dat';put parms cov;
%mend isuitei;
%isuitei(integral);
```

```
***** 入力データセット1 *****
study:研究
time : 生存時間
censor : 打ち切り情報 (1:生存、途中打ち切り 2:死亡)
treat : 治療法(1:control 2:治療群)
*****
```

実行例

次に示す入力データセット(integral)を用いて実際にメタ・アナリシスを行った結果を示す。研究数(study)は14であり、解析対象(ID)は10225例である。

入力データセット[integral] (一部)

ID	study	time	...	censor	treat	...
1	A	5		1	1	
2	A	4.5		2	0	
3	B	3.8		1	0	
4	D	3.2	...	2	1	...
.	
.	
.	

結果

OBS	study	PARMS	COV
1	A	-0.01781	0.01951
2	B	0.03073	0.01094
3	C	-0.90903	0.19206
4	D	0.07822	0.01854
5	E	-0.17745	0.01529
6	F	-0.30461	0.01637
7	G	-0.20787	0.10648
8	H	-0.43480	0.01110
9	I	-0.13118	0.00887
10	J	0.11171	0.11142
11	K	-0.26999	0.01260
12	L	-0.11993	0.01402
13	M	-0.14921	0.01413
14	N	-0.98747	0.08537

各研究の対数ハザード比

対数ハザード比
の推定分散

マクロを実行することにより各研究の効果の大きさ(PARMS)と分散(COV)を含んだデータセットが作成される。

2. 均一性の検定統計量 $Q_1 = w_i \times (\log HR_i - \log HR)$ 、研究間のバラツキの大きさ τ^2 、重み付き平均 HR_{DSL} を計算し、95%信頼区間をもとめ、最後に有意性の検定を行う。

*****マクロDSL*****

```

*input2=入力データセット2;
%macro DSL(input2);
data weight;set &input2;dummy=1;
*---- [漸近分散法に基づく均質性の統計量(Q1)を計算]-----;
*---- 各研究の重み(wi)を計算-----;
wi=1/(VlogHRi);
wlogHRi=wi*logHRi;
proc summary;var wi wlogHRi;output out=sumweight sum=;
*---- 漸近分散法に基づく統合ハザード比(HR)の推定---;
data HR;set sumweight; logHR=wlogHRi/wi;
HR=exp(logHR);
data cul;set HR;dummy=1;logHRv=logHR;keep logHRv dummy;
data cul;merge weight cul;by dummy;
*---- 漸近分散法に基づく均質性の検定--;
data Q1;set cul;q1=wi*(logHRi-logHRv)**2;

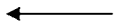
```

```

proc summary;var q1 dummy;output out=outQ1 sum=Q1 k;
*---- [変量モデル:DerSimonian-Laird法に基づく統合ハザード比の推定]-----;
data DSL; set weight;wi2=wi**2;
proc summary;var wi wi2;output out=sumweight2 sum=w w2;
*---- 研究間のバラツキの大きさ(tau2)の推定---;
data tau2; merge outQ1 sumweight2;
tau2=(Q1-(k-1))/(w-(w2)/w);
if tau2>=0 then tau2=tau2;
      else tau2=0;

```

均質性の検定
統計量 Q1



τ^2 の計算



```

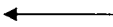
data cul;set tau2;dummy=1;
*----各研究の重み(wiDSL)---;
data wiDSL;merge &input2 cul;by dummy;
wiDSL=1/(VlogHRi + tau2);keep wiDSL dummy;
*----統合ハザード比(HRDSL)の推定---;
data HRDSL;merge &input2 wiDSL;by dummy;
wlogHR=wiDSL*logHRi;
proc summary ;var wlogHR wiDSL;output out=sumHRDSL sum=wlogHR wDSL;
data HRDSL;set sumHRDSL;
logHRDSL=wlogHR/wDSL;
HRdsl=exp(logHRDSL);
*----有意性の検定---;
Q2=logHRDSL**2*wDSL;
p=1-probchi(Q2,1);
*---- (信頼区間の推定) --;
risk=exp(logHRDSL);
risklower=exp(logHRDSL-1.96*(1/wDSL)**.5);
riskupper=exp(logHRDSL+1.96*(1/wDSL)**.5);
%mend DSL;
%DSL(data);

```

統合ハザード比



95%信頼区間



```

***** 入力データセット2 *****
[本プログラムではマクロisuiteiによってもとめている]
logHRi: 第i試験の効果の推定値 [本プログラムでは対数ハザード比(PARMS)]
VlogHRi: 第i試験の効果の推定値の分散[本プログラムではCOV]
*****

```

実行例

入力データセット

ここでは、マクロ `isuitei` によって各研究の効果を推定したデータセット `data` を用いる。マクロ DSL の実行結果は次のようになる。

結果

OBS	Q2	p	risk	risklower	riskupper
1	10.7090	.001066174	0.83099	0.74377	0.92844

参考のため、固定効果モデルによるメタ・アナリシスを行った SAS プログラムとその実行結果を示す。PHREG プロシジャにおいて研究施設を STRATA 文で指定している。

```
*****固定効果モデルによるメタ・アナリシス*****  
proc phreg data=inputdata outest=estby covout ;  
model time*censor(1)=treat/rl:strata study;  
*****
```

結果

Analysis of Maximum Likelihood Estimates			
Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits	
treat	0.840	0.783	0.901

変量効果モデルによるメタ・アナリシスは固定効果モデルと比べてバラツキ τ^2 を考慮する分、信頼区間が広がること分かる。

6. まとめ

本稿ではエンドポイントを生存時間とし、対象データとして個別データが得られることを想定した DerSimonian-Laird 法を用いて統合ハザード比をもとめる SAS マクロを示したが、マクロ DSL の入力内容を各研究のハザード比からオッズ比に変更することによりハザード比だけでなくオッズ比等の幅広い指標を統合することができる。このように簡単なプログラムによりメタ・アナリシスは実施できるがメタ・アナリシスの宿命上、結果の解釈は簡単ではない。メタ・アナリシスによって得られた結果はあくまで探索的な解析で新たな研究によって検証が必要がある。

参考文献

- 丹後俊郎 (2002) メタ・アナリシス入門 朝倉書店
- 大橋靖雄、浜田知久馬 (1995) 生存時間解析 東京大学出版会
- 竹内啓、市川伸一、大橋靖雄、岸本淳司、浜田知久馬 (1993) SASによるデータ解析入門 東京大学出版会
- 丹後俊郎 (2000) 統計モデル入門 194-195 朝倉書店
- 浜田知久馬 (1995) SASによるメタ・アナリシス SUGI-J 241-254

日本SASユーザー会 (SUGI-J)

メタ・アナリシスにおける公表バイアスの評価 trim-and-fill 法の SAS マクロの作成

○松岡 伸篤 浜田 知久馬
東京理科大学大学院工学研究科経営工学専攻

Evaluation of publication bias in meta-analysis
Developing SAS macro for “trim-and-fill” method

○Nobushige Matsuoka and Chikuma Hamada
Graduate School of Engineering Tokyo University of Science
1-3,kagurazaka, Shinjyuku-ku, Tokyo 162-8601

要 旨

メタ・アナリシスにおける公表バイアスの影響を評価する手法として Duval and Tweedie(2000a,b)により trim-and-fill 法が提案されている。trim-and-fill 法はメタ・アナリシスの対象研究を逐次的に削除して統合効果を推定するため、SAS でプログラムを作成する際、異なったデータセットで同じ作業を繰り返す必要がある。そこで、trim-and-fill 法を DATA ステップ、PROC ステップなどを組み合わせて、1つの手続きとして実施するため SAS のマクロを作成した。

キーワード：メタ・アナリシス 公表バイアス trim-and-fill 法 funnel プロット SAS マクロ

1. はじめに

メタ・アナリシスを行なう際、非常に大きな問題となるのが公表バイアス(publication bias)である。薬剤の有効性を評価する研究を行なった際、有意な結果が得られなかった研究の結果は投稿されにくく、投稿されたとしても受理されにくい傾向がある。そのため、公表されている複数の独立な研究結果にのみ基づいてメタ・アナリシスを行なうと、結果が有意な方向に偏る。この偏りが公表バイアスであり、メタ・アナリシスを行なう際には公表バイアスの影響について検討する必要がある。そこで、本論文では、funnel プロットに基づき公表バイアスの影響を評価する trim-and-fill 法の SAS マクロを作成した。

2. funnel プロット

公表バイアスを視覚的に検討する方法として funnel プロットと呼ばれる散布図が用いられる。funnel プロットとは、

- 横軸： オッズ比、ハザード比などの効果の推定値
- 縦軸： 効果の推定値の推定精度(標本サイズ または 推定値の標準誤差の逆数)

として、メタ・アナリシスの対象としている各研究の結果をプロットしたものである。標本サイズが大きいと推定値のばらつきは小さく、真の効果に近くなる。一方、標本サイズが小さいときは、公表バイアスが存在しなければ真の効果を中心に左右対称にばらつくため、全体として funnel プロットは漏斗(funnel)を逆さまにしたような左右対称形になる。

次の図 1、2は横軸を対照群に対する薬剤群のハザード比とした funnel プロットである(中央より右は薬剤効果なし、左は薬剤効果あり)。

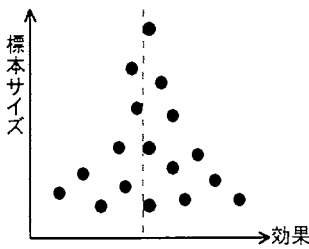


図 1: 公表バイアスなし

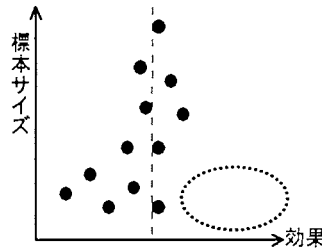


図 2: 公表バイアスあり

図 1 の funnel プロットは左右対称に近いので、公表バイアスは見られないと判断できるが、図 2 の funnel プロットは右側の点が欠け(点線で囲んだ部分)、左右非対称になっている。つまり、ネガティブな結果となった研究が公表されていないと考えられる。したがって、公表バイアスの影響が懸念される。

3. trim-and-fill 法

trim-and-fill 法とは公表バイアスが存在しなければ funnel プロットが左右対称になるという特性を前提としている。funnel プロットを左右対称にするために必要な未公表研究の数を Duval and Tweedie 法により計算し、仮想的に左右対称になるように研究を追加することにより、公表バイアスの影響を除いた上でメタ・アナリシスによって統合した効果を推定しなおすことができる。具体的には、funnel プロットが対称形になるまで繰り返し研究を削除(trim)して未公表研究数を推定し、推定された研究数だけ funnel プロットが左右対称になるように配置(fill)して未公表の研究の点と見なす(impute)方法である。

3. 1. Duval and Tweedie 法

この方法は、funnel プロットから未公表研究数を推定する方法である。ここで、手順の説明のために以下のような仮定を設ける。

メタ・アナリシスの解析対象として K 個の研究を収集した時、funnel プロットの右端の点より k_0 個の研究が未公表となっている(公表バイアスが存在する)とする。また、各研究の効果の推定値を $\hat{\theta}_i$ とし、真の効果 θ が既知であるとする。

このとき、

$$y_i = \hat{\theta}_i - \theta$$

とおき、絶対値 $|y_i|$ の順位を r_i とし、次の 2 つの統計量を定義する。

- γ : y_i が負かつ最大の順位 K まで連なっている連の長さ
- $T_K = \sum_{y_i < 0} r_i$ (符号付 Wilcoxon 順位和)

この 2 つの統計量を利用した未公表研究数 k_0 の推定量として以下の 3 つが Duval and Tweedie により提案されている。

$$R_0 = \gamma - 1$$

$$L_0 = \frac{4T_K - K(K+1)}{2K-1}$$

$$Q_0 = K - \frac{1}{2} - \sqrt{2K^2 - 4T_K + \frac{1}{4}}$$

なお、これらは研究数の推定量なので最も近い整数に丸める必要がある。

※ これらの導出については参考文献[3]参照。

※ 3つの統計量のどの値を用いるかについては、 k_0 が K の25%以上と想定される場合には L_0 、そうでなければ R_0 を用いることが推奨されている。(参考文献[3]参照)

また、funnelプロットの左方向がネガティブな結果である場合には

- γ : y_i が正かつ最大の順位 K まで連なっている連の長さ
- $T_K = \sum_{v_i > 0} r_i$ (符号付 Wilcoxon 順位和)

と変更する。

3. 2. trim-and-fill 法のアルゴリズム

trim-and-fill 法は前述したように Duval and Tweedie 法を用いて、「左右対称な funnel プロット」を仮想的に再生する方法である。Duval and Tweedie 法では真の効果 θ を既知と仮定しているため、漸近的一致性が成り立つ。しかし、現実には真の効果というものは未知の値である。そこで、trim-and-fill 法では変量効果モデルである DerSimonian-Laird 法により推定した統合効果を真の効果の推定値として用い、反復収束法を適用する。アルゴリズムは次のようになる。

Step 1. 初期推定値 $\hat{\theta}^{(1)}$ を DerSimonian-Laird 法により推定し、 $y_i^{(1)} = \hat{\theta}_i - \hat{\theta}^{(1)}$ とし、Duval and Tweedie 法により $\hat{k}_0^{(1)}$ を推定する。 k_0 が K の25%以上と想定される場合には L_0 、そうでなければ R_0 を \hat{k}_0 として本マクロでは用いる。

Step 2. funnelプロットの最左端の方に位置する点から $\hat{k}_0^{(1)}$ 個の点を除く(trim)。残りの点から同様に $\hat{\theta}^{(2)}$ を推定し、 $y_i^{(2)} = \hat{\theta}_i - \hat{\theta}^{(2)}$ とおいて $\hat{k}_0^{(2)}$ を推定する。

Step 3. 同様の作業を $\hat{\theta}^{(j-1)} = \hat{\theta}^{(j)}$ となるまで繰り返す。つまり、 $\hat{k}_0^{(j)} = 0$ となった時点で終了となる。最終的に除かれた点の総数を $\hat{k}_0 (= \sum_{j=1}^J \hat{k}_0^{(j)})$ とおく。

Step 4. 左端の最大値から \hat{k}_0 個のデータを $\hat{\theta}^{(j)}$ の回りに左右対称に配置(fill)して、対称な funnel プロットを再生する。推定誤差の値は対称なデータと同じ値を採用する。

Step 5. 再生された $K + \hat{k}_0$ 個のデータに基づいて統合効果を推定しなおす。

4. trim-and-fill 法の SAS マクロの作成

4. 1. プログラムの作成方針

4.2 節で示したように、trim-and-fill 法は「統合効果を DerSimonian-Laird 法で推定⇒未公表研究数を推定し、点を削除⇒統合効果を DerSimonian-Laird 法で推定⇒・・・」という作業を反復しなければならない。したがって、SASにより trim-and-fill 法を実行するためには複数のプロシジャを1つのまとまりとして登録し、反復するマクロを作成する必要がある。trim-and-fill 法では解析対象となる研究数が逐次的に変化するが、本マクロではこの履歴が追えるように、各段階ごとに異なったデータセット名を与えた。また、各段階での結果を funnel プロットとして図示した。

次節で、作成したプログラムを示し、解説する。

4. 2. SAS マクロを用いた trim-and-fill アルゴリズム

プログラムの構造を示す。trim-and-fill アルゴリズムをマクロ(マクロ名:TandF)として定義する。さらにマクロ TandF の内部に DerSimonian-Laird 法をマクロ(マクロ名:DSL)として定義する。なお、もう1つマクロ(マクロ名:names)を定義するが、このマクロは Step 3(3.2 節)での最終的に取り除かれた点を計算するためである。

入力データセットの変数は

- 各研究の効果(ハザード比)の推定値の対数をとったもの
- 各研究の効果の推定値の対数をとったものの分散

の2つを含み、それぞれの変数名は logHRi と VlogHRi と固定する。

以下に、trim-and-fill アルゴリズムのプログラムを示す。

```

/***** Trim and Fill 法のマクロ *****/
*----変数の説明----*;
* N      : 研究数;
* i      : 入力データセット No.(初期入力データセットを1番とする。);
* input  : 入力データセット名;
* HRi    : 各研究での効果の推定値(本論文ではハザード比);
* HRdl   : DerSimonian-Laird 法により統合効果の推定値(本論文ではハザード比);
*なお、入力データセットの変数は(logHRi , VlogHRi);
/*****/

```

```

%macro TandF(N, input);
%do i=1 %to &N;
%macro DSL(input, i);
data &input&i;set &input&i;dumy=1;
data weight;set &input&i;dumy=1;
/*---- 漸近分散法に基づく均質性の統計量(Q1)を計算 ----*/
/*---- 各研究の重み(wi)を計算 ----*/
wi=1/(VlogHRi);
wlogHRi=wi*logHRi;
proc summary;var wi wlogHRi;output out=sumweight sum=;
/*---- 漸近分散法に基づく統合ハザード比(HR)の推定 ----*/
data HR; set sumweight;
logHR=wlogHRi/wi;
HR=exp(logHR);

data cul;set HR;dumy=1;
logHRv=logHR;
keep logHRv dumy;
data cul;merge weight cul;by dumy;
/*---- 漸近分散法に基づく均質性の検定 ----*/
data Q1;set cul;
q1=wi*(logHRi-logHRv)**2;
proc summary;
var q1 dumy;output out=outQ1 sum=Q1 k;
/*---- 変数モデルに基づく統合ハザード比の推定 ----*/
data DSL;set weight;
wi2=wi**2;
proc summary;var wi wi2;output out=sumweight2 sum=w w2;
/*---- 研究間のバラツキの大きさ(tau2)の推定 ----*/
data tau2;merge outQ1 sumweight2;
tau2=(Q1-(k-1))/(w-(w2)/w);
if tau2>=0 then tau2=tau2;
else tau2=0;
data cul;set tau2;dumy=1;
/*---- 各研究の重み(wiDSL) ----*/
data wiDSL;merge &input&i cul;by dumy;
wiDSL=1/(VlogHRi+tau2);keep wiDSL dumy;
/*---- 統合ハザード比(HRDSL) ----*/
data HRDSL;merge &input&i wiDSL;by dumy;
wlogHR=wiDSL*logHRi;
proc summary;
var wlogHR wiDSL;
output out=sumHRDSL sum=wlogHR wDSL;
data HRDSL;dumy=1;set sumHRDSL;

```

①マクロ TandF を定義し、内部にマクロ DSL を定義する。

```

logHRDSL=wlogHR/wDSL;
HRdl=exp(logHRDSL);
CIU=exp(log(HRdl)-1.96*(1/wDSL)**.5);
CIL=exp(log(HRdl)+1.96*(1/wDSL)**.5);
call symput('HRdl',HRdl);
proc print data=HRDSL;run;

```

②DerSimonian-Laird 法
による統合効果(HRdl)の
推定結果

```

/*---- funnel plot の作成 ----*/
data funnel;merge &input&i HRDSL;by dummy;
se=sqrt(VlogHRi);
se2=1/se;
HRi=exp(logHRi);
if 1<=_n<=14 then dummy=1*dummy;
else dummy=dummy*2;
data funnel;set funnel;
title1 h=5 c=blue f=swissb 'funnel plot';
axis1 label=(f=swissb h=4 a=90 '1/Standard error');
axis2 label=(f=swissb h=4 'Hazard ratio')
      order=0.3 to 1.4 by 0.1;
legend1 label=none
      value=(h=3 c=black);
proc print data=funnel;run;
proc format;
value dummy 1='original data' 2='imputed data';
run;
proc gplot data =funnel;
plot se2*HRi=dummy/href=&HRdl
      href=1 LH=(2) } ③
      frame
      legend=legend1
      vaxis=axis1
      haxis=axis2;

```

③funnel プロット上に統合効果の推定
値を点線で表示。

```

format dummy dummy.;
symbol1 v=dot h=3;
symbol2 v=circle h=3;
%mend DSL;
%DSL(&input,&i);
/*---- tirm and fill アルゴリズム ----*/
data dsyi;merge HRDSL funnel;by dummy;
data dsyi; set dsyi;
yi=HRi-HRdl;
y=abs(yi);
proc rank data=dsyi out=rankdata;
var y;
ranks wscore;
/*---- Duval and Tweedie 法による R0 の計算 ----*/
data stepR1;set rankdata;
if yi>0 then wscore=1*wscore;
else wscore=-1*wscore;
data stepR2;set stepR1;
keep wscore;
proc transpose data=stepR2 out=stepR3;
proc means data=stepR3;
var wscore;
output out=stepR2 min=min max=max;
data dsR0;set stepR2;
gamma=-1*(min+max);
R0=gamma-1;
/*---- Duval and Tweedie 法による L0 の計算 ----*/
data stepL1;set rankdata;
if yi>0 then Tk=0*wscore;
else Tk=1*wscore;
proc summary ;var Tk dummy;output out=stepL2 sum=Tk K;
data dsL0;set stepL2;
L0=(4*Tk-K*(K+1))/(2*K-1);
L=int(L0);
if abs(L0-L)<=0.5 then L=L;

```

④マクロ DSL の定義終了と呼び出し。
do ループで i=1 からとしているので、初期
入力データセット(inputdata1)から入力

```

else L=L+1;
/*---- Duval and Tweedie 法による Q0 の計算 ----*/
data dsQ0;set stepL2;
Q0=K-1/2-sqrt(2*(K**2)-4*Tk+1/4);
/*---- k0 の決定 ----*/
data dsk&i;merge dsR0 dsL0;
if L/K>=0.25 then call symput('k0',L);
else call symput('k0',R0);
data dsk&i;set dsk&i;
k0=&k0;
if &k0=0 %then %goto out;
proc print data=dsk&i;
proc sort data=inputdata&i;by logHRi;
/*---- k0 個の点を削除(trim) ----*/
data &input&eval(&i+1);merge &input&i dsk&i;
if 1<=_n_<=&k0 then delete;
keep dummy logHRi VlogHRi;
&end;
out;

```

⑤

⑤if~then goto 文により、k0 が 0 となった時点で do ループから脱出。

⑥

⑥k0 個だけ研究を削除し、入力データセットを逐次的に更新する。

```

%macro names(name,number);
do n=1 %to &number;
name&n
&end;
%mend names;

data dsk;
set %names(dsk,&i);
keep k0;
/*---- 最終的に取り除かれた点の総数 ----*/
proc means data=dsk;
var k0;
output out=dsk sum=sumk0;
data dsk;dummy=1;set dsk;
proc print data=dsk;

```

```

/*---- 未公表論文の imputed value のハザード比を計算 ----*/
%let i=1;
data impute;merge &input&i dsk;by dummy;
if 0<=_n_<=sumk0;
keep dummy logHRi VlogHRi;
data impute2;set impute;
HRi=exp(logHRi);
data impute3;merge HRDSL impute2;by dummy;
x=HRd1-HRi;
HRi=x+HRd1;
logHRi=log(HRi);
keep HRi logHRi VlogHRi dummy;

```

```

/*---- 未公表論文を再生したデータセット(filldata)の作成 ----*/
data filldata;set &input&i impute3;
keep logHRi VlogHRi dummy;
/*---- 統合ハザード比を推定し直すために再生したデータセット(filldata)をマクロ DSL に入力 ----*/
%DSL(filldata);
%TandF(14,inputdata);
proc print;run;

```

⑦

⑦マクロ TandF の定義終了。
解析対象の研究数と入力データセット名を指定し、マクロ TandF を呼び出す。

5. プログラムの実行例

解析対象研究数が 14 研究、効果指標がハザード比であるデータセット(データセット名:inputdata1)に対して、trim-and-fill 法の SAS マクロを実行した。プログラムを実行するには、4.2 節の⑦で示したように研究数と入力データセット名をマクロ TandF において指定すればよい。

5. 1. 入力データセット

次の表 1 に示す初期入力データセット(inputdata1)を用いて、4.2 節の SAS マクロを実行する。

表 1: 初期入力データセット(inputdata1)

OBS	logHRi	VlogHRi
1	-0.98747	0.08537
2	-0.90903	0.19206
3	-0.43617	0.01103
.	.	.
.	.	.
.	.	.
14	0.11171	0.11142

このデータセットにおいて logHRi が小さいものから各反復段階で推定された k_0 個だけ、逐次的にデータが削除される。

5. 2. 結果の出力

各反復段階ごとの結果の出力を示す。

(1) 14 研究での結果

- データセット HRDSL (統合ハザード比の推定結果)

OBS	dummy	...	_FREQ_	...	logHRDSL	HRdl	CIU	CIL
1	1	...	14	...	-0.18620	0.83011	0.74322	0.92717

- データセット dsk1 (k_0 の推定値)

OBS	...	_FREQ_	...	R0	...	L0	...	k0
1	...	14	...	1	...	-0.22222	...	1

- 14 研究での funnel プロット

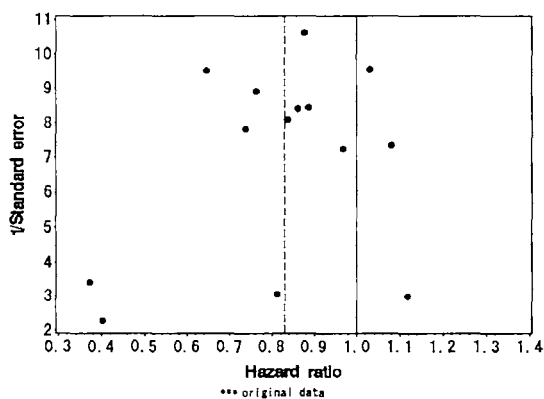


図 3: 14 研究での funnel プロット

(2) 14 研究から1研究除いた 13 研究での結果

- データセット HRDSL

OBS	dummy	...	_FREQ_	...	logHRDSL	HRdl	CIU	CIL
1	1	...	13	...	-0.16121	0.85112	0.77286	0.93729

- データセット dsk2

OBS	...	_FREQ_	...	R0	...	L0	...	k0
1	...	13	...	0	...	-0.4	...	0

※ k_0 が 0 と推定されたので、反復終了となる。

- 13 研究での funnel プロット

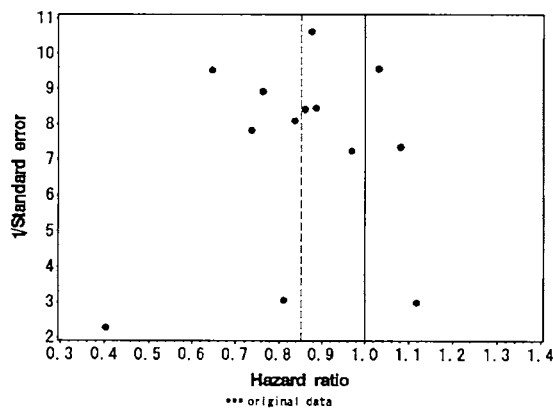


図 4: 13 研究での funnel プロット

(3) 14 研究に1研究追加した 15 研究での結果

- データセット HRDSL

OBS	dummy	...	_FREQ_	...	logHRDSL	HRdl	CIU	CIL
1	1	...	15	...	-0.17294	0.84119	0.75263	0.94017

- 再生された 15 研究での funnel プロット

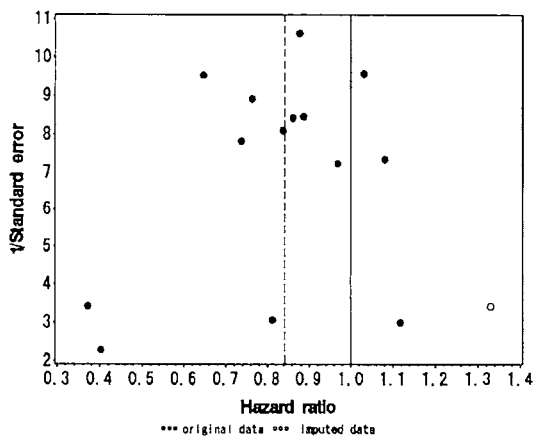


図 5: 再生された 15 研究での funnel プロット

5. 3. 結果のまとめ

表 1 に trim-and-fill 法を適用した結果をまとめる。

表 1: trim-and-fill 法の結果

研究数	ハザード比	95%信頼区間
K=14	0.830	0.743~0.927
K=13	0.851	0.773~0.937
K=12	0.841	0.753~0.940

未公表研究数は 1 研究と推定された。再生する前の 14 研究での統合ハザード比が 0.830 であったのに対して、再生された 15 研究での統合ハザード比を推定しなおすと 0.841 修正されたが違いは大きくない。この結果から、今回対象としたデータは公表バイアスの影響が小さく、また公表バイアスの影響を考慮しても統合効果の推定値はほとんど影響を受けないといえる。

6. まとめ

trim-and-fill 法は公表バイアスの有無のみならず、バイアスの影響を調整した上で統合効果を推定しなおすことにより公表バイアスの影響を評価できる。trim-and-fill 法は反復作業を行なうので、SAS マクロによりプログラムを作成した。本マクロでは、4.2 節⑦に示したように研究数と入力データセット名の指定のみ行なうことにより、メタ・アナリシスの解析対象に対して公表バイアスを評価することができる。また、今回はメタ・アナリシスの効果指標としてハザード比を対象としたが、入力データセットで効果の指標をオッズ比等に変更することにより、他の指標を用いたメタ・アナリシスにおける公表バイアスの評価も可能である。

【参考文献】

- [1] 丹後 敏郎、2002 “メタ・アナリシス入門。” 朝倉出版。
- [2] Sue Duval and Richard Tweedie.,2000 ”Trim-and-fill:A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis” Biometrics,56,pp.455-463
- [3] Sue Duval and Richard Tweedie.,2000 “A Nonparametric “Trim and Fill” Method of Accounting for Publication Bias in Meta-Analysis” Journal of American Statistical Association, Vol.95, No.499, pp.89-97
- [4] 松村智恵子・余田昭夫・田崎武信、2001 “trim-and-fill 法によって公表バイアスを探る” 癌臨床研究・生物統計誌 Vol.21,No.1,July2001, pp.25-38
- [5] 竹内啓 監修、市川伸一・大橋靖雄・岸本淳一・浜田知久馬 著、1993 “SAS によるデータ解析入門[第 2 版].” 東京大学出版会。

ポスターセッション
統計教育

看護系大学における疫学・生物統計学教育の実態調査

田中 司朗

東京大学医学系研究科生物統計学

Research on education of Epidemiology and Biostatistics

Shiro Tanaka

Biostatistics, School of Health Sciences and Nursing, University of Tokyo.

要旨 看護系大学における疫学・生物統計学教育の実態について、国公立大学62校及び私立大学27校を対象に、自記式調査票を用いて調査を行った。調査の目的は、担当教官の背景・教官が授業を行っていく上で問題と感じている点・講義と実習の内容を把握し、教官の背景と抱えている問題点の関連性を検討する事とした。対象校89校のうち50校から61通の回答が得られ、担当教官の専門分野や所属学会などの背景、抱えている問題点、講義・実習・卒業論文指導の内容などが明らかになった。特に、疫学・生物統計学を専門としている教官が講義している大学は少なく、工学部・薬学部・理学部数学科などの他学部所属の教官に頼っている事や教科書に対する要望が強い事、学生に学ぶ意欲や数学とパソコン・情報処理の能力が足りないと感じている教官が多い事が示された。

キーワード： 疫学、生物統計学、教育、調査票

1. はじめに

疫学は特定の集団における健康に関連する状況あるいは事象の、分布あるいは規定因子に関する研究を行う学問であり、生物統計学は医学研究におけるデータの取り方、解釈の方法などを考える応用統計学である。質の高い医療を提供するために科学的根拠を医療現場で有効に活用するエビデンスに基づく医療 (EBM) が注目を浴びている。看護課程の学部学生にとっても EB M を実践するために論文や統計資料を読み解く機会が増えており、疫学・生物統計学を学ぶ事が重要となってきた。

看護系の専門学校及び大学のカリキュラムは基礎分野、専門基礎分野、専門分野に分類され、指導要領の上では疫学・生物統計学に関する授業は一般教養に該当する基礎分野に位置づけられる。これは1997年の大学・専門大学看護課程におけるカリキュラム及び看護師国家試験出題基準によるものである。保健師の指導要領では疫学・保健統計という名で必修とされている。疫学・生物統計学は看護師国家試験にはあまり出題されていないなど、カリキュラム上、大きく取り上げられていないのが現状である。

一般に、統計学を教える教官の不足は以前からも指摘されており、全大学の学部教育における統計学の教員数は0.5%にすぎず、統計学の学部教育は統計学を専門としない教官に依存しているといわれている。しかし疫学・生物統計学の学部教育を対象とした研究はほとんどなされておらず、文献などから実態を把握する事も難しい。大きく取り上げられていない。

2. 目的

本研究では看護系大学における疫学・生物統計学教育の実態を把握するため、自記式調査票を用いて調査を行う。目的は、疫学・生物統計学の授業を担当する教官の背景、教官が授業を行っていく上での問題と感じている点、看護系大学においての疫学・生物統計学の講義・実習の内容を把握し、教官の背景と抱えている問題点に関連があるかどうかを検討することとする。

3. 対象

看護系の教育課程を持つ国公立大学62校及び私立大学27校(計89校)の疫学・生物統計学の授業を担当している教官を調査対象とした。各大学のシラバスやホームページなどから担当教官をあらかじめ調べ、原則として疫学・生物統計学担当の教官を調査した。

4. 方法

調査計画書に基づき2002年7月下旬に調査票3部と調査を依頼する手紙及び返送用の封筒・切手を、対象校の疫学・生物統計学担当の教官宛に郵送により送付した。教官名が不明な場合は学科長(若しくは学部長)宛とした。調査票を3部送ったのは一つの大学に複数の担当教官がいる事も考えられたためである。発送締め切り期日は8月31日とし、返送されてこなかった大学には電話で再度依頼した。集計後、調査票を回収できた教官には結果を報告書にまとめ12月下旬に送付した。

調査には表紙を含めA4用紙11枚、33項目からなる自記式調査票を用いた。調査票の表紙には依頼文を載せ、プライバシーは守る事、調査結果は報告書にまとめてフィードバックする事、連絡先を明記した。質問項目の内容については表1に示す。

表1、調査票の質問項目

問1	教官の背景(現在の所属、学位と最終学歴、専門分野、所属学会)
問2	疫学・生物統計学教育についての考え方(どのような状況で必要になるか)
問3	教官が授業を行う上で問題と感じている点(教材、設備、教官、生徒など)
問4	講義内容(カリキュラム上の位置付け、教科書や統計パッケージ、扱う分野)
問5	実習内容(カリキュラム上の位置付け、教科書や統計パッケージ、扱う分野)
問6	卒業論文指導の内容(指導形態、参考書や統計パッケージ、扱う内容やテーマ)

5. 結果

5. 1. 回収状況

対象校89校のうち50校から61通の回答が得られた。回答数が対象校の数より大きいのは複数の教官が回答してきた大学があったためである。回収率は56% (50/89)であった。回収できた大学とできなかった大学とに地域差や国公立・私立の割合の差は見られなかった。

5. 2. 担当教官の背景

担当教官の専門分野を表2に示す。疫学・生物統計学を専門としている教官は12人(20%)、公衆衛生学を専門としている教官は10人(17%)、看護学の他の分野を専門としている教官は13人(22%)であった。他学部からの教官は18人(31%)であり、他学部における統計(数理統計、確率論、経済統計)、工学部、薬学部出身で医療情報、心理学を専門とする教官がみられた。

また、所属学会については表3のとおりである。看護に関する学会は20人(33%)の教官が所属しており、日本看護科学学会14人(23%)、日本看護研究学会16人(26%)が主であった。疫学・公衆衛生学に関する学会には40人(66%)の教官が所属しており、日本疫学会18人(30%)、日本公衆衛生学会39人(64%)が主であった。統計に関する学会に所属している教官は11人(18%)と少なかった。情報に関する学会には18人(30%)が所属しており、医療情報学会14人(23%)が主であった。心理に関する学会に所属している教官は7人(11%)と少なかった。

また、看護学生にとって今後どのような状況で疫学・生物統計学が必要になると考えられているかを教官に質問した結果、ほとんどの教官が疫学・生物統計学を教える必要があると答え、しかも52人(85%)と多くの教官が学生に十分身につけていないと回答した。

表2、担当教官の専門分野 (N=61)

看護学部	40	(69%)
疫学と生物統計学の両方	2	(3%)
疫学	7	(12%)
生物統計学	3	(5%)
公衆衛生学	10	(17%)
看護学科のその他の分野	13	(22%)
看護学(専門は不明)	5	(9%)
他学部	18	(31%)
数理統計または確率論	3	(5%)
経済統計	1	(2%)
工学系、薬学系の医療情報	4	(7%)
薬学部のその他の分野	1	(2%)
工学部のその他の分野	3	(5%)
数学科の統計以外の分野	3	(5%)
農学	1	(2%)
心理学	2	(3%)
無回答	3	

表3、担当教官の所属学会 (N=61)

看護に関する学会	20	(33%)
日本看護学会	3	(5%)
日本看護科学学会	14	(23%)
日本看護研究学会	16	(26%)
その他	3	(5%)
疫学・公衆衛生に関係する学会	40	(66%)
日本疫学会	18	(30%)
日本公衆衛生学会	39	(64%)
その他	14	(23%)
統計に関する学会	11	(18%)
日本統計学会	7	(11%)
応用統計学会	3	(5%)
計量生物学会	6	(10%)
行動計量学会	5	(8%)
その他	2	(3%)
情報に関する学会	18	(30%)
医療情報学会	14	(23%)
その他	9	(15%)
心理学に関する学会	7	(11%)
日本心理学会	4	(7%)
その他	6	(10%)

5. 3. 教官が授業を行っていく上で問題と感じている点

教官の感じている問題点を、表4のように質問項目に対し問題意識をもっているとした教官数で示す。ここでは60%以上の教官が問題意識をもっていたものを取り挙げる。

教材については39人(65%)の教官が「調査・統計実習に適した教材が少ない」と回答した。設備について問題であると回答した教官は少なかった。教官・スタッフの人数に関しては46人(75%)の教官が「チューターなど授業を手伝ってくれるスタッフの人数が不足している」と回答した。学生について「学生の疫学・(生物)統計学を学ぶ意欲が足りない」40人(67%)、「学生に授業を行う上で前提となる知識・能力が足りない」47人(78%)といった回答があり、学生に足りない能力として「数学」45人(74%)、「パソコン・情報処理」23人(38%)が挙げられた。また、教官については37人(62%)の教官が「教官自身にもっと学ばなければならない分野がある」と回答し、教官自身に足りない能力としては「数学」17人(28%)、「医学」12人(20%)、「パソコン・情報処理」14人(23%)、「調査や研究の実践」21人(34%)が挙げられた。また、自由回答では、「看護に関する実例を挙げた教科書がない」19人(31%)、「統計学の重要性を知らない教官が多い」5人(8%)という意見があった。

表4. 担当教官が授業を行う上で感じている問題点 (N=61)

教材について		
調査・統計実習に適した教材が少ない	39	(65%)
問題集のようなものが少ない	35	(58%)
外国の良い教科書がなかなか翻訳されない	23	(39%)
内容(難度・分野など)の適切な教科書がない	34	(56%)
設備について		
パソコンやインターネットなどの設備の不足	8	(13%)
学生が使用できるような統計ソフトの不足	15	(26%)
文献を利用する環境が整っていない	9	(15%)
液晶プロジェクターやOHPなど映像関係の設備の不足	9	(15%)
教官・スタッフの人数について		
担当教官の人数の不足	32	(52%)
授業を手伝ってくれるスタッフの人数の不足	46	(75%)
学生の疫学・生物統計学を学ぶ意欲が足りない	40	(67%)
学生に前提となる知識・能力が足りない	47	(78%)
・語学*	6	(10%)
・数学*	45	(74%)
・医学*	10	(16%)
・パソコン・情報処理*	23	(38%)
教官について		
担当教官の間で意見交換ができる場がない	24	(40%)
こういった分野・難度まで教えればよいのか分からない	26	(43%)
ご自身にもっと学ばなければならない分野がある	37	(62%)
・語学*	4	(7%)
・数学*	17	(28%)
・医学*	12	(20%)
・パソコン・情報処理*	14	(23%)
・調査や研究の実践*	21	(34%)

*この分野について知識・能力が足りないと思うと回答した教官数を挙げた。

5. 4. 看護系大学における疫学・生物統計学の講義・実習の内容

疫学・生物統計学講義の必修・選択と実習の有無、講義時間を表5に示した。複数の授業がある場合の講義時間は、時間が長いものをその大学の講義時間とした。ほとんどの大学で疫学・生物統計学の講義が必修となっており、その講義時間は90分の講義が週1回で15週前後、すなわち1350分前後の講義が組まれている大学が23校(51%)と多かった。実習に関しては31校(62%)の大学で行われていた。

用いられている教科書・統計パッケージを表6に示した。教科書を用いていた大学は30校あった。特によく使われている教科書はなかった。複数の大学で使われていた教科書・教材を挙げると、南山堂「保健統計・疫学」、南江堂「疫学 基礎から学ぶために」、医学書院「ナースのための疫学」、厚生統計協会「厚生統計テキストブック」がそれぞれ2校ずつで用いられていた。また17校(36%)で講義がプリントのみによって行われていた。統計パッケージはSPSSが多く21校(42%)で用いられていた。

表5、授業の必修・選択、講義時間 (N=50)

必修・選択と実習の有無			
必修の講義あり		44	(88%)
なし		6	(12%)
実習あり		31	(62%)
必修の実習あり		24	(77%)
なし		7	(23%)
講義時間			
270分から1350分		8	(18%)
1350分		23	(51%)
1350分から4050分		14	(31%)
無回答		5	

表6、教材・統計パッケージ (N=50)

教材			
教科書を用いている		30	(64%)
プリントのみ		17	(36%)
無回答		3	
統計パッケージ			
SPSS		21	(42%)
HALBOU		5	(10%)
JMP		4	(8%)
STATVIEW		2	(4%)
SAS		2	(4%)

講義・実習の分野に関しては「疫学」と「生物統計学」など講義が分かれている事も考えられるため、その大学の授業全体で網羅できている分野にまとめて集計し、表7に示した。「2. 統計における基本概念」に含めた分野や、「4. 統計解析」のうちの基本的な分野に関しては90%前後の大学で講義していたが、「1. 疫学における基本概念」や「3. 医学・疫学研究デザイン」に含めた分野については70%前後となっていた。「4. 統計解析」のうちメタアナリシスや生存時間解析に触れている大学は約30%と少なかった。表には示していないが、実習で触れている分野については「メールやワープロなどパソコンの基本的な使用」21校(78%)、「Excelなどの表計算ソフト」26校(93%)、「統計ソフトを用いた統計処理」25校(89%)、「実際のデータ解析」20校(77%)の割合が大きかった。

表7、講義のなかで触れられている分野(N=50)

	触れない	軽く・詳しく触れる	無回答
1. 疫学における基本概念			
疫学の定義, 目的, 対象, 歴史	14 (31%)	31 (69%)	5
がんや感染症など各々の疾患における疫学	17 (38%)	28 (62%)	5
因果関係についての解説	7 (16%)	38 (84%)	5
罹患率や有病率など疾病頻度の指標	10 (22%)	35 (78%)	5
オッズ比、相対危険、寄与危険など曝露効果の指標	12 (27%)	33 (73%)	5
敏感度と特異度や予測価など検査の特性	13 (30%)	32 (73%)	6
偏りと交絡	9 (20%)	35 (80%)	6
マッチング	13 (30%)	31 (70%)	6
生命表や人口動態調査などの保健統計資料	8 (19%)	35 (81%)	7
2. 統計における基本概念			
質的データと量的データ(データと尺度と型)	1 (2%)	47 (98%)	2
正確度と精度	8 (17%)	39 (83%)	3
無作為抽出と無作為割り付け・外的と内的妥当性	4 (9%)	42 (89%)	3
平均値や分散など代表的な統計量	1 (2%)	47 (98%)	2
正規分布やポアソン分布など代表的な分布	2 (4%)	45 (96%)	3
検定と推定	1 (2%)	46 (98%)	3
3. 医学・疫学研究デザイン			
観察と介入、縦断と横断などの分類	8 (17%)	38 (81%)	3
ケースコントロール研究とコホート研究	8 (17%)	38 (81%)	3
臨床試験	14 (30%)	32 (68%)	3
地域介入研究と地域相関研究	16 (34%)	30 (64%)	3
4. 統計解析			
データの記述とグラフ表示	1 (2%)	47 (98%)	2
t検定	0 (0%)	47 (100%)	3
χ^2 検定	0 (0%)	47 (100%)	3
ウィルコクソン検定	12 (26%)	35 (74%)	3
平均値の差の信頼区間	5 (11%)	42 (89%)	3
分散分析	10 (21%)	38 (81%)	3
相関係数	2 (4%)	44 (94%)	3
重回帰分析	16 (36%)	31 (69%)	5
ロジスティック回帰分析	22 (49%)	24 (53%)	5
層別解析	22 (49%)	23 (51%)	5
SMRなどの標準化	20 (44%)	27 (60%)	5
メタ・アナリシス	31 (66%)	14 (30%)	3
生存時間解析	31 (69%)	14 (31%)	5

5. 5. 教官の背景と抱えている問題点の関連性

疫学・生物統計学を専門としている教官と専門ではない教官で抱えている問題点がどう違うかを調べるため、「それについて問題と思う・思わない」と「疫学・生物統計学を専門としている・していない」とで比較した結果、疫学・生物統計学を専門としている・専門としていないで割合が異なっていた問題点は「教官自身にもっと学ばなければならない分野がある」(特に語学、医学と調査や研究の実践)であり、有意差はみられないものの特に割合の異なっていた問題点は「チューターなど授業を手伝ってくれるスタッフの人数が不足している」、「学生の疫学・(生物)統計学を学ぶ意欲が足りない」、「学生に授業を行う上で前提となる知識・能力が足りない」(特に数学とパソコン・情報処理)であった。いずれも専門としていない教官のほうが問題であると答えた割合が大きかった。また、専門でない教官をさらに「公衆衛生学」、「看護のその他の分野が専門」、「他学部」に分類してみると、他学部に所属している教官が「内容(難度・分野など)の適切な教科書がない」と答えた割合がやや小さく、「学生に数学とパソコン・情報処理の能力が足りない」と答えた割合がやや大きかった。

専門としている教官と専門としていない教官で講義している分野がどう違うかを比較した結果、「因果関係についての解説」、「罹患率や有病率など疾病頻度の指標」、「オッズ比、相対危険、寄与危険など曝露効果の指標」、「敏感度と特異度や予測値など検査の特性」、「生命表や人口動態調査などの保健統計資料」、「SMRなどの標準化」など、「疫学における基本概念に含めた分野」、「医学・疫学研究デザイン」に含めた分野は割合の異なるものが多かった。いずれも専門としている教官のほうが教えている割合が大きかった。

また専門でない教官をさらに「公衆衛生学」、「看護のその他の分野が専門」、「他学部」で分けてみると、特に他学部に所属している教官が疫学の基本概念や研究デザインに関する分野、看護・医療で特に使われている解析手法についてあまり講義していなかった。

6. 考察

教官の背景については、ほとんどの教官が疫学・生物統計学を教える必要があると考えていた事や、他の教官に疫学・生物統計学の重要性が認知されていないという意見が5件あった事などから、疫学・生物統計学の教育は重要であると考えられているようである。しかし、疫学・生物統計学を専門としている教官が講義している大学は21%しかない一方で、工学部・薬学部・理学部数学科などの他学部からの教官に頼るケースが31%あった。また、疫学・公衆衛生学に関する学会に所属している教官が66%いるものの、そのうちのほとんどの教官は日本公衆衛生学会に所属しており、日本疫学会などの疫学に関する学会に所属している教官は少なかった。統計に関する学会に所属している教官が特に少なく18%しかいなかった事も含め、疫学・生物統計学を教える教官の不足が伺えた。

授業を行う上で問題点については、学生に教科書を購入させずプリントで授業を行っている大学は36%と予想していたよりも少なく、講義は教科書を中心に行われており、「内容(難度・分野など)の適切な教科書がない」と56%の教官が回答した事、自由回答で看護に関する実例を挙げた教科

書がないという意見が19件あった事など、教科書に対する要望が少なからず見られた。多数の大学で用いられている教科書はなかったが、もし標準的な教科書があれば分量・難度ともにどの程度教えればいいのかの基準となり教官にとっても講義しやすくなると考えられる。また、統計パッケージが高価であるという意見が3件あった。実習等で用いられている統計パッケージとしてはSPSSが多かった。国内ではSPSSが広まっている事に加え、比較的安価なため導入しやすいとと考えられる。

教官・スタッフの人数が不足している事を問題と感じている教官も多く、若手を育成する必要があるという意見も3件挙げられるなど、教官自身も教官の不足を問題と感じているという事が分かった。また、多くの教官が学生の意欲・数学とパソコン・情報処理の能力の不足を感じていた。

講義・実習の内容については、ほとんどの大学で疫学・生物統計学の授業が必修になっており講義時間も充分とられており、大学の設備に不満は見られず、制度・設備に関してはどの大学でもある程度充実していると考えられる。また、手伝ってくれるチューターがいないという意見も多かったが、この事が実習の充実しない一つの原因とも考えられる。疫学・生物統計学を専門とする教官が増えれば、研究室の院生などをスタッフとして用いる事ができるようになると思われる。疫学における基本概念や研究デザインに関する分野は70%前後の大学でしか教えられておらず、もっと講義で取り上げられる事が望まれる。

教官の背景と抱えている問題点や講義内容との関連性については、学生に数学とパソコン・情報処理の能力が足りないという意見が、特に他学部にも所属している教官の中に多かった。その理由として数理的な面を重視して教えているかもしれない事、看護学生のなかに数学を十分に履修していないものもいる事、他の学部の学生と比較してしまう事が考えられる。疫学の基本概念や研究デザインに関する分野、看護・医療で特によく使われている解析手法については疫学・生物統計学を専門とする教官の方がよく教えているようであった。この事からも疫学・生物統計学を専門とする教官が今後増えていく事が望まれる。

7. 結論

本研究では看護系大学における疫学・生物統計学教育の実態について自記式調査票を用いて調査を行った。疫学・生物統計学を専門としている教官が疫学・生物統計学を講義している大学は少なく、工学部・薬学部・理学部数学科などの他学部にも所属している教官に頼っている事が分かった。また、良い教科書・実習用の教材・問題集がなく、特に看護に関する事例を挙げた教科書が望まれている事、教官やチューターの人数が不足している事、学生の意欲、数学やパソコン・情報処理の能力が足りない事が問題点として挙げられた。疫学の基本概念や研究デザインに関する分野、看護・医療で特によく使われている解析手法については講義している大学がやや少ないようであり、特に工学部・薬学部・理学部数学科などの他学部にも所属している教官はそれらの分野について講義で触れてない傾向があった。また、他学部にも所属している教官に学生に数学とパソコン・情報処理の能力が足りないと感じているものが多かった。

参考文献

- 1) John M. Last 編、疫学辞典第3版、2000
- 2) 厚生省健康政策局、医療技術評価推進検討会報告書、1999
- 3) 看護問題研究会監修、新訂看護教育カリキュラム、第一法規、1997
- 4) 村山征勝、大学における統計学の教育・研究環境とその問題点、統計数理、1995、vol.43、no.2、367-75
- 5) 宮下光令、笹原朋代、Evidence-Based Nursing 誌について、Quality Nursing、2001、vol.7、no.10、841-8
- 6) 宮下光令、笹原朋代、数間恵子、わが国の看護研究論文に用いられている統計手法について、Quality Nursing、2001、vol.7、no.10、849-54
- 7) 丹後俊郎、消化器病学に関する研究論文での統計的方法について、日消誌、1992、vol.89(1)、90-96、vol.89(2)、561-8、日消誌、1993、vol.90(1)、75-82、vol.90(8)、1722-8
- 8) 丹後俊郎、研究の種類に応じたデータのまとめ方、日消誌、1995、vol.95(5)、412-8
- 9) 浜田知久馬、臨床統計 FAQ、臨床医薬、1999、15 卷 10 号、1583-99
- 10) 浜田知久馬、統計パッケージを誤用しないために、臨床麻酔、1999、vol.23、no.10、1651-6
- 11) 山内一史、看護における情報学;何をどう教えるか、看護展望、2000、vol.25、no.13、1476-86
- 12) 佐藤俊哉、調査票の作成、保健の科学、1995、第37巻、第2号、72-6
- 13) 真部昌子、「国試出題基準」はどのような看護婦を求めているのか、看護教育、2000、vol.41、202-7
- 14) 川島みどり、今、求められる基礎教育の質、看護教育、1997、vol.38、874-86

ポスターセッション システム

日本SASユーザー会 (SUGI-J)

SAS を用いた XML データの作成 —ODM ver. 1.1 対応—

○岡下 邦博 進藤 三富子

株式会社日本アルトマーク

統計解析部

Making XML data files for ODM ver 1.1 in SAS System

Kunihiro Okashita Satoko Shindo

Ultmarc inc.

Statistical Analysis Division

要 旨

2002年4月にCDISC(Clinical Data Interchange Standards Consortium)は、XML文書を用いた臨床データの標準仕様としてODM(Operational Data Model)のVersion1.1を公開した。臨床データのあり方を模索する中で今後有力なトレンドと思われるXML形式に注目し、ODM Version1.1形式のデータからSASを用いて作成・運用する社内システムを開発する方向性について社内で検討した結果を報告したい。

キーワード： Base SAS XML ODM ver. 1.1

1. CDISC ODM(Operational Data Model)

米国の非営利団体「Clinical Data Interchange Standards Consortium(CDISC)」が2000年10月に策定した仕様。現在Version1.1(2002年4月策定)。XMLを用いた電子的なデータ交換方式の標準化を行っている。調査票の形式にこだわらず、XMLデータを作成する事が可能になった。またSASにも関連した仕様も含まれており、報告書を作成する上での解析用SASデータセットを作成できる。

1.1 CDISC ODM データ構成

- I. 調査概要(Study Attributes)
- II. 調査内容(Study)
 - 1. 調査・プロトコール名等(Global Variables)
 - 2. 基本項目(Basic Definitions)
 - 単位(Measurement Unit)
 - 3. ファイル構成(Metadata Version)
 - ① データバージョン管理(Include Prior Metadata Version)
 - ② 患者登録情報(Protocol Study Events) → 症例番号 or 登録番号
 - ③ フォーム名(Forms for Study Event) → 患者背景、有害事象…
 - ④ フォーム詳細(Form Definitions) → 患者背景:性別、年齢…
 - i. グループキー項目(Item Group Reference for Form)
 - ii. グループ項目(Item Groups) → SAS データセット名及び構成
 - ⑤ 変数項目(Items) → SAS 変数名
 - 追加情報(Additional Information for Item)→効果判定・服薬状況等
 - ⑥ コードリスト(Code List) → SAS フォーマット
- III. 管理情報(Administration Data)
- IV. キー項目(Reference Data)
- V. 臨床データ情報(Clinical Data Study)
- VI. 入力データ(Subject Data)

具体的なモデル構成及び仕様については DTD ファイル等を参照してほしい。

2. SAS システムにおける XML データ

SAS システムでは、V8より評価版ではあるものの、XML への出力及び読込が可能になった。日本語での解説がほとんどないため、開発にあたっての情報収集は海外のサイトが中心となった。

SAS→XML の変換は以下の方法で実行する。

- 1.DATA ステップ
- 2.SAS SCL によるデータ編集
- 3.XML LIBNAME エンジン(評価版)
SAS XML LIBNAME ENGINE(SXLE) Module(レジストリ必要)
→CDISC 対応
- 4.ODM Markup(評価版)

XML→SAS の変換の変換は以下の方法で行う。

- 1.DATA ステップ
- 2.SAS SCL によるデータ編集
- 3.XML LIBNAME エンジン
XMLMap オプション使用(評価版)
※XMLMap:XPath 及び XPointer の仕様に基づく XML ファイル。

3. 開発手順

3.1 開発目的

入力された SAS データセットから Web 上に帳票を表示するシステムを作成する。SAS データセットは作業内容によってデータ仕様が異なるため、CDISC の ODM を標準仕様とし、仕様に従って XML ファイルを作成する。作成された XML は Java 等にて Web 上に表示される。

3.2 開発内容

- ・フロー作成
- ・仕様書作成
- ・SAS データ定義書
- ・ODM データ定義書
※共に Excel で作成し、SAS プログラムでデータ取込
再利用できるように標準フォーマット作成
- ・プログラム開発
- ・SAS→XML データ変換(SAS) ※DATA ステップで XML 出力
- ・運用画面(HTA)
※HTA(HTML Application) : HTML 形式のアプリケーションツール

4. 問題点及び今後の課題

4.1 問題点

- ・ XML出力が Clinical Data しか対応していないこと
- ・ 正確な仕様書作成の必要があること
 - 1 つでも誤りがあれば、XML ファイル作成に与える影響が大きい
- ・ 変換後のデータの検証方法が確定していないこと
 - SAS Contents リストと仕様書との検証
 - 症例数が多くなると XML 形式では膨大な量になる
- ・ 今後、XML バージョンアップに伴って対応をとっていく必要があること
- ・ XML ファイルのボリュームが SAS データセットよりかなり大きいため、Standalone ではシステムに負担がかかること
- ・ SAS フォーマットカタログに未対応であるため、現状は SAS フォーマットなしで出力していること

4.2 今後の課題

- ・ データ管理体制の確立を目指す
 - 現在 SAS データセットをマスタデータとしているが、ODM とどちらかにするかを検討する必要がある。
 - SAS の場合、Audit Trail を用いた管理方法が考えられる
 - ODM の場合、データ入力・修正を含めたシステムの開発が必要となる
- ・ Clinical Data 以外のデータ処理が未対応のため、今後対応を広げていく必要がある
- ・ DATA ステップ以外の XML データ変換方法を模索する必要がある。次期日本語バージョンにて標準装備されることを期待する
- ・ 他のデータベース (MedDRA, MEDIS 等) の連携をとる必要がある
- ・ テストデータは英語で行ったため日本語対応のテスト・運用が必要となる
- ・ XML→SAS への対応については SDM, ADaM 用 SAS データセット作成、XMLMap の利用などが考えられるため、今後検討をすすめていく必要がある

参考文献

- 1) 「解析用データセットのあり方 - CDISC を意識して -」, SUGI-J 2002 論文集, 長谷川要, 本山佳代子, 小崎昌昭, 外城靖子 麒麟麦酒株式会社 医薬カンパニー 開発本部 開発推進部
- 2) CDISC ホームページ <http://www.cdisc.org/>
- 3) 「Introduction to the CDISC Operational Data Model Version 1.1 (Final)」

- <http://www.cdisc.org/models/odm/v1.1/ODM1-1-0-Intro.pdf>
- 4) 「Overview of the CDISC Operational Data Model for Clinical Data Acquisition and Archive (based on CDISC DTD 1.1 Final)
<http://www.cdisc.org/models/odm/v1.1/ODM1-1-0-Overview.pdf>
 - 5) 「Specification for the Operational Data Model (ODM)」
<http://www.cdisc.org/models/odm/v1.1/ODM1-1-0.html>
 - 6) 「Overview of Techniques for Reading and Writing ODM Data」 2001/11/6
<http://www.cdisc.org/pdf/CDISCReadWriteODM27.pdf>
 - 7) 「Operational Data Model Proof of Concept Demonstration 11th Annual European Workshop on Clinical Data Management」 Testing and Applications Team, CDISC Clinical Data Connectathon, 2001/10/30
http://www.cbtech.com/Paris_CTHON/
 - 8) 「CDISC, the new XML standard for Clinical Data」, XML4Pharma, Computer Chemistry Consultancy
<http://www.compchemcons.com/CDISC/index.html>
 - 9) 「XML and SAS : An Advanced Tutorial」, SUGI25, paper 13-25, Greg Barnes Nelson, STATPROBE Technologies, Cary, NC
<http://www2.sas.com/proceedings/sugi25/25/aa/25p013.pdf>
 - 10) 「XML Roadmap for the SAS System」, SUGA02, Rudy Gyzen, Qual I. T. Services
<http://www.sas.com/offices/asiapacific/sp/suga/2002/presentations/SUGA02--TS3--XMLroadmap--RudyGyzen--QUALIT.pdf>
 - 11) 「XML Resources」, SAS Institute, Inc., Technical Support, Base SAS Community
<http://support.sas.com/rnd/base/index-xml-resources.html>

日本SASユーザー会 (SUGI-J)

SAS データセットのエクスポート

羽田野 実

SAS Institute Japan 株式会社

カスタマーサービス本部 プロフェッショナルサービス第 1 部

Exporting a SAS data set

Makoto Hatano

Professional Service Department 1, Customer Services Division,

SAS Institute Japan Ltd.

要 旨

基幹システムデータ、実験データなど SAS System にインポートされたデータを加工、集計、分析された結果を他のシステムやアプリケーションにエクスポート(出力)する場合があります。エクスポート先として Oracle などのデータベースシステム、Microsoft EXCEL、XML ファイル、カンマ区切りファイルなどがある。本論文では、Windows 版 SAS 8.2 において追加・拡張された機能を用いて SAS データセットを Microsoft EXCEL にエクスポートする方法について、プログラムコードを例示しながら記述する。

キーワード： BASE、ACCESS、EXPORT、EXCEL

1. はじめに

SAS データセットを Microsoft EXCEL(以下 EXCEL)にエクスポートするには、DDE(Dynamic Data Exchange)を始め表 1.1 に示す方法などがある。これらの方法のいくつかについて記述する。

表 1.1 SAS データセットのエクスポート方法

方法	記述
DDE	FILENAME ステートメント DDE エンジン
OLE オートメーション	SCL(SAS Component Language)
DBLOAD プロシジャ	
EXPORT プロシジャ	
ODS(Output Delivery System)	SAS 8.2
ODBC	SAS 8.2、LIBNAME ステートメント ODBC エンジン
OLE DB	SAS 8.2、LIBNAME ステートメント OLEDB エンジン

2. DDE

DDE(Dynamic Data Exchange 動的データ交換)とは、Microsoft Windows オペレーティングシステムファミリに実装されているプロセス間通信(IPC)の形式である。FILENAME ステートメントに DDE エンジンを指定して DATA ステップで SAS データセットを EXCEL ファイルに出力できる。プログラム例を図 2.1 に記述する。

```

/* EXCELの起動*/
options noxwait noxsync;
x 'start excel';
data _null_;
    rc= sleep(2);
run;
/* ファイル参照名の割り当て (DDEエンジン) */
filename class dde 'excel|Sheet1!r1c1:r20c5';
/* SASデータセットのEXCELへの出力 */
data _null_;
    set sashelp.class;
    file class;
    if _n_ = 1 then
        put '名前' '09'x '性別' '09'x '年齢' '09'x
            '身長 (インチ)' '09'x '体重 (ポンド)' '09'x;
    if index(trim(name), '09'x) or
        index(trim(name), '') then do;
        temp= put(name, $quote200.);
    
```



```

    put temp $ +(-1) @;
end;
else
    put name $ +(-1) @;
if index(trim(sex), '09'x) or
    index(trim(sex), '') then do;
    temp= put(sex, $quote200.);
    put '09'x temp $ +(-1) @;
end;
else
    put '09'x sex $ +(-1) @;
if age > .z then do;
    temp= left(put(age, best12.));
    put '09'x temp $ +(-1) @;
end;
else
    put '09'x @;
if height > .z then do;
    temp= left(put(height, best12.));
    put '09'x temp $ +(-1) @;
end;
else
    put '09'x @;
if weight > .z then do;
    temp= left(put(weight, best12.));
    put '09'x temp $ +(-1);
end;
else
    put '09'x;
run;

```

図 2.1 DDE のプログラム例

また、図 2.2 のコードのように EXCEL コマンドを発行してシート追加、シート名、罫線追加など様々な制御ができる。

```

filename xcmd dde 'excel|system';

data _null_;
    put '[close.all]'; /* EXCEL ファイルのクローズ */
    put '[new(1)]';    /* シートの追加 */
run;

```

図 2.2 EXCEL コマンド発行プログラム例

3. OLE オートメーション

OLE オートメーション(OLE Automation)を使用して SAS データセットを EXCEL ファイルに出力できる。この場合、SAS/AF の SCL(SAS Component Language)でプログラムを記述する必要がある。使用するクラスは、SASHELP.FSP.HAUTO.CLASS である。

4. DBLOAD プロシジャ

DBLOAD プロシジャで SAS データセットを EXCEL ファイルに出力するプログラム例を図 4.1 に記述する。

```

proc dbload dbms= excel
    data= sashelp.class
    path= 'e:\temp\class.xls';
    putnames y;
    limit= 0;
    label;
    reset all;
    load;
run;

```

図 4.1 DBLOAD プロシジャのプログラム例

5. EXPORT プロシジャ

EXPORT プロシジャで SAS データセットを EXCEL ファイルに出力するプログラム例を図 5.1 に記述する。

```
proc export dbms= excel
    data= sashelp.class
    outfile= 'e:¥temp¥class.xls';
run;
```

図 5.1 EXPORT プロシジャのプログラム例

6. ODS

SAS 8.2 で追加された ODS(Output Delivery System)で SAS データセットを EXCEL ファイルに出力するプログラム例を図 6.1 に記述する。

```
title;
footnote;
ods listing close;
ods html file= 'e:¥temp¥class.xls';
proc sql;
    select *
    from sashelp.class;
quit;
ods html close;
ods listing;
```

図 6.1 ODS のプログラム例

EXCEL ファイルへの出力の見映え(罫線、色など)は、TEMPLATE プロシジャや ODS タグを用いてカスタマイズできる。

7. ODBC

SAS 8.2 で追加された SAS/ACCESS LIBNAME ステートメントオプションでエンジンに ODBC(Open Database Connectivity)を指定することにより、SAS データセットを EXCEL ファイルに出力するプログラム例を図 7.1 に記述する。

```
libname xlslib odbc noprompt= "dsn=Excel Files;
                                dbq=excelfile.xls";
proc copy in= sasuser out= xlslib;
    select class crime fitness /mt= data;
run;
```

図 7.1 ODBC のプログラム例

8. OLE DB

SAS 8.2 で追加された SAS/ACCESS LIBNAME ステートメントオプションでエンジンに OLEDB を指定することにより、SAS データセットを EXCEL ファイルに出力するプログラム例を図 8.1 に記述する。

```
libname xlslib oledb provider= "Microsoft.Jet.OLEDB.4.0"
                                properties=('data source='e:¥temp¥excel.xls')
                                provider_string= "Excel 8.0; HDR=YES;";
proc append base= xlslib.class
    data= sasuser.class;
run;
proc append base= xlslib.crime
    data= sasuser.crime;
run;
```

図 8.1 OLE DB のプログラム例

9. まとめ

SAS 8.2 で追加された ODS 及び SAS/ACCESS LIBNAME ステートメントオプションにより、SAS データセットの EXCEL ファイルへの出力が、より容易に、より SAS プログラムライクになった。今後 SAS System 9 以降も更なるエンハンスが期待される。

ポスターセッション
経営・経済

日本SASユーザー会 (SUGI-J)

労働市場の時系列分析 ～JMPを利用して～

浦澤 浩一

八千代銀行/青山学院大学

Time-series Analysis of Labor Market ～User of JMP～

Yachiyo Bank/Aoyamagakuin University

Kouichi Urasawa

要 旨

JMPを利用し、日本の労働市場をマクロ分析する。(集計データにて)
UV分析、総供給曲線、オークン係数、フィリップス曲線、個用意調整速度係数を
理論と実証を交えながら分析する。

JMP、Labor Market、Econometrics

キーワード：

目次

1. はじめに

- | | | |
|----|------------------|-----|
| 1章 | 労働市場におけるマクロ経済分析 | まとめ |
| 1節 | 日本の労働市場における経験的背景 | |
| 2節 | UV分析 | |
| 2章 | マクロ分析 | |
| 1節 | 総供給曲線 | |
| 2節 | オークン係数 | |
| 3節 | 雇用調整係数の推計 | |

はじめに

1998年頃より、米国の失業率もついには逆転するという未曾有な状況に陥った。このような深刻な長期の不況下において、体力のない企業は吸収され、また企業内においても早期希望退職制度や高齢者世代を中心にリストラが活発に行われるように、長期雇用制の維持が困難になってきている。それと同様に雇用システムの特徴の一つとして、年功序列賃金制度が能力主義システムに転換してきている。³このような経済停滞を端に発する雇用の変化の主要な原因と見るのが妥当であるのだが、他にも要因となるものもあるであろう。1990年代においては、機械部品や繊維衣料品の東南アジア諸国での生産拠点の国際化、合併などによる産業界の再編、女性の高学歴化・機会均等法における社会進出、インターネットやオートメーション化にみられる急速な技術革新とそれに対応できる企業内訓練の困難性、中途採用者の増加やフリーアルバイト、ワークシェアリング⁴など職業形態の変化など、社会環境・構造変化が顕著であった時代である。そして雇用の潮流は、旧来前の雇用制度の特徴とされてきたものが、企業の欲する人材とその欲する能力を有する労働者とのミスマッチ⁵を排除した、流動性のある労働市場へと変化しているのではないか。

日本の雇用の安定性に変化が見受けられる一つつまり日本的雇用の安定性とは、雇用量的変化が小さいことである労働保蔵を行なっていることである。つまり失業率の低位安定性も含め、終身雇用制や年功序列賃金制にみられる、長期にわたる企業と労働者の「暗黙の契約」の論理

3. 長期雇用制度や年功序列制度は日本的なものではなく、長期雇用制度は西欧諸国でも多くみられ、米国においては先任権制度による若年者のレイオフは多いが、長期雇用は一般的である。年功序列制度に関しては、企業への定着という企業戦略、基幹労働力を確保する為に永年勤続を優遇し、熟練のスキルを積みながら昇進する人的資本理論で説明できる。但し査定制度(成績)や能力によって、賃金の相違はあるのである。雇用制度については、中馬(1987)、中村・大橋(2002)に詳しい。人的資本論は、J.Mincer(1974)、G.Becker(1964)参照。

4. オランダモデルが目立っており、日本においては、兵庫県が検討している。兵庫県経営者協会「ワークシェアリング検討委員会報告」参照されたい。またオランダ、ドイツ、フランスのワークシェアリングの効果として計量分析している内閣府『世界経済の潮流』2002秋を参照されたい。

5. ミスマッチはマクロ分析の章 UV 分析でも触れるが、最近では松下モデルと呼ばれる、インターン制度をもちいた学生と企業の相互理解(ミスマッチの解消)が今後主流になると言われている。インターンシップ事例 松下電器産業「松下ウォーミングアップ・プログラム」短時間の面接・選考というプロセスでは困難な「企業と学生の相互理解に基づいた採用と就職」を主にしている。採用の際に基準についても「学歴・潜在能力・総合力」から「能力(初速のスピード)、スキル、専門性」に特化しつつあるようである。また富士ゼロックスはweb上にて「FXDBS ワークショップ」を開催。期間は長いもので、NECの2ヶ月～1年間と長期のもので研究開発を行なうものや、ジョンソンのマーケティング、日立製作所などの営業・経理・法務・人事を経験するものが多く、2～3週間程度が主流である。採用人数については、松下電器産業の150人からWOWOWの4人など様々であるが、平均して20～30人である。大学側の積極的な取り組みは、平成10年度143校(23.7%)平成11年度186校(29.9%)平成12年度250校(38.5%)と徐々に増加傾向にある。(日本経済新聞 2001年3月28日、フジゼロックスホームページより)

から、労働の流動性が低く、その反面、賃金や労働時間の変化が激しいことで安定性を保ってきたのである。特に日本は企業特殊のスキル⁶を蓄積していく長期雇用体制を形成していると考えられており、日本の高度成長を支え、1980年代世界が注目していた独自の雇用形成されたといわれる。そこで、本論文は時系列分析を主体とした、労働市場のマクロ分析をおこなう。

1章. 労働市場におけるマクロ経済の分析

1章1節 日本の労働市場における経験的背景

この1節では、労働市場を構成する基本的事実の変化と、景気変動（GDP）と失業率や労働時間などと相関があるのか概観する。その結果によっては、本論での企業内労働保蔵という観点からも補完できる。まず、2001年12月には完全失業率は5.5%と過去に稀に見る高い水準を示し、平成不況の長期停滞を示している。特に今後日本は、少子高齢化による人口構成は経験的事実であり、労働市場の先行き懸念が心配される。1985年から2001年までの変化をみると、0歳～14歳までの非労働人口は、2,603万人から1,828万人と半分近く減少し、65歳以上の高齢者は、1,246万人から2,286万人と1.8倍程度増加している。以下表1・1より、人口比では、1998年に逆転している。労働人口比率にしても、男子の労働

表1・1 年齢(0-14歳、65歳以上人口)、男女労働力の人口比率

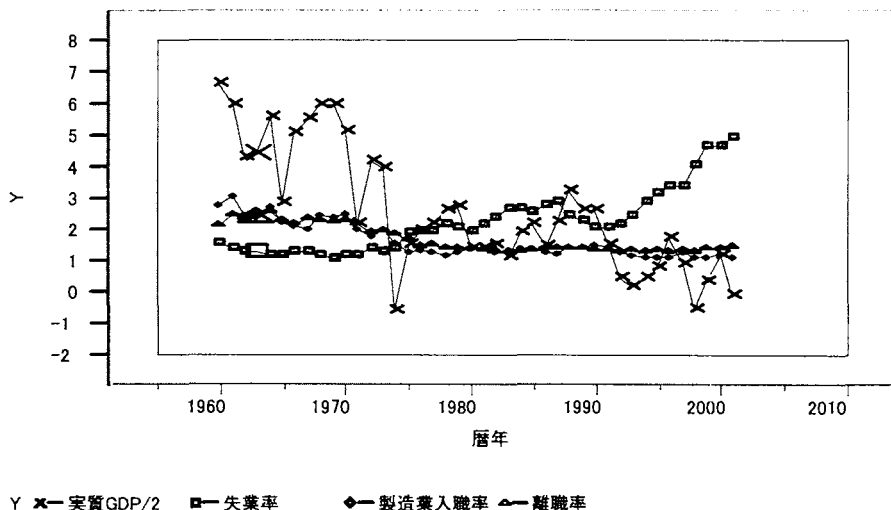
暦年	0-14歳人		65歳以上		男子労働人		女子労働人		
	口比率	人口比率	口比率	人口比率	口比率	人口比率	口比率	人口比率	
1982	23.0	9.6	61.0	39.0	1993	17.2	13.0	59.3	40.7
1983	22.5	9.8	60.5	39.5	1994	16.7	13.5	59.5	40.5
1984	22.0	9.9	60.4	39.6	1995	16.3	14.0	59.5	40.5
1985	21.5	10.3	60.3	39.7	1996	15.9	14.5	59.5	40.5
1986	20.9	10.6	60.2	39.8	1997	15.6	15.1	59.5	40.5
1987	20.2	10.9	60.1	39.9	1998	15.0	15.3	59.3	40.7
1989	19.5	11.2	59.7	40.3	1999	15.1	16.2	59.3	40.7
1990	18.8	11.6	59.6	40.4	2000	14.8	16.7	59.4	40.6
1991	18.2	12.0	59.4	40.6	2001	14.6	17.3	59.3	40.7
1992	17.7	12.6	59.2	40.8	2002	14.4	18.0	59.1	40.9

総務省統計局「人口統計月報」「労働力調査」より、年齢別人口／総人口にて算出、男子・女子労働人口／労働力人口(就業者数+完全失業者)にて算出

5. Hashimoto and Raisian(1985)(1992)より、企業特殊スキル(勤続年数)は、他の企業に転職するとその経験が失われてしまうスキルとして、一般的スキル(年齢-教育年数-6)と区別される。J.Mincer(1974)に詳しい。

市場参加が減少、女子は上昇傾向にある。この点は進学率の上昇や、男女機会均等法や育児法の法改正も含め、男女共労働市場の参加率に影響しているであろう。以上のように、時代とともに変化が見取れるのであるが、景気の変動によって、失業率や雇用などは変化していたのであろうか、図1・1によると

図1・1



* データについては、『経済統計要覧』2002 CD-ROM版と、『世界の潮流』2002を使用した。

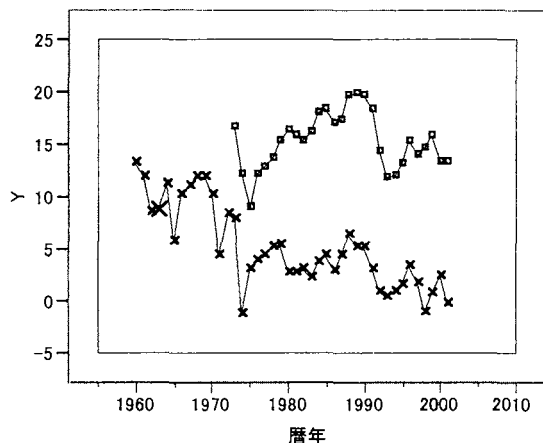
景気変動によって、失業率や雇用の流動性を示す離職・入職率においては、相関がないように感じられる。表1・2において簡単な多変量解析を行なった。景気と失業率において、相関係数が -0.475 であり相関がないと判断できることから説明できない。では入職・離職が労働の流動性への代理変数と捉えるならば、労働力の需要である入職についてはどうであろうか。中村(2002)は長期的雇用関係の変化を類推する為、失業と全産業の入職率と離職率をグラフ化している。高度成長期を通して雇用者の長期勤続化が進行するとともに、定着化も進行し、その傾向は変わってないとし、バブル崩壊後も失業率とは逆に安定している。つまり若年層と高齢者層の失業が高い要因としている。本論文での実証では、製造業の入職・離職率を用いて行なっているが、図1・1から、中村と同様に低位安定を示している。相関係数においては、景気変動と入職率が 0.5439 と高く、離職率が 0.1873 と比較すると低い為、景気好況期には積極的採用を行い、不況期には解雇を行なわない、労働保蔵が行なわれていると考えられる。但し、1990年代からその間隔に開きが生じ、企業は解雇(離職)をせざるを得ない状況と判断される。更に、景気変動において製造業は、所定外労働時間において調整していることが、相関係数 0.5104 であることからわかる。図1・2からも明らかにも景気変動と同調していることが示されている。大日(1995)によるとGNPと労働者数の変化の相関は低く、労働保蔵の傾向と労働時間との相関が高いことが示されている。

表1・2多変量解析

相関係数

	実質GDP成長率	失業率	製造業入職率	離職率	製造業所定外労働時間
実質GDP成長率	1.0000	-0.4758	0.5439	0.1873	0.5104
失業率	-0.4758	1.0000	-0.6613	-0.4186	-0.0759
製造業入職率	0.5439	-0.6613	1.0000	0.6486	0.3991
離職率	0.1873	-0.4186	0.6486	1.0000	-0.2684
製造業所定外労働時間	0.5104	-0.0759	0.3991	-0.2684	1.0000

図1・2



Y x—実質GDP成長率 □—製造業所定外労働時間

1章2節 UV分析

製造業の入職・離職は労働市場の需給を示す指標の一つであるが、UV分析によって、より詳細に分析できる。UV分析とは、労働市場の需給状況の変化を表す指標の一つであり、労働力供給を雇用失業率（Unemployment rate）で表し、労働力需要を欠員率（Vacancy）で表すことで、失業を需要不足型と構造的失業に分析できる。構造的失業とは、欠員が増加しても失業率が減少しない状況での失業であり、企業と労働者の技能・経験などのミスマッチから生じる失業や、転職がすぐに欠員補充という形では達成できない時間的ラグが生じる失業である。失業率と欠員率が等しくなることは、労働市場の需給が均衡していることなので、その時に発生する失業率は、構造的失業と解釈することができる。そしてプロットしたデータから以下のことが読み取れる。推計データとして、2つの異なったデータを用いた。

厚生労働省『職業安定業務統計』完全失業率と求人倍率と、雇用失業率⁷と欠員率⁸（労働力調査）のデータを用いて推計した。

U=V 労働需要供給均衡

U↑>V↓であるならば、生産量・労働の需要不足による失業の多い労働市場

U↑=V↑であるならば、労働者と企業におけるミスマッチが引き起こす構造的失業労働市場

U↓<V↑であるならば、需要過多であり、労働力不足である労働市場（バブル期）

U↓=V↓であるならば、完全雇用状態である労働市場（≠0）

下記図1・3をみると、2つのデータを時系列の動きを辿ってみると、大体同じ動きをしている。ただし完全失業率と求人倍率のデータは、まとまったデータではあるがその計測危うさを指摘する点もあり、⁸参照 雇用失業率と欠員率での分析が実経済を表している数字とよいて判断する。1963年～80年頃は、失業率も欠員率も低く、労働の需給は均衡していたことが伺え（左下枠）、1989年のバブル期は失業率が低く、且つ欠員率が高いという労働不足であったこと（右下枠）が示されている。そのバブル崩壊後の1991年頃から、右下から左上方向（1995年）へ向かい、経済不況化における生産や労働需要の不足からくる失業率の高さが伺える。2000年頃より右上方向に転じ、高失業率で職を求めているが、希望する仕事とのミスマッチが生じてきていることがわかる。

図1・3

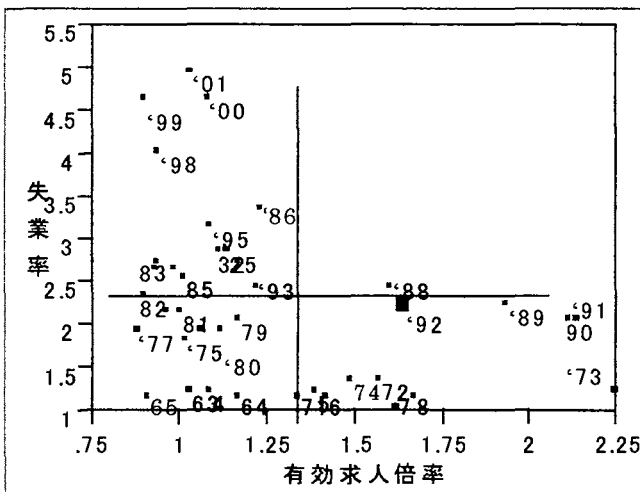
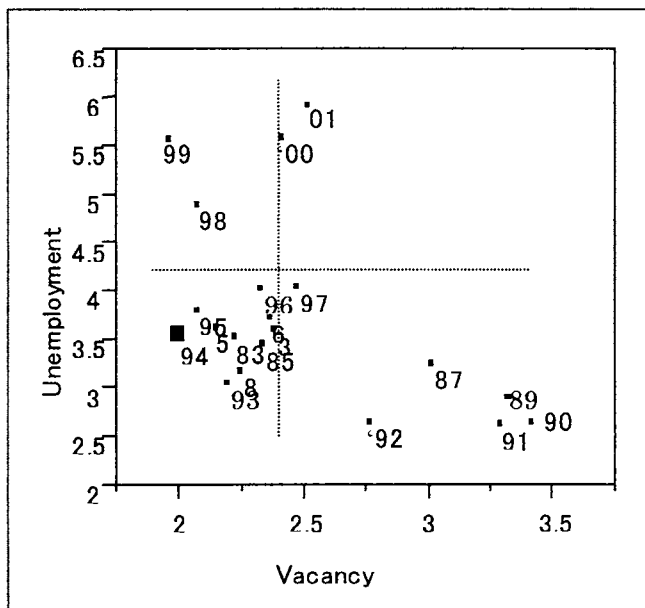


図1・3において、データに関し上図は、1963年から2001年までの暦年データを使用し、『経済統計年鑑』のCD-ROMデータから作成した。下図は年内閣府政策統括官『世界の潮流2002』のデータ1980年から2001にて作成した。

7.8. 労働省による雇用失業率、欠員率は以下の計算で行なっている。

雇用失業率=完全失業者数/(完全失業者数+雇業者数) 欠員率=(有効求人数-就職件数)/{(有効求人数-就職件数)+雇業者数}。有効求人数-就職件数は未充足求人数としても表されている。厚生労働省の職業安定所による求人倍率は、昨今のインターネットや人材派遣業が発展したことで、正確さという点で欠けている可能性がある。また完全失業率も完全失業者/労働者数であるため難点があると判断した方が妥当である。しかしまとまったデータとしては有効である為、本論文にも使用した。



2章 マクロ分析

2章1節 総供給曲線

ケインズ型の考え方より、総供給曲線は物価水準(P)と産出量(Y=GNP)の関係を表すものであり、総需要関数にもよることから、その傾きを時系列に計測することで、日本の労働市場の変化を実証できる。ただし日本がどのようなものであるかは比較対象を必要とする為、米国を対象とする。特に総供給曲線の傾きが、フィリップ曲線(賃金調整速度)の傾きをオークン係数で除したものと等しいことから、労働市場の変化に数字として何か見えてくるはずである。まず、総供給曲線を分析し、オークン係数、雇用調整係数を推計する。まず総供給曲線が右下がりであることから簡単に説明するならば、中谷(1981)を始め多くの教科書に取り上げられているが、黒坂(1988)、黒縄(1998)、坂井(1998)によると、

$Q=F(N)$ subject to $F' > 0$ $F'' < 0$ Q =生産量、 N =労働投入量である生産関数であり、 N による1階微分が正で、2階微分が負であることは労働の限界生産物MPL(Marginal product of labor)が労働投入量を増加するとともに逓減する。

生産物の価格(P)と名目賃金(W)が与えられている時の利潤(π)は

$$\pi = PF'(N) - WN \quad \dots \dots (1)$$

であり、利潤最大に行動するならば、 N で微分する。すると

$$PF'(N) = W$$

で表され、生産物1単位追加生産されるときに、得られる価格(物価水準)と限界生産物(労働投入量)の限界費用が導きだすことができる。

$P = W / F'(N)$ と変形できる。価格(物価水準)は名目賃金に労働投入量の逆数を掛けたも

のである。つまり、賃金を一定としたと仮定すると、物価水準の上昇によって雇用と生産量 (FN) は増加する、総供給曲線は物価水準と産出量において右上がりである。

そこで坂井や黒坂は、名目 GNP と実質 GNP と GNP デフレーター (物価) のデータを用い、名目 GNP の変化によって、総需要曲線が総供給曲線上を動くならば、物価と名目 GNP の弾力性と実質 GNP と名目 GNP の需要量 (生産量) の弾力性の比が総供給曲線の傾きを求めることができるとした。総供給曲線の推定式は様々にあるが、黒坂(1988)は、日本の失業率が低いことから、フィリップ曲線の傾きとオークン係数の逆数により、総供給曲線の傾きが推計できる、Dornbusch and Fischer 流の求め方¹より、下記の式による名目 GNP と実質 GNP、GNP デフレーターより求めた方がより好ましいとしている。

$$\log P(\text{GNP デフレーター}) = \alpha + \beta \log Y(\text{名目 GNP}) \cdots \cdots (1)$$

$$\log E(\text{実質 GNP}) = \gamma + \delta \log Y(\text{名目 GNP}) \cdots \cdots (2)$$

によって、両辺を Y (名目 GNP) で微分すると、 δ / β が実質 GNP の GNP デフレーター (物価) の弾力性であり、総供給曲線の傾きである。³

この結果はどう解釈したらよいだろうか。まず米国との比較から行なうと、黒坂(1985)や黒縄(1998)との数値の違いは推計年代やデータ数、四半期か暦年による違いによるが、日本の方が傾きはキツイ (大きい) ことが解る。但し、黒縄に関しては、暦年データで推計しデータ数上の制約からか 1970 年代以降傾きは米国と比較し緩いという結果となっている。では傾きが大きいということはどういうことであろうか。右下がり方向への供給曲線の移動がある場合 (ショックなど⁴)、価格や生産 (GNP) への影響が小さいことであり、上述にも示したが、総供給曲線は労働投入による限界費用 = 賃金と一致する。つまり景気 (産出量) によって、物価の上昇 = 賃金の上昇などの影響は生じないことを示している。逆に需要曲線はショックに対し、生産 (GNP) よりも物価の変動がおおきくなる。つまり現在のデフレ下においては、需要ショックが大きいと考えられる。以上から日本は供給サイドに優位であり、労働の限界生産性は逓減することから、生産の波に対して、伸縮的な賃金調整を行ないながら、固定的な雇用体制を築き、労働力保蔵を行っていたと解釈される。企業では不況期には解雇することはせず、遊休していることで、日本的と呼ばれる年功序列制や終身雇用制によって経済が成り立っていたことがわかる。

では、日本の時系列でみた場合はどうであろうか、全期が傾き 2.498 であり、バブル崩壊前

1.ケインズ型とは、『一般理論』第 5 編 20 章にて、「有効需要の変化に対する物価の弾力性と産出量の和は 1 に等しい、有効需要はある部分は産出量や物価に影響を及ぼすことから影響力は消失する」ことからである。

2.Dornbusch and Fischer のロジックに関しては、坂本・中山他訳『マクロ経済学 (下)』や、黒坂(1988)、黒縄(1998)を参照されたい。

3.両辺を Y で微分すると(1)式は $d \log P / dY = dP / P / dY = \beta \cdot (d \log Y / dY)$ 、(2)式は $dE / dY = dE / E / dY = \delta \cdot (d \log Y / dY)$ によって、 $dP / P / dE / E = \delta / \beta$ が導き出される。

4.石油危機などの供給ショックによって、物価高と失業によるスタグフレーションなどである。

表 1-3

日本とアメリカ—総供給曲線の傾き—

日本 暦年						アメリカ			
	β	δ	δ/β (傾き)	R^2	n	β	δ	δ/β (傾き)	R^2
1980・1期～200 1・3期 (全期)	0.287 (26.51)	0.7172 (93.63)	2.498	0.89 0.99	87				
1980・1期～1990・ 4期	0.3384 (14.27)	0.6774 (49.02)	2.001	0.84 0.98	40	0.4456 (26.2)	0.4458 (26.14)	1.000	0.82 0.99
1990・4期～2001・ 3期	0.1394 (1.80)	0.8750 (15.10)	6.276	0.67 0.83	47	0.4336 (6.82)	0.6271 (37.65)	1.446	0.94 0.94

* 日本のデータは東洋経済新報社『経済統計年鑑』、内閣府『世界の潮流2002』の季節調整済四半期データを使用した。データの加工分析を行なった。決定係数 R^2 に関して項の上が(1)式、下が(2)式の決定係数である。

* ()内はt値であり、日本の β 値0.139以外は、すべて5%有意である。

黒坂(1988)の推計結果

日本 暦年						アメリカ			
	β	δ	δ/β (傾き)	R^2	n	β	δ	δ/β (傾き)	R^2
1960～1985	0.3592 (24.54)	0.6481 (43.69)	1.8042873	0.99 0.99		0.4766 (17.75)	0.5245 (19.45)	1.1005036	0.99 0.98

武縄(1998)の推計結果

日本	δ/β (傾き)	アメリカ	δ/β (傾き)
1970～1979	2.11083	1970～1979	2.009
1980～1989	0.46772	1980～1989	1.2171
1990～1995	0.81917	1990～1995	1.2559

後で推計してみたが、バブル崩壊前後は、黒縄と同様傾きはキツくなるという結果である。(2.0→6.27、0.46→0.81) この解釈については、景気が好況時の方が、転職リスクが少なく労働需要が大きい、バブル崩壊後は、需要サイドによる価格＝賃金の伸縮的効果が大きくなり労働者側からも移動しないという伸縮的賃金・固定的な労働市場体制であると考えられる。但し、

供給曲線の傾きだけでは言及できないので、以下の分析と合わせて考える必要がある。供給曲線の傾きがキツイことは、需要ショックは数量よりも価格によって吸収される度合いが大きくなることは述べたが、1990年代におけるデフレ下の経済においては、名目貨幣量増大により迅速に物価上昇をもたらすマネタリスト的な価格調整メカニズムが有効であると付け加えておく。

2章2節 オークン係数

オークン係数はいわゆる生産の変動（実質 GNP）に対して雇用調整（完全失業率）がどれだけ行なわれるかの調整度の逆数である。坂井（1998）は実質 GNP 成長率と失業率のデータを使用しており、浜田・黒坂（1984）においても推計方法として詳しい。本論文は、浜田・黒坂（1984）の推定式を使用する。

$$\ln(100-u) = \alpha + \beta_1 \ln Y + \beta_2 t$$

u=完全失業率 Y=実質 GDP t=代理変数（タイム・トレンド）

(100-u) が対数変換した雇用率であり、1%雇用率が変化すると、β%産出量に変化する。その係数βの逆数が雇用の弾力性であることから、係数の逆数がオークン係数となる。オークン係数が大きい程、雇用の調整がされず産出量に対して雇用の保蔵度＝労働者の安定性が高いことがいえる。逆に低い（低下傾向）にあるならば、調整がされる＝流動的になってきたと言えるのである。

表1・4の結果より、1980年代は米国の方が労働保蔵の傾向を示し、内部労働市場を形成していたことが伺え、その他の期間は日本の方が係数は高い。また日本の時系列を見ても、1980年前半から1990年バブル崩壊前までは、労働保蔵による内部労働市場を形成する傾向を示し、バブル崩壊後の1990前半は景気が下がりつつも、解雇など雇用調整にタイム・ラグが生じることから、産出量が小でも、低失業率により16.6という数字が算出されたのである。労働保蔵＝安定性を余儀なくされたのである。1995年以降は1980年代の数値に近似し雇用調整が弾力的傾向とみることが出来る。さらにフィリップス曲線の傾きが、総供給曲線の傾きとオークン係数の積で導き出せることから、その係数は1980年代と1990年代を比較するに約6倍（12.9→73.8）、アメリカは約3倍（6.9→17.1）であった。フィリップス曲線とは、失業者で代理される労働市場の需給状態が名目賃金に及ぼす効果を示す非線形の曲線であることから、名目賃金の伸縮性、硬直性による賃金の調整速度を表すものとして使われる。日本の結果は6倍（12.95→73.84）ということであるから、労働の安定性や内部労働市場を形成しながら、名目賃金の変化が大きいという形で賃金調整がされていることが示される。また昨今、実力主義や能力主義に代表される雇用システムの推移をみる限り名目賃金の変化は、雇用の流動性を示すことを示す要素とも捉えることができるのではないかと考える。

表 1・4

オークン係数の推計結果とフィリップス曲線の傾きの導出

日本

	全 期	前 期	後 期	①	②	③	④
	1980・1～2001・3	1980・1～1990・4	1991・1～2001・3	80・1～'85・4	86・1～'90・4	91・1～'95・4	95・4～'01・3
β	0.1168	0.1545	0.085	0.1288	0.1069	0.06	0.1169
t値	(26.06)**	(19.29)**	(3.76)**	(4.08)**	(3.05)**	(0.99)	(3.06)**
$1/\beta$	8.5616438	6.47249191	11.7647059	7.7639752	9.354537	16.666667	8.5543199
R^2	0.954	0.902	0.971	0.899	0.917	0.918	0.927
δ/β	2.498	2.001	6.276				
フィリップ	21.39	12.95	73.84				

アメリカ

	1980・1～1989・4	1991・1～2001・3
	β	0.1441
t値	(8.79)**	(5.47)**
$1/\beta$	6.93962526	11.8343195
R^2	0.67	0.74
δ/β	1	1.446
フィリップ	6.94	17.11

黒坂(1988)による推計結果

	日 本	ア メ リ カ
	期 間	期 間
	1960～1985	1960～1985
β	0.0271	0.359
t値	(10.29)	(5.56)
$1/\beta$	36.900369	2.78551532
R^2	0.97	0.95

* データに関しては、東洋経済新報社『経済統計年鑑 2002』CD-ROM、内閣府『世界の潮流 2002 春』より四半期データを使用した。またフィリップ曲線は、総供給曲線の傾き×オークン係数で求められることから導出した。黒坂(1988)参照

2 章 3 節 雇用調整係数の推計

景気変動によって生産量が調整される時、どうしても雇用量の影響には時間的ラグが生じる。前節ではオークン係数を用いて、生産量に対する雇用の調整（保蔵）という観点から述べた。雇用調整係数は労働需要係数とも呼ばれ、実際には計測できない最適雇用量 L^* を達成するため、ラグ付き内生変数として、1期前の雇用量 L_{t-1} と、現在の雇用量 L_t とを用いることで、雇用調整速度 λ を求めることである。つまり雇用調整には、雇用調整コスト⁵や人員整理などのタイム・ラグが生じる為、1期前の雇用量のある部分 λ でしか調整できないことから、部分調整モデルと呼ばれる推計式を用いる。山本(1995)や坂井(1998)⁶は計量分析の立場から定式化し、黒坂(1985)理論と分析を行なっている。野田(2002)は、赤字企業の雇用調整速度を計測している。本論文は黒坂のモデルを使用している。

最適雇用量は 当期の生産量とで決まる生産関数を考えると良い。生産量とは、賃金と労働時間としたならば、対数変換した次式が導かれる。

$$L_t^* = \alpha + \beta Y_t + u \quad \dots \dots (1)$$

$$L_t^* = \alpha + \beta Y_t + W_t + LT + u \quad \dots (1)'$$

1期だけのラグを用いた部分調整モデル

$$L_t - L_{t-1} = \lambda(L_t^* - L_{t-1}) \dots \dots (2)$$

(2)式に(1)式を代入すると

$$L_t = \lambda\alpha + \lambda\beta Y_t + (1-\lambda)L_{t-1} + \lambda u \dots (3)$$

変数などは坂井(1998)を用いるが、特に日本の経済基盤である製造業の各データを使用することで雇用調整の現状を示すことができる。また変数 Y である生産指数に関しても、データの制約上製造業が望ましい。データは、『経済統計要覧』『日本経済統計年鑑』の1962年から2001年の四半期データ(季節調整済)を使用した。

推定式

$$\ln L_t = \alpha + \beta \ln Y_t + \ln(W_t / P_t) + (1-\lambda)L_{t-1} + T \dots \dots (4)$$

L =製造業常用雇用指数(30人以上) Y =鉱工業生産指数 W =製造業賃金指数(30人以上)
 P =消費者物価指数 T =タイム・トレンド

5.雇用調整費用の存在、つまり募集費や新規採用時の担当官や採用者へのOJTによる指導者の機会費用など多くの費用が必要である。その為、需要の減少としても直ぐに解雇することは将来の「販売可能性」の危険性や雇用の不安定性からくる生産性の低下の危険性も吟味しなければならない。

6.計量分析の立場から詳しく解説されている。坂井(1998)は黒坂(1985)のモデルに近い。

表2・5

雇用調整速度係数の推計結果

期間/係数	α	β (lnY)	log(W/P)	λ	T	1- λ	R ²	n
1973.1~2001.1	0.2711 (3.97)**	0.0746 (8.91)**	-0.05 (-13.88)**	0.9228 (54.93)**	-0.0003 (-5.31)**	0.0772	0.977	115
A) 1973.3~1984.4	0.4619 (3.45)**	0.0837 (4.27)**	-0.0477 (-9.68)**	0.8722 (28.03)**	-0.0004 (-2.37)**	0.1278	0.942	66
B) 1990.1~2002.1	0.7268 (2.88)**	0.0557 (2.64)**	-0.0526 (-10.59)**	0.85321 (20.60)**	-0.0006 (-3.87)**	0.14679	0.991	48
I '73.3~'83.4	0.5633 (2.82)**	0.0704 (3.43)**	-0.0446 (-7.54)**	0.8592 (18.93)**	-0.0004 (-2.00)**	0.1408	0.958	43
II '84.1~'93.4	1.1827 (3.54)**	0.0916 (4.28)**	-0.0589 (-8.96)**	0.7105 (9.62)**	0.0003 (-1.45)	0.2895	0.961	40
III '94.1~'02.1	1.1437 (2.76)**	0.0784 (4.54)**	-0.0439 (-10.18)**	0.7467 (9.03)**	-0.001 (-3.18)**	0.2533	0.994	32

* 推計データは、東洋経済新報社『経済統計要年鑑 2002』CD-ROM より、季節調整済四半期データを使用した。期間の年度横の数字は、1~4 期を示している。W=製造業賃金は、消費者物価指数で除し実質化してある。

*.マーク** は 5%有意である。

以上の結果から、黒坂(1988)のアメリカのデータを借りるならば、1960年~1985年データで速度係数は、0.66であり日本はその時期に照らし合わせると0.140~0.289であるから、アメリカと比較することで日本は、かなり小さいことが解る。つまり生産変動に対して、雇用調整をする中で最適な雇用調整をアメリカは6%雇用調整を行なっていることを示し、日本は最適雇用に14%~28%されている=雇用調整を景気に対して行わないことを証明している。時系列でみるとバブル崩壊前後には、雇用調整が活発であったこと、そして1995年以降も景気変動(生産量)に対して、雇用の調整が製造業の分野で行なわれていることが示されている。ただし、現在も終身雇用制度が多く企業の企業でみられ、中村(2002)のように高齢者や若年層に雇用調整が行なわれているとも考えられる。

まとめ

バブル崩壊による雇用の構造変化において、依然日本はバブル崩壊時のような景気後退期を経験している。そしてその構造変化はむしろ、労働市場と景気にはタイム・ラグがあることから現在も続いているようである。ただし1990年を堺に変化が生じているのは上記分析から明

らかである。日本の労働市場の歴史にも構造変化の生じた時期は依然あった。1970年代の2度に渡る供給ショックである石油危機であるが、黒坂(1988)は、オークン係数の労働生産性の観点から論及している。石油危機以後から労働供給の調整として、労働保蔵を行ないながら労働時間による雇用調整を行なうようになったというこの時に構造変化が生じたと述べている。図1・2からも支持される。

本論文からの結論として、供給曲線、オークン係数、フィリップス曲線、雇用調整速度を分析して、時間・名目賃金調整をすることで、日本は労働保蔵をしながら安定的である日本的と呼ばれる雇用体制を築きあげてきたようである。ただしバブル崩壊後の景気低迷は、雇用のミスマッチを生じさせ構造的失業が多くなってきていること、さらに雇用調整速度が速くなりつつあり、流動的にならざるを得ない状況にあると考えられる。実力主義や能力主義による企業の利潤中心的な労働市場を形成しながらの労働市場の柔軟性・弾力的なものと、企業の基幹的な労働力保蔵(安定性)を形作る雇用システムの共存がもたらされる労働市場と考えられる。ただしその結論にはミクロ的な分析も必要とする。本論では述べないが、時系列の賃金関数分析により、勤続年数や年齢などの稼得(賃金)の効果は下がってきているし、転職コストも下がってきている。少子高齢化を迎える日本において労働問題は重要な課題であり、ライフサイクルにおいても大きな要素を閉めている。今後更に集計データによるマクロ的分析と個票データによるミクロ的分析の両方の検討が必要であろう。

Reference

- 荒井一博、大橋勇雄 『労働経済学』 有斐社 1989
- 黒坂佳央 『マクロ経済学と日本の労働市場供給サイドの分析』 東洋経済新報社 1988
- 島田晴雄 『労働経済学』 岩波書店 1986
- 大日康史,有賀健 人的資本の形成と労働保蔵～RBC理論の日本労働市場への応用～
1995 「ファイナシャル・レビュー」May1995 大蔵省財政金融研究所
- 山本拓 『計量経済学』 株式会社 新世社 1995 pp.177～
- 野田知彦 「労使関係と赤字調整モデル」 経済研究 Vol.53, No.1, Jan.2002
- 奥西好夫 「アメリカの労働経済学」 『日本労働研究雑誌』 1996, No.431 pp.43-51
- 坂井吉良 『SASによる経済学入門』 シーピーエー出版株式会社 1998
- Mincer.J(1971) 「Schooling Experience and Earning」 NBER and Columbia University Press 1971, pp3, pp11, pp83～
- Gary S Becker(1975) 「Human Capital A theoretical and empirical analysis with special reference to education」 佐野陽子訳 東洋経済新報社 1975
- Gary S.Becker 「Investment in human capital:A theoretical analysis」
- Hashimoto Masanori and Raisian John(1985) 「Employment Tenure and Earnings Profiles in Japan and United States」 『Amerikan Economic Review』 Vol.75, no.4, September, 1985
- Msasanori Hashimoto and John Raisian 「Emplyment Tenure and Earnings Profiles in Japan and United States:Reply」 『Amerikan Economic Review』 Vol.80, no.1 March 1992

日本SASユーザー会（SUGI-J）

アジルな Supply Chain を実現する予測プロセスの自動化

－SAS® High-Performance Forecasting のご紹介

松舘 学

SAS Institute Japan 株式会社

カスタマーサービス本部 プロフェッショナルサービス第2部

Automation of Forecasting Process enabling Agile Supply Chain

－ Introduction of SAS® High-Performance Forecasting

Manabu Matsudate

Professional Service No.2 Dept. Customer Services Div.

SAS Institute Japan Ltd.

要 旨

時系列予測にまつわるあらゆるステップを自動化し、サプライ・チェーン運営の高速化を実現する SAS High-Performance Forecasting を実現するテクノロジーとプロシジャの利用法の紹介。また、予測自動化のビジネス応用事例として SAS® Supply Chain Intelligence ソリューションを取り上げる。

キーワード： サプライ・チェーン、予測の自動化、SAS High-Performance Forecasting, HPF プロシジャ

1. はじめに: SAS Supply Chain Intelligence

昨今の厳しい経済環境の中で、企業は競争優位の確立のため、さまざまなビジネス上の戦略を模索している。その中でも、製造・流通・消費財業界では、サプライ・チェーン・マネジメント(以下SCM)改革の先進事例に注目が集まっている。

サプライ・チェーン・マネジメントとは、企業内はもちろん、企業間のコラボレーションによって「調達」「生産」「在庫管理」「輸送」「販売」という全てのビジネス・プロセスを統合することによる「全体最適」を目指し、競争優位を築こうとする試みである。SAS社は、SCMにおける計画業務に欠かせない解析・予測・最適化に関わるケイパビリティを提供し、SCMをさらに支援・強化する、SAS Supply Chain Intelligenceを提唱している。

SAS Supply Chain Intelligenceは、サプライヤ戦略を最適化するSAS® Supplier Relationship Management、生産・品質工程管理のSAS® Process Intelligence、コスト管理のSAS® Value Chain Analytics、需要予測・在庫管理・価格最適化のSAS® Demand Intelligenceの4つのソリューションから成るビジネス・スイートである。

SAS Demand Intelligence は、需要予測のSAS® Demand Planning、在庫管理・補充コストを最小化し、サービス・レベルにあった在庫補充を提供するSAS® Inventory Replenishment Planning、価格や値下げの最適化を行なうSAS® Price Optimization からなる。

消費者の趣向が多様化し、商品数が増加し、商品のライフサイクルが短期化する今日、生産計画・在庫計画の立案は戦略的に重要な課題である。単に統計ツールを利用して、需要予測の精度を上げていくことだけでなく、需要動向を常時把握して、予期しない販売動向に俊敏に対応できるように、柔軟なサプライ・チェーンを構築していくことがキー・サクセス・ファクターとなる。

SAS Demand Planning は、数千・数万に及ぶ膨大な商品に対して高速かつ高精度の予測を自動化することで予測担当者の作業を大幅に軽減する。同時に、特殊な需要パターンを示す商品など、商品特性に応じた高度な解析手法によって対話型の予測を可能にしている。予測の自動化には、SAS High-Performance Forecasting がコア技術として利用されている。本稿では、特に予測の自動化技術にフォーカスし、SAS High-Performance Forecasting(以下 HPF)で用いられる大容量・高速・高精度の自動予測のアルゴリズムとサンプル・プログラムを紹介する。

2. 予測の自動化

2-1. SAS Demand Planning

SAS Demand Planningはビジネス上のあらゆるレベルの意思決定をサポートする、スケーラビリティのある需要予測を可能にする。配送センターから店舗という地域軸、商品分類コード・SKUなどの商品軸まであらゆるレベルで、日々の販売状況を自動かつ正確に需要予測を行なう。また、新商品・短ライフサイクルの商品など、予測の難しいものについては、対話型の予測インターフェースを利用でき、SAS社の培ってきた高度な分析力を活かして正確な予測を可能にした。また、セールス・プロモーションの効果も加味してシミュレートできる。このように、SAS社の提供する最適化手法を用いて、1つ1つの対象に対して需要予測を行なうため、どんなに詳細なレベルでも需要の動向を把握できる。加えて、予期しない需要変化の対応するため、エージェントが常に予測精度をモニタリングしており、一定の精度を下回ると予測担当者にメールなどでアラートを発する。これにより、膨大な商品数の予測精度の監視作業を軽減し、システムによる自動運転のリスクを軽減する。

顧客のニーズに合致したサービスを提供するために、サプライ・チェーンにおいて重要なフェーズである需要予測を正確に実現するSAS Demand Planningは、企業の収益性向上に多大な貢献をする。

2-2. 予測自動化の必要性

予測は、企業内計画プロセスにおける意思決定の根拠となる。資金調達、生産・在庫計画、資材配分、予算、販売ノルマ、キャンペーン、調達活動などの決定は、将来の予測に基づいている。根拠となる予測がより正確であればあるほど、より高度な意思決定が可能になる。

従来型のサプライ・チェーンでは、需要予測が工場や配送センター、店舗と別々のユニットでそれぞれ独立して行なわれており、その予測も経験則に基づくものであるなどデータによる根拠に乏しいものである。このため、店舗などの下流での需要予測の変動が、上流である工場に至る頃には増幅して伝わり、過剰な在庫・仕掛り在庫の増加をもたらし、企業の収益を圧迫している。この効果を鞭のしなりに例えてブル・ウィップ効果と呼んでいる。在庫は帳簿上流動資産に計上されるが、簡単にキャッシュ・フローを生み出すことは困難であると同時に、在庫の維持費などの間接コスト増大を引き起こす。

ブル・ウィップ効果による見込み発注を回避し、過剰在庫を削減するための正確な需要予測を行なうには、サプライ・チェーンのプロセスを統合し、企業内だけでなく企業間のコラボレーションを通じて予測情報を共有することが理想的である。需要予測における課題は、こうした情報の共有化のハードルだけではなく、膨大な商品数の存在もあげられる。予測を行なう対象商品が数万を超える場合があり、これが障害となり全ての予測

を正確に行なえないことがある。

実際こうした、生産・在庫・販売計画にあたって ERP や POS システムによって蓄積されたトランザクション・データを利用して、数千・数万にも及ぶ対象の需要予測を行ないたいというビジネス・ニーズは多い。予測の対象が僅かであれば、熟練した分析者は、これまでの業務上の経験や勘を活かし、予測ソフトウェアを利用してさまざまな時系列モデルを適用しながら、精度の高い予測を行なうことができる。実際 SAS 社では SAS/ETS などの高度な統計処理のための諸機能を提供してきた。

しかしながら、コンビニエンス・ストアのように商品数が平均約3000品目を超えるような大容量の予測となると、上記のような、分析者が一つ一つの商品に対して予測を行うのは非常に労力の強い作業となる。このような場合、予測精度の低下を抑えつつ、データの最適化や時系列分析をある程度の自動化が有効である。

実際に予測の自動化が必要となってくるのは、次のようなケースが考えられる。

- 予測対象が数千・数万にも及ぶ
- 頻繁に予測を行なう必要がある
- 予測に利用したいデータがあらかじめ時系列データになっておらず、データ加工を行なう必要がある

こうしたニーズに応えるため、SAS 社は SAS High-Performance Forecasting を開発した。HPF は、予測に関連するあらゆるステップ(①トランザクション・データを時系列データに変換 ②予測モデル構築 ③最適モデルによる予測 ④予測値算出 ⑤モデル適合度検証)を自動化し、大容量・高速・高精度の予測を実現する。

もちろん、コンビニエンス・ストアのような小売店での需要予測に限らず、銀行 ATM への現金補充における需要予測など、幅広いビジネス上の活用法が考えられる。

なお、本稿では誌面の都合上、時系列モデルの統計的な解説は行なわない。時系列モデルについては、A.C.ハーベイ⁶⁾などを参照されたい。

SAS/ETS と SAS High-Performance Forecasting との比較

SAS/ETS プロシージャ	SAS/ETS 時系列予測システム	SAS High-Performance Forecasting
統計専門家	初級～中級レベルの担当者	担当者
対話式による様々な分析	対話式・自動	自動
バッチ	GUI	バッチ
精度重視・少数の対象	精度と効率	効率重視・膨大な対象

3. 予測の自動化プロセス

3-1. 予測の自動化とは

予測の自動化とは、分析者への負荷を最小限にするよう、システムが自動で予測を行なうことである。予測の自動化プロセスでは、それぞれの時系列・候補となるモデルに対して、予測結果がもっともよく当てはまるようにパラメータ推定の最適化を独立して行なう。したがって各々の時系列に対して、何種類もモデルをあてはめる。

こうした予測の自動化と、通常の前測は性格が大いに異なる。分析者が一つ一つ行う予測は、最適な予測をゴールとしている。しかし、予測の自動化のゴールは、もっとも頑丈(robust)なモデルを選び、それぞれの予

測精度がある程度得られることである。

実際のビジネスでの運用においては、商品のライフステージ、特性に応じて予測手法を適宜変えていくことが望ましい。たとえば小売業では、予測の対象の商品にランクをつけ(サービス・レベルとリンクさせてもよい)、重要度の低い商品については、自動で予測を行ない、その中で精度が低いと判断されたものに対しては、SAS/ETS を利用して、さらに精度の向上のため予測モデルの修正を行なう、というケースが考えられる。また、重要度の高い商品についてははじめから分析者が対話的に分析を行なっていく、といった商品特性別の利用が理想的である。

3-2. 予測のステップ

次に、実際にどのように予測が行なわれるかを、6つのステップにわけて見ていく。

(1) データ最適化

分析対象データを適切な形式に変換するステップであり、予測を行なう上で非常に重要なステップである。Web ログデータや POS のトランザクションなどは時間間隔が一定でないデータを、時系列モデルを適用するために、一定の時間間隔をもった時系列データに変換する。また、欠損値が存在すると正確な予測が行えないため、平均値などを利用して統計的に欠損値を推定して補完を行う。

(2) 最適モデル候補選択

アルゴリズムは、時系列データの特徴を把握し、そのデータに妥当性のあるモデルのリストを抽出するという流れになる。トレンドのある時系列にはトレンドに対応した時系列モデル、季節性がある時系列には季節性に対応した時系列モデルのリストを選択する。非線形なデータは、線形なデータに変換しなければならない。また、需要が断絶的である^{ひげ}間歇需要の場合には、クロス法を用いて予測を行なうよう指定も可能である。

経験的に、時系列データは季節性やトレンドといった観点から特徴つけることができる。したがって、時系列モデルを適用する際はこれに留意する必要がある。

また、時系列モデルを当てはめるにあたって、変換を行なう必要性が生じる場合がある。その際には自動で、対数変換・平方根変換・ロジスティック変換・Box-Cox 変換などを行ない、最適なモデルを判断するよう指定可能である。

(3) 予測モデル選択

(2)最適モデル選択のステップで抽出された妥当性のある時系列モデル候補それぞれを、時系列データに適用し、予測値を算出する。その中でもっとも当てはまりの良いモデルを1つ採択する。

モデルの選択に当たって、このアルゴリズムではデータをサブセットして抽出し、それに対してモデルを複数適用し、適合度統計量を比較することで、最適なモデルを決定する。

サブセットされるデータは、最新のデータから遡って最初の欠損値が登場するまでのデータである。なお、データ量が少なく、モデルの適合度が低いと考えられる場合は、サブセットをせずに全てのデータを利用してモデルを適用する。サブセット・データによる予測モデルの検討を行うことで、厳密な精度と予測対象データの容量をある程度トレード・オフし、予測モデル選択にかかる処理時間を短縮できる。このサブセット・データは、ニューラル・ネットワークにおけるトレーニング・データに類似するものと考えると理解を助けるかも知れない。

(4) 予測値算出

(3)のステップで選択されたモデルを、時系列データに適用し予測を行なう。その際に、前ステップで行なったようなサブセット・データだけではなく、データ全件を利用して予測値を算出する。サブセット・データはモデルの選択にのみ使用されるもので、予測モデルが一意に特定された後では予測に用いない。

また、予測に基づいて意思決定を行なう際は、予測値、信頼区間の上限信頼限界および下限信頼限界などのうち、どの幅で予測値とするか判断する必要がある。たとえば、小売店において、サービス・レベルが低く過剰在庫を抱えるリスクが大きい場合は下限信頼限界を用い、サービス・レベルが高く売れ筋商品など販売機会損失を回避したい場合には上限信頼限界を用いるなど、臨機応変の対応が必要となる。また、HPF のアウトプットを、SAS Inventory Replenishment Planning の入力データとして、在庫最適化計画と連携が可能である。

(5) 適合度検証

採用された予測モデルを用いた予測結果から、パラメータ推定値および適合度統計量を計算し、モデルの当てはまり具合を確認する。

適合度統計量(statistics of fit)は、実績値とモデルを比較することによって、予測モデルの適合度を確かめるための統計量である。平均平方誤差 Mean square error (MSE)、平均絶対誤差率 Mean Absolute Percentage Error (MAPE)、赤池の情報量基準(AIC)などを用いて比較される。

モデルの適合度を確認し、満足いくものであれば、これを予測基準値として採用する。実際の販売計画では、この予測基準値に対して、セールスプロモーションの効果などのマーケティング効果を調整して、最終的な予測値を導出する。

(6) パフォーマンス(予測精度検証)

(1)から(5)のステップまでは予測プロセスであったが、このステップは、実績値と予測値を比較するレビュー・プロセスである。予測をもとに生産・販売計画が立案され、商品が実際に顧客に販売され、その販売実績が明らかになった後に、果たして予測がどれだけ正確であったかを検討する。パフォーマンスは、次の点を考慮して行なうと良い。

- 予測精度はどうであったか
- 予測精度の低下の原因はなにか
- 予測精度が前回の予測に比べて低下した場合、何がおこったのか

精度検証をする際の判断基準として、前述した適合度統計量を用いることで、モデルどれだけ適合していたかを判断できる。また、グラフを書くことによって、実績値が予測値の信頼区間内に存在したかどうかで判断することができる。

過去当てはめたモデルにおいて精度が良好であったが、その後同じモデルで行なった予測では精度が低下していたならば、流行が廃れた、競合他社のプロモーションなどの不規則なイベントが起こった、マーケットシェアが低下した、あるいは単純にそのモデルでは適合しきれなくなった等と、他のデータを利用して判断することができる。

このように、モデルの精度を確認し、精度が悪かった場合にはその因果関係を考察し、モデルの精度を繰り返し高めていくことが業務では求められる。

4. サンプル・プログラム

次に、SAS High-Performance Forecasting の機能である HPF プロシジャを利用したプログラムのサンプルを紹介する。

サンプル・データは、小売店における商品ごとの販売履歴(1994.1~1999.1)データである。すでに時系列データに変換されている。このデータを利用して予測を行なうには、下記のプログラムを実行する。入力データセットは Sales、出力データセットは nextyear である。ID は日付変数の date を指定し、月間隔なので interval=month オプションを指定する。予測の対象が全ての商品なので、forecast=_ALL_ で指定している。

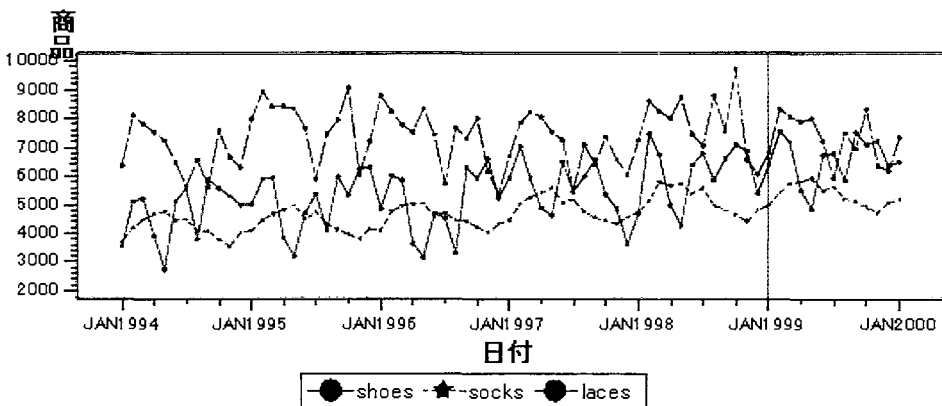
販売実績

	1-Jan-94	3557	3718	6368.8	575	987	10.82	15	102.6	12410	15013
	1-Feb-94	5128	4174	8123.2	565	1000	12.12	15.1	99.9	13556	12413
	1-Mar-94	5222	4482	7807.2	406	1005	11.78	15.3	102	11063	12752
	1-Apr-94	3925	4665	7543.2	266	1043	12.24	11.3	79.5	11799	16222
	1-May-94	2750	4759	7242	194	1074	13.46	14.7	96	14497	13622
	1-Jun-94	5117	4469	6486.8	290	1042	12.49	13.4	98	14999	13193
	1-Jul-94	5570	4497	5654.3	583	1110	14.14	21.8	104.8	15711	17201
	1-Aug-94	3812	4103	6584.5	625	1054	14.28	19	107.5	12584	9455
	1-Sep-94	5917	4076	5635.2	654	1086	13.42	19.1	115.7	13036	11649
	1-Oct-94	5575	3801	7567.1	583	1054	14.41	17.3	113.2	15307	15269

プログラムの例

```
PROC HPF data=Sales out=nextyear;
  ID date INTERVAL=month;
  FORECAST _ALL_ ; run;
```

モデルと実績値(1999年2月以降を予測)



このような簡単なプログラムで、自動的に予測モデルを当てはめ、予測値を出力することができる。オプションをさまざまに設定することで、出力データセットには、パラメータ推定値や適合度統計量、信頼区間などのデータを含めることができる。

なお、このグラフでは予測期間を指定しなかったためデフォルトで 12 期分の予測値が出力されている (1999.2 以降の部分)。

5. まとめ

本稿では、SAS High-Performance Forecasting で用いられる予測の自動化のテクニックを実際の業務に即した形で紹介した。このテクニックによって、大容量・高速の予測自動化を実現した。サブセット・データを用いた予測モデルの選択プロセスによって高速化を実現し、実際の業務予測担当者の作業を大幅に軽減し、サプライ・チェーン・サイクルのますますの短縮化の期待に応えることが可能となる。また、高度なサプライ・チェーンにおける需要予測でのキー・サクセス・ファクターは、単に予測精度を上げるのではなく、精度を常にモニタリングして変化を即時に捉えるシステムを整備しつつ、予測バケットを短期化してその変化に対応できる生産プロセスなどの社内プロセスを再構築することである。

ⁱ ドイツの4大銀行のひとつで、世界トップ30に入るコメルツ銀行は、SAS社の予測ソフトウェアを使用して、市場予測と最適化供給計画を注文処理に統合し、支店や現金支払機への現金供給を最適化するアプリケーションを開発。コメルツ銀行では、SAS社の予測ソリューションにより、それまで現金支払機(ATM)に滞っていた現金や物流コストを最大18%まで削減できた。

ⁱⁱ A.C.ハーベイ著 国友直人・山本拓訳「時系列モデル入門」東京大学出版会,1985

*SAS Publishing

[参考文献]

- SAS High-Performance Forecasting Webサイト
<http://www.sas.com/technologies/analytics/forecasting/hpf/index.html>
- SAS White Paper "Large-Scale Automatic Forecasting -Millions of Forecasts with SAS High-Performance Forecasting " SAS Publishing, 2002
- SAS High-Performance Forecasting Procedure Reference
Ref. <http://support.sas.com/rnd/app/da/new/hpf/hpf.pdf>
- What's New in SAS 9.1 SAS High-Performance Forecasting

```
PROC HPF option;  
    BY 変数名 ;  
    FORECAST 変数リスト / option ;  
    ID 変数名 INTERVAL=interval option ;
```

PROC HPF ステートメント

BACK= n 予測が開始されるオブザベーションを指定。デフォルトは BACK=0。 (9.1)

DATA= SAS-data-set 入力データセットの指定。

LEAD= n 何期先まで予測するか指定。デフォルトは 12 期。

MAXERROR= number エラーの許容範囲を指定。デフォルトは 50。

NOOUTALL 予測期間の最終オブザベーションのみをデータセットに出力。

OUT= SAS-data-set 出力先 SAS データセットの指定

OUTEST= SAS-data-set パラメータ推定値および適合度統計量を含むデータセットを出力

OUTFOR= SAS-data-set 実績値と予測値、95%信頼区間を含むデータセット出力

OUTSEANSON= SAS-data-set 季節統計量を含むデータセットを出力

OUTSTAT= SAS-data-set 適合度統計量を含むデータセットを出力

OUTSUM= SAS-data-set 各種統計量および予測値を含むデータセットを出力

OUTTREND= SAS-data-set トレンドに関する統計量を含むデータセットを出力

PLOT= option | (options) ODS 出力をカスタマイズ。9.1 では評価版。詳細はプロシジャ・レファレンス参照。

PRINT= option | (options) アウトプット画面に出力する内容を指定。デフォルトでは、画面出力されない。

ESTIMATE...パラメータ推定値適合度統計量 **SUMMARY...** 要約統計量

FORECASTS...実績値と予測値、95%信頼区間 **TRENDS...** トレンド統計量

SEASONS... 季節統計量 **ALL...** 全て出力

STATISTICS...適合度統計量

PERFORMANCE ...パフォーマンス時の統計量を表示(9.1)

PERFORMANCESUMMARY ...パフォーマンス時の要約を BY グループごとに表示(9.1)

PERFORMANCEOVERALL ...パフォーマンスの時の要約を全ての BY グループで表示(9.1)

STATES...バックキャスト、イニシャル、ファイナルのステイトを表示(9.1)

PRINTDETAILS アウトプット画面に詳しい結果を出力

SEASONALITY= number 季節サイクルを指定。デフォルトは 1。

SORTNAMES FORECAST ステートメントで指定された変数名をソートして表示(9.1)

STARTSUM=n 予測を何期目から開始するか指定する。Lead=n オプションの値と併用。

BY ステートメント

BY グループ名 グループごとに予測を行なう。

FORECAST ステートメント

FORECAST 変数リスト / option ;

ACCUMULATE= option タイムバケットの積み上げ・配分を指定。

ALPHA= number 有意水準を指定。閾値は 0~1。デフォルトは、0.05 で信頼区間 95%。

HOLDOUT= number 予測モデル選択に利用されるサンプルの期間を指定。サンプル抽出される期間は、最新のデータから欠損値が存在しない区間まで。なお、HPF では予測モデルの決定に利用される期間を指定するのであり、最適な予測モデルが選択された後には、実績値を全て利用して予測を行なう。

HOLDOUTPCT= number サンプル期間の長さを全体の何パーセントにするか指定。(9.1)

INTERMITTENT= number 平均の欠損値間隔が intermittent オプションで指定した値より大きい時、間歇需要と判断して IDM/クロストン法を利用する。デフォルトは、1.25。BESTALL オプションと併せて使用する。

MEDIAN 中央値で予測を行なう。無指定ならば、平均で予測を行なう。

MODEL= モデル名

予測に際して、あらかじめ時系列モデルを指定することが可能。デフォルトは、BEST。

NONE...モデルを適用しない

DAMPTREND...ダンプトレンド指数平滑化

SIMPLE...指数平滑化

SEASONAL...季節調整平滑化

DOUBLE...2重(ブラウン)トレンド平滑化

WINTERS...乗法型ウィンターズ

LINEAR...線形(ホルト)指数平滑化

ADDWINTERS...加法型ウィンターズ

BEST...平滑化法全て (SIMPLE, DOUBLE, LINEAR, DAMPTREND, SEASONAL, WINTERS, ADDWINTERS)

BESTN...季節性のないモデル (SIMPLE, DOUBLE, LINEAR, DAMPTREND)

BESTS...季節性のあるモデル (SEASONAL, WINTERS, ADDWINTERS)

IDM|CROSTON...クロストン法。間歇需要(需要が断続的なケース)に用いる。(9.1 から IDM に呼称変更)

BESTALL...全てのモデルから選択(IDM|CROSTON, BEST)

NBACKCAST= n バックキャストを初期化するオブザベーション数を指定(9.1)

REPLACEBACK BACK=option で指定されたデータを、OUT=データセットで置き換える(9.1)

REPLACEMISSING... 実績値中の欠損値を、予測値で補完する

SELECT= option 適合度統計量を指定。デフォルトは、RMSE。

SSE, MSE, UMSE, RMSE, URMSE, MAPE, RSQUARE, ADJRSQ, AADJRSQ, RWRSQ, AIC, SBC, APC,

MAXERR, MINERR, MINPE, MAXPE, ME, MPE が利用可能。詳細はプロシジャ・レファレンスを参照のこと。

SETMISSING= option | number ID ステートメントの項を参照。

TRANSFORM= option データセットの変換方法を指定。MODEL=CROSTON の時は無効。

NONE...変換しない。デフォルト。

LOGISTIC... ロジスティック変換

LOG...対数変換

BOXCOX(n)... Box Cox 変換。-5 ≤ n ≤ 5

SQRT...平方根変換

AUTO... None または Log から自動で変換法を決定。

USE= option OUT=及び OUTSUM=データセットにおいて出力される予測値を特定する。

PREDICT...予測値を用いる。デフォルト。

UPPER...上限信頼限界を用いる。

LOWER...下限信頼限界を用いる。

ZEROMISS= option ID ステートメントの項を参照。

ID ステートメント

ID 変数 **INTERVAL=**interval option

ACCUMULATE= option タイムバケット毎にどのように累算するかを指定。

NONE...そのままの値を用いる。デフォルト。

N...欠損以外の数

TOTAL...合計

NMISS...欠損の数

AVERAGE | **AVG**...平均

NOBS...オブザベーション数

MINIMUM | **MIN**...最小値

FIRST...最初の値

MEDIAN | **MED**...中央値

LAST...最後の値

MAXIMUM | **MAX**...最大値

STDDEV | **STD**...標準偏差

ALIGN=option... 出力での SAS 日付の表記位置を指定。

BEGINNING|**BEG**|**B**(デフォルト), **MIDDLE**|**MID**|**M**, **ENDING**|**END**|**E** が利用可能。

END=option... データセット中の日付値の終了日を指定。

INTERVAL=option 日付値の間隔を指定。YEAR, SEMIYEAR, QTR, MONTH,

SEMIMONTH, TENDAY, WEEK, DAY, HOUR, MINUTE, SECOND が使用可能。

NOTSORTED 変数をソートして出力しない(9.1)

SETMISSING= option | number ...欠損値の補完方法を指定。

MISSING(欠損値として扱う), **AVERAGE**|**AVG**, **MINIMUM**|**MIN**, **MEDIAN**|**MED**,

MAXIMUM|**MAX**, **FIRST**, **LAST**, **PREVIOUS**|**PREV**, **NEXT** が使用可能。

START= option...データセット中の日付値の開始日を指定。

ZEROMISS= option 最初、あるいは最後の「0」の値の扱いを指定。NONE,LEFT,RIGHT,BOTH が指定可能。

IDM ステートメント(9.1)

IDM options;

9.1 では、間歇需要 Intermittent Demand Model の機能が大幅に強化された。詳細はプロシジャ・レファレンスを参照のこと。

INTERVAL=(smoothing-model-options)

SIZE=(smoothing-model-options)

AVERAGE=(smoothing-model-options)

BASE=AUTO | number

SAS High-Performance Forecasting の HPF プロシジャのシンタクスは、SAS® System 9 を元に記述されている。(9.1)と記述されているものは、SAS® 9.1 からの新機能。

ポスターセッション
調査・マーケティング

日本SASユーザー会 (SUGI-J)

地方における実演芸術鑑賞の実態

— 県民芸術劇場(兵庫県)の来場者調査より —

有馬 昌宏

神戸商科大学商経学部管理科学科

An Attempt to Grasp the Demand Structure of Performing Arts in Hyogo Prefecture

Masahiro Arima

Kobe University of Commerce

要 旨

兵庫県で実施されている、地方での実演芸術の鑑賞機会を提供するための県民芸術劇場への来場者を対象とするアンケート調査を平成 12 年度と 13 年度に実施し、この調査データを SAS および JMP を用いて分析し、レジャー白書や社会生活基本調査といった全国規模の無作為標本による調査の結果と比較対照することにより、地方における実演芸術のライブでの鑑賞活動の実態を明らかにしようと試みている。

キーワード： アンケート調査、JMP、SAS 8.2、TABULATE プロシジャ

1. はじめに

わが国経済社会が成熟し、交通基盤や情報通信基盤が整備・高度化されるとともに、我々の生活様式は大きく変化し、我々の活動内容は多様化して行動範囲も拡大してきている。この生活様式の変化と現状を総務庁統計局の「平成 13 年(2001 年)社会生活基本調査」に基づいて生活時間(15 歳以上の人々の週全体の平均時間)の観点から見てみると、1次活動(睡眠、食事など生理的に必要な活動)は 10 時間 34 分、2次活動(仕事、家事など社会生活を営む上で義務的な性格の強い活動)は 7 時間 00 分であるのに対し、自由時間である3次活動は 6 時間 26 分と 1 日の生活時間の 4 分の 1 を超えており、25 年前の昭和 51 年(1976 年)調査と比較すると3次活動に充てられる時間は 59 分も増加してきている。しかし、この余暇活動時間の変化や余暇活動内容の変化は、人々の年齢・性別・職業・居住地域などの違いを超えて一律に進行してきているものではない。例えば、3 次活動の中の積極的余暇活動を構成している「趣味・娯楽」の種目別にみた行動者率は、表 1 に示すように 47 都道府県の間でかなりの変動を示している。

本研究は、このような状況を踏まえ、兵庫県が推進している「こころ豊かな地域社会づくり」のための

表1 趣味・娯楽の都道府県別の行動者率(範囲と変動係数)

	最高		兵庫県	最低		変動係数 (%)
全ジャンル	埼玉県	89.3	86.3	77.2	青森県	3.9
スポーツ観覧	福岡県	24.9	19.3	13.1	徳島県	15.2
美術鑑賞	東京都	28.9	20.5	10.9	沖縄県	20.4
演芸・演劇・舞踊鑑賞	東京都	22.5	16.8	11.0	愛媛県	17.2
映画鑑賞	東京都	44.3	36.2	19.5	秋田県	17.6
クラシック音楽鑑賞	東京都	12.9	10.5	6.1	和歌山県	17.1
ポピュラー音楽鑑賞	東京・京都	15.6	13.5	9.8	青森県	13.8
楽器演奏	滋賀県	13.4	11.1	8.1	徳島県	10.7
邦楽	石川県	2.3	1.4	0.9	群馬県	21.5
カラオケ	埼玉県	44.8	39.1	28.8	青森県	10.4

平成13年社会生活基本調査報告より作成。

一つの基盤として、積極的余暇活動を構成する「趣味・娯楽」の中でも特に実演芸術の鑑賞関連活動に注目し、兵庫県が県内で展開している『兵庫県民芸術劇場』への来場者調査を実施し、社会生活基本調査やレジャー白書で知られる「余暇活動に関する調査」などの全国調査との比較を通じて地域社会における芸術・文化の現状と課題を把握し、今後の県や市町の芸術・文化行政立案と評価のための基礎資料を提供することを試みるものである。

2. 既存の芸術・文化統計の状況と本研究の概要

わが国の芸術に関する統計情報の整備状況は、芸術・文化に関連する活動を客観的に把握して計量的に分析することの必要性や要求が、経済学や社会学を中心とした社会科学分野での学問的関心からだけでなく、文化政策やアートマネジメント(芸術経営)などの実務的な観点からも高まってきているにもかかわらず、1976年から5年毎に実施されている総務庁統計局の「社会生活基本調査」、財団法人自由時間デザイン協会(旧名は財団法人余暇開発センター)による1976年から毎年実施されて「レジャー白書」として公表されている「余暇活動に関する調査」などが継続的に実施されているだけで、その整備は1970年代後半になって漸く着手されはじめた段階であるといえる。

しかも、上記の調査は、全国の約7.7万世帯、約20万人を対象とする「社会生活基本調査(2001年調査)」から全国の3千人を対象とする「余暇活動に関する調査」まで、いずれも無作為標本に基づく調査ではあるが、残念ながらマイクロの標本データは提供されておらず、報告書などで公表されているマクロの統計データから分析を行うことができるのみであった。なお、マイクロの標本データによる分析としては、1996年の社会生活基本調査のデータを用いて社会人の音楽鑑賞活動の有無を規定する要因と児童・生徒のクラシック音楽鑑賞活動を規定する要因の分析を行った有馬[3]、2000年の全国消費実態調査のデータを用いて実演芸術の鑑賞構造を分析しようとした有馬・周防[5]などがある。また、マイクロの標本データを用いた研究の可能性については、有馬[4]を参照されたい。

また、上記の各種全国調査は芸術活動のみに焦点を当てた調査ではないため、芸術の享受に関する詳細な分析を試みようとしても、ジャンル(種目)を細分化した各芸術分野での芸術の享受量(ある

いは需要量)や享受のスタイルについてまで踏み込んだ設問がされていないという問題点やサンプル数に伴う問題点があった。

例えば、「レジャー白書」として公表されている「余暇活動に関する調査」を利用して兵庫県での観劇、演芸鑑賞、音楽会・コンサートなどについての参加率の変化をみると、表2に示すように、母集団の大きさ(2000年の国勢調査による兵庫県の15歳以上人口は4,716,433人)に対してサンプル数が100前後と非常に小さいことの影響を受け、推定された兵庫県での参加率は大きく変動しており、この数字で兵庫県での鑑賞の実態を代表させることには無理がある。

したがって、芸術活動の細分化された各ジャンルの需要や芸術の享受者(需要者)の享受スタイルなども把握できる詳細な標本データを得て、さらに芸術の享受活動に影響を及ぼす個人属性などでブレイクダウンを重ねる細かな分析にも耐えうるだけの標本数を確保できる調査を限られた予算の範囲内で行おうとすれば、調査対象を特定の社会階層や特定の地域に限定した上で、標本抽出にも工夫を凝らした独自の調査を実施することが必要になる。

こうした観点から、兵庫県が県内で展開している『兵庫県民芸術劇場』事業の一般公演事業に着目し、『県民芸術劇場』を所管している兵庫県県民生活部芸術文化課ならびに財団法人兵庫県芸術文化協会の協力を得て、『兵庫県民芸術劇場』の一般公演入場者を対象とする来場者調査を実施し、消費支出の観点から芸術・文化関連活動の実態を全国的に把握できる「家計調査」と行動の有無や頻度の観点から芸術・文化関連活動の実態を全国のおよび都道府県別に把握できる「社会生活基本調査」の分析結果と比較対照しながら、兵庫県民の芸術・文化活動の特徴を解明していくことを目的とする本研究を企画した。

3. 兵庫県「県民芸術劇場」の来場者調査からみた実演芸術鑑賞の実態

3.1 調査の概要

兵庫県では、平成3年度(1991年度)より、県民に優れた舞台芸術を身近に鑑賞できるように、兵庫県と市町等がその経費を一部負担し、県内芸術団体等の協力を得て公立文化施設を会場とする「県民芸術劇場」事業を展開してきている。平成12年度は、県内37市町で延べ38回の「県民芸術劇場」が開催され、14,612人の県民が、また平成13年度は、県内32市町で延べ33回の「県民芸術劇場」が開催され、11,004人の県民が舞台芸術を鑑賞している。

我々は、限られた経費の中で兵庫県民の芸術鑑賞活動の実態を把握すべく、「県民芸術劇場」を所管している兵庫県県民政策部芸術文化課と財団法人兵庫県芸術文化協会の協力を得て、平成12年度は、平成12年8月6日に南淡町立文化会館で開催された「南淡町立文化体育館竣工記念ミュージカル」から平成13年3月11日に芦屋市のルナ・ホールで開催された「新世紀の響き ハイドン・オラトリオ」までの14の県民芸術劇場の公演について、平成13年度は我々のミスにより調査票の授受ができなかった2つの公演を除く30の公演について、来場者調査を実施した。有効回答者数は、平成12年度調査で4,499人、平成13年度調査で3,512人であった。

なお、調査票はA4サイズの用紙に両面印刷されたもので、①情報入手先(15の回答選択肢からの

表2 「余暇活動に関する調査(レジャー白書)」による芸術活動参加率の推移

ビデオの鑑賞(レンタルを含む)														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	30.8	36.9	44.4	42.9	41.6	44.3	43.2	42.2	40.8	41.0	45.4	45.7	42.6	43.4
東京	38.9	37.7	49.0	51.7	44.8	48.7	50.3	48.8	44.0	44.4	46.6	49.9	44.9	44.6
大阪	24.4	37.2	44.9	35.4	40.1	37.6	44.1	33.5	39.9	39.5	47.1	33.7	40.7	42.4
兵庫	38.4	36.5	44.6	35.1	42.3	39.3	47.1	44.6	42.9	50.9	47.0	54.5	40.0	37.6
観劇(テレビは除く)														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	12.6	12.5	12.1	11.9	12.6	12.8	12.8	11.7	12.2	12.7	11.5	10.5	12.6	11.7
東京	18.8	21.2	22.3	20.8	20.4	24.0	23.4	23.2	22.7	21.3	18.9	20.8	23.2	21.3
大阪	13.7	16.6	12.0	18.5	12.5	14.5	17.5	9.9	11.0	16.4	9.6	11.9	12.5	12.7
兵庫	16.2	13.6	12.2	10.3	12.7	15.3	12.4	10.1	15.5	20.3	14.9	14.1	12.6	12.8
演芸鑑賞(テレビは除く)														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	4.6	4.9	4.0	5.1	5.1	6.0	5.3	5.4	5.5	4.7	4.8	5.5	4.6	5.2
東京	4.3	4.6	6.5	6.2	5.6	6.9	6.1	9.2	8.6	7.9	5.9	13.9	6.7	8.1
大阪	2.6	7.0	3.9	4.3	5.6	7.5	6.0	6.5	5.5	2.4	6.7	4.7	6.5	7.8
兵庫	4.0	8.3	2.0	5.6	6.3	5.2	4.4	3.6	3.5	3.3	9.0	4.1	3.7	5.5
音楽会・コンサートなど														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	19.8	19.8	19.1	21.5	20.9	22.0	22.3	20.8	22.3	20.0	20.9	20.4	23.6	22.5
東京	25.0	24.5	26.7	29.3	24.4	31.4	33.5	26.4	31.0	28.1	25.2	30.6	34.4	29.1
大阪	11.8	18.4	14.2	21.8	15.2	15.4	18.3	15.2	19.1	15.7	14.6	11.7	18.5	15.6
兵庫	19.2	28.2	17.9	15.5	18.1	21.6	14.1	16.8	20.5	24.5	23.0	22.8	23.7	17.4
音楽鑑賞(CD・レコード・テープ・FMなど)														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	31.3	32.0	34.1	37.9	39.6	41.7	40.1	39.1	39.2	38.8	41.9	39.8	40.2	40.6
東京	42.6	35.6	43.6	45.9	43.4	48.2	48.6	48.8	46.8	45.8	46.6	48.1	43.9	47.6
大阪	21.8	30.4	28.6	33.2	34.8	38.0	39.7	31.5	32.7	33.9	36.4	28.1	37.5	28.3
兵庫	41.4	36.7	32.8	30.6	29.8	36.7	30.0	38.6	40.8	36.0	45.4	48.3	39.3	38.5
サンプル数														
	1987	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
全国	3,205	3,379	3,527	3,529	3,408	3,470	3,416	3,467	3,420	3,413	3,461	3,426	2,421	2,431
東京	352	326	337	403	349	384	377	377	361	367	323	338	314	296
大阪	271	285	288	264	266	266	266	234	259	245	235	233	216	205
兵庫	99	132	147	153	165	165	164	149	152	161	149	152	135	109

1) 昭和62年より、それまでの全国5万人以上都市に居住する15歳以上男女で3,000サンプルに加え、5万人未満都市及び郡部の1,000サンプルを加えた計4,000サンプルによる調査を実施してきたが、平成12年調査(平成12年12月実施)より、再び調査対象を以前の都市部3,000サンプルに戻している。

2) 参加率とは、ある余暇活動を1年間に1回以上おこなった人(回答者)の割合である。

出典:各年版の『レジャー白書』から作成した。

無制限複数回答)、②同行者(6の回答選択肢から単一回答)、③県民芸術劇場鑑賞の有無(2の回答選択肢から単一回答)、④過去を通算したジャンル別実演芸術鑑賞経験の有無(22の回答選択肢から無制限複数回答)、⑤過去1年に限定したジャンル別実演芸術鑑賞経験の有無(22の回答選択肢から無制限複数回答)、⑥稽古事や学校・地域での主体的文化・芸術活動のジャンル別経験の有無(19の回答選択肢から無制限複数回答)、⑦公演内容の評価(4の回答選択肢から単一回答および理由と感想の自由回答)、⑧入場料金の評価(3の回答選択肢から単一回答)、⑨今後の県民芸術劇場のジャンル別鑑賞希望の有無(16の回答選択肢から無制限複数回答)、⑩県民芸術劇場に対する意見・感想(自由回答)、⑪年齢(8の回答選択肢から単一回答)、⑫職業(13の回答選択肢から単一回答)、⑬性別(2の回答選択肢から単一回答)、⑭住所(市町村名を自由回答)、の14の質問から構成されている。

3.2 調査データの入力と解析

調査データは、有馬[1]と有馬[2]の方法に従って、マイクロソフト社の表計算ソフト Excel を利用して公演別に入力し、SAS によって入力エラーチェックを行うとともに無制限複数回答形式の質問に対するデータを各回答選択肢に対応する0-1型の2値データに変換し、公演別に作成されたSASデータファイルの結合を行った。その上で、JMP を利用して性別と職業、および職業と年齢のクロス集計を行い、矛盾する回答がないかのデータの論理的エラーチェックを行った。また、データの単純集計も JMP を利用して行った。ただし、調査票の④、⑤、⑥、⑨の無制限複数回答形式の質問については、これらの質問の間のクロス集計や、年齢や職業や性別との間のクロス集計を JMP を利用して行おうとすれば、例えば⑫の職業(回答選択肢の数は13)と⑤の過去1年に限定した実演芸術鑑賞経験ジャンル(回答選択肢の数は22)の間では13または22のモザイク図と分割表からクロス集計表を作成する手間があるように、かなり面倒な作業が必要となる。この手間は、無制限複数回答形式だけでなく、回答数が制限されている制限複数回答形式の質問についても同様である。そこで、複数回答形式の質問を対象とするクロス集計については、有馬[1]で紹介した TABULATE プロシジャを利用して一括してクロス集計表を作成するプログラムを作成し、SAS で実行することとした。なお、JMP では SAS で作成した SAS データセットがそのままの形式で読み込めるが、JMP で作成したデータセットを SAS で利用しようとするれば、1) JMP でデータセットを保存する際に通常の「JMP データテーブル」としてではなく「SAS 移送ファイル」として保存し、2) SAS のプログラムで DATA ステップでデータを読み込む前に COPY プロシジャで移送ファイルを変換する、という手順が必要なので注意されたい。この手続きは、SAS 社のウェブサイトのテクニカサポートのページ(<http://www.sas.com/offices/asiapacific/japan/service/technical/faq/list/body/pc021.html>)で紹介されている。

3.3 集計結果の分析

現在、公演実施市町から我々のもとに郵送されてきた平成12年度と平成13年度の回収された調査票のデータ入力作業は完了しており、全国を対象とした無作為調査である社会生活基本調査やレジ

表3 県民芸術劇場来場者調査の単純集計結果

性別

	2000年度	2001年度
1. 男性	26.9%	25.9%
2. 女性	67.8%	66.6%
3. 不明・無回答	5.4%	7.5%
サンプル数	4,499	3,512

年齢

	2000年度	2001年度
1. 10歳未満	1.8%	3.0%
2. 10歳代	12.4%	19.0%
3. 20歳代	6.1%	5.6%
4. 30歳代	11.7%	10.3%
5. 40歳代	19.1%	16.7%
6. 50歳代	21.3%	19.8%
7. 60歳代	17.7%	14.5%
8. 70歳代以上	6.8%	5.7%
9. 不明・無回答	3.2%	5.3%
サンプル数	4,499	3,512

仕事

	2000年度	2001年度
1. 会社員(常勤)	16.8%	14.6%
2. パートタイム	9.5%	7.1%
3. 自営業	7.0%	5.9%
4. 公務員	7.4%	7.0%
5. 教員	3.1%	2.7%
6. 専業主婦	21.3%	19.0%
7. 小学生	3.3%	4.7%
8. 中学生	8.1%	14.3%
9. 高校生	1.6%	1.7%
10. 大学・短大生	1.1%	1.1%
11. 専門学校生	0.1%	0.2%
12. 無職	10.6%	8.6%
13. その他	3.7%	3.8%
14. 不明・無回答	6.5%	9.3%
サンプル数	4,499	3,512

Q3 これまでに県民芸術劇場で公演の鑑賞をされたことがありますか。

	2000年度	2001年度
1. 今回がはじめて	48.9%	41.8%
2. これまでに鑑賞した	46.5%	54.0%
3. 不明・無回答	4.7%	4.2%
サンプル数	4,499	3,512

Q7 本日の催しのご感想はいかがでしたか。

	2000年度	2001年度
1. とてもよかった	54.4%	53.4%
2. よかった	26.6%	27.6%
3. ふつう	3.4%	5.0%
4. よくなかった	0.2%	0.6%
5. 不明・無回答	15.4%	13.4%
サンプル数	4,499	3,512

Q8 入場料金はいかがでしたか。

	2000年度	2001年度
1. 安い	21.2%	20.5%
2. ふつう	53.1%	53.4%
3. 高い	3.9%	5.9%
4. 不明・無回答	9.3%	20.3%
サンプル数	4,499	3,512

Q1. 本日の催しを何でお知りになりましたか。(複数回答)

	2000年度	2001年度
1. ポスター	17.1%	16.9%
2. ちらし	17.2%	18.0%
3. 新聞	5.6%	5.0%
4. 県広報	2.0%	2.3%
5. 市・町広報	26.1%	21.6%
6. 情報誌	3.4%	2.6%
7. 出演者から	5.7%	8.1%
8. 会場からの案内	10.6%	9.5%
9. 雑誌	0.2%	0.8%
10. 知人から	22.4%	21.0%
11. テレビ・ラジオ・CATV	1.0%	1.1%
12. 有線放送	1.4%	3.2%
13. 学校やPTAからの案内	11.3%	15.2%
14. 回覧板	2.1%	1.4%
15. その他	6.3%	5.6%
サンプル数	4,499	3,512

Q4. 本日の公演を除き、これまでに劇場・ホールでご覧になった演目があれば、該当する演目すべてに○をお付けください。

	2000年度	2001年度
1. オーケストラ	42.8%	42.5%
2. 器楽演奏	29.0%	30.1%
3. 室内楽	28.1%	27.3%
4. 声楽	22.1%	21.9%
5. 合唱	30.7%	32.9%
6. オペラ	12.9%	13.6%
7. バレエ	19.0%	19.5%
8. ミュージカル	29.5%	33.7%
9. 演劇	28.6%	29.1%
10. 歌舞伎	13.2%	13.1%
11. 文楽	6.6%	6.7%
12. 能・狂言	13.5%	15.7%
13. 邦楽	7.2%	7.9%
14. 邦舞	3.7%	4.1%
15. 民謡	7.3%	8.4%
16. ジャズ	18.4%	17.2%
17. ロック	7.5%	6.5%
18. ポップス	17.6%	17.7%
19. 歌謡曲・演歌	25.2%	25.0%
20. 落語・漫才	22.9%	24.4%
21. その他	4.4%	5.8%
22. 見たことがない	—	10.3%
サンプル数	4,499	3,512

Q6. 以下にあげる学校のクラブ・公民館・カルチャーセンターなどで行われている活動や個人で行うお稽古事について、参加したり習ったりしたことのあるものすべてに○をお付けください。

	2000年度	2001年度
1. ピアノ	22.9%	25.5%
2. 電子オルガン	5.0%	5.1%
3. ブラスバンド・オーケストラ	8.3%	7.2%
4. コーラス・声楽	18.2%	18.4%
5. 邦楽	3.7%	3.2%
6. 民謡	1.6%	2.2%
7. 邦舞・おどり	3.2%	3.5%
8. バレエ・モダンダンス	3.4%	4.2%
9. 社交ダンス	6.3%	5.2%
10. 華道	19.5%	18.6%
11. 茶道	16.7%	15.9%
12. 和歌・俳句	3.1%	3.8%
13. 詩・文芸	1.2%	1.5%
14. 絵画	8.1%	7.5%
15. 書道・習字	26.4%	25.6%
16. ロック・ジャズなどのバンド	2.4%	1.6%
17. カラオケ	4.3%	4.6%
18. その他	9.6%	8.3%
19. したことがない	—	24.3%
サンプル数	4,499	3,512

Q2. 今日はどなたとお越しになりましたか。

	2000年度	2001年度
1. 家族	48.2%	47.5%
2. 職場の人	3.8%	4.8%
3. 学校の人	8.6%	13.8%
4. 近所の人	13.1%	12.2%
5. 一人で	15.5%	12.9%
6. その他	11.2%	10.0%
サンプル数	4,499	3,512

Q5. 本日の公演を除き、過去1年間に限定して劇場・ホールでご覧になった演目があれば、該当する演目すべて○をお付けください。

	2000年度	2001年度
1. オーケストラ	22.1%	18.8%
2. 器楽演奏	15.1%	13.8%
3. 室内楽	14.0%	12.1%
4. 声楽	10.2%	9.3%
5. 合唱	16.2%	15.4%
6. オペラ	4.8%	4.0%
7. バレエ	6.0%	6.4%
8. ミュージカル	11.8%	15.1%
9. 演劇	10.6%	9.8%
10. 歌舞伎	5.0%	4.2%
11. 文楽	1.5%	1.3%
12. 能・狂言	5.2%	6.5%
13. 邦楽	2.6%	3.0%
14. 邦舞	1.3%	1.3%
15. 民謡	3.8%	3.4%
16. ジャズ	8.4%	7.9%
17. ロック	2.1%	2.1%
18. ポップス	6.5%	6.5%
19. 歌謡曲・演歌	10.3%	10.6%
20. 落語・漫才	7.8%	8.4%
21. その他	3.6%	3.1%
22. 見たことがない	—	17.3%
サンプル数	4,499	3,512

Q9. 今後、「県民芸術劇場」でご覧になりたい演目を下記ジャンルから3つお選びください。

	2000年度	2001年度
1. オーケストラ	35.2%	32.1%
2. 器楽演奏	15.9%	13.2%
3. 室内楽	16.2%	14.0%
4. 声楽	7.7%	7.2%
5. 合唱	9.4%	9.9%
6. オペラ	10.4%	10.6%
7. バレエ	14.0%	14.9%
8. ミュージカル	31.3%	33.3%
9. 演劇	17.7%	17.8%
10. 歌舞伎	11.2%	10.6%
11. 文楽	5.1%	4.0%
12. 能・狂言	11.5%	10.7%
13. 邦楽	5.1%	5.0%
14. 邦舞	2.1%	1.9%
15. ジャズ	20.3%	15.8%
16. その他	3.9%	3.4%
17. 特にない	—	5.6%
サンプル数	4,499	3,512

表4 県民芸術劇場来場者調査のクロス集計結果(過去1年の鑑賞の有無×性別と年齢)

平成12年度(2000年度)

	男	女	10歳未満	10歳代	20歳代	30歳代	40歳代	50歳代	60歳代	70歳以上	全体
1. オーケストラ	25.2	21.4	16.0	13.0	19.6	15.6	21.3	27.7	30.5	23.2	22.5
2. 器楽演奏	16.2	15.1	8.0	11.0	15.3	11.7	13.3	15.3	22.6	22.1	15.4
3. 室内楽	16.3	13.7	6.7	7.8	12.2	12.9	15.9	13.3	20.8	17.3	14.4
4. 声楽	10.8	10.3	4.0	3.5	4.7	5.3	8.2	12.6	19.0	17.7	10.4
5. 合唱	14.3	17.6	12.0	11.4	7.5	11.9	14.5	17.3	27.7	20.5	16.6
6. オペラ	4.8	4.9	1.3	1.9	4.7	3.0	3.9	4.5	9.1	8.3	4.8
7. バレエ	4.3	6.9	16.0	4.8	4.3	8.1	5.1	5.1	8.0	6.3	6.2
8. ミュージカル	9.0	13.2	18.7	6.5	9.4	16.4	13.3	10.7	13.6	11.4	12.0
9. 演劇	9.6	11.4	10.7	8.2	9.0	9.5	10.9	12.0	12.2	13.0	10.8
10. 歌舞伎	3.0	5.9	1.3	0.9	1.2	2.8	4.6	5.7	10.1	7.9	5.0
11. 文楽	1.3	1.6	0.0	0.4	0.8	0.4	0.5	1.7	2.9	5.5	1.5
12. 能・狂言	4.2	5.8	2.7	1.1	2.4	4.4	3.7	7.5	8.3	8.3	5.3
13. 邦楽	2.9	2.7	1.3	0.7	2.8	1.8	2.7	4.0	2.8	4.7	2.7
14. 邦舞	0.7	1.6	0.0	0.2	0.0	0.2	1.2	1.5	2.5	3.9	1.3
15. 民謡	3.9	3.6	1.3	0.9	0.8	0.6	1.8	5.0	6.5	13.0	3.7
16. ジャズ	10.4	7.5	5.3	4.7	8.2	8.1	10.4	9.8	8.2	6.3	8.4
17. ロック	2.7	1.9	2.7	2.4	7.8	3.4	2.1	1.3	0.7	0.0	2.1
18. ポップス	7.3	6.8	2.7	3.7	7.5	7.1	10.8	7.6	5.3	4.3	7.0
19. 歌謡曲・演歌	8.4	10.9	2.7	0.4	1.6	5.7	10.5	16.6	13.7	15.8	10.1
20. 落語・漫才	7.5	7.8	2.7	2.2	2.8	5.3	7.4	10.6	9.8	15.4	7.7
21. その他	3.8	3.6	10.7	6.5	5.5	4.7	3.4	2.1	2.6	1.2	3.7
22. 観たことがない	66.5	69.0	62.7	47.1	60.0	64.4	69.0	75.3	79.5	71.3	68.3
サンプル数	1,172	2,923	75	537	255	506	828	918	722	254	4,095

平成13年度(2001年度)

	男	女	10歳未満	10歳代	20歳代	30歳代	40歳代	50歳代	60歳代	70歳以上	全体
1. オーケストラ	23.0	17.0	11.9	10.9	13.0	13.7	21.7	22.0	27.8	23.2	18.7
2. 器楽演奏	13.2	14.5	6.9	9.5	9.3	13.1	16.3	15.9	18.1	19.0	14.1
3. 室内楽	13.5	11.8	9.9	3.9	9.3	11.7	12.7	16.2	18.1	19.0	12.3
4. 声楽	7.7	10.0	3.0	3.1	7.3	7.9	10.9	10.1	17.2	13.4	9.3
5. 合唱	14.6	15.9	4.0	10.1	11.4	10.8	15.4	17.2	26.0	24.7	15.5
6. オペラ	2.8	4.5	1.0	3.7	2.6	3.5	4.1	4.2	4.9	6.3	4.0
7. バレエ	4.2	7.5	14.9	4.8	3.1	13.7	6.3	5.7	5.1	4.9	6.5
8. ミュージカル	10.3	17.2	24.8	10.9	15.5	23.3	15.4	13.9	14.3	16.2	15.2
9. 演劇	7.0	10.9	14.9	5.9	7.3	12.0	10.6	10.8	12.1	7.0	9.8
10. 歌舞伎	2.5	5.3	1.0	0.3	1.6	1.5	3.4	5.6	11.0	16.2	4.5
11. 文楽	1.3	1.3	2.0	0.2	0.5	0.0	1.1	2.0	2.9	2.1	1.3
12. 能・狂言	4.7	6.9	2.0	2.0	2.6	3.8	5.6	8.2	12.8	12.7	6.3
13. 邦楽	1.9	3.7	0.0	0.9	3.6	3.5	3.6	4.3	3.7	5.6	3.2
14. 邦舞	0.6	1.7	0.0	0.6	0.0	0.3	1.1	1.7	2.6	5.6	1.4
15. 民謡	3.7	2.7	0.0	0.5	0.5	1.8	2.0	3.4	8.4	7.8	3.0
16. ジャズ	10.9	7.2	3.0	2.0	6.2	10.5	10.9	10.8	10.1	9.2	8.2
17. ロック	3.0	1.9	1.0	1.6	6.7	4.1	1.4	2.3	0.9	1.4	2.2
18. ポップス	5.7	7.1	3.0	2.3	9.3	7.9	10.7	8.2	6.6	0.7	6.7
19. 歌謡曲・演歌	8.1	11.3	2.0	0.8	2.6	7.0	10.7	16.4	19.4	20.4	10.4
20. 落語・漫才	9.1	8.3	3.0	2.6	3.6	8.5	11.6	10.5	11.5	14.8	8.5
21. その他	2.7	3.2	5.9	4.7	1.6	5.0	2.9	1.4	2.4	1.4	3.1
22. 観たことがない	24.3	15.3	22.8	35.9	23.3	19.5	12.9	11.9	6.4	4.9	17.9
サンプル数	882	2,200	101	644	193	343	559	647	454	142	3,083

ジャー白書での集計結果とも比較対照しながら分析作業を進めているところであるが、我々の調査は県民芸術劇場の来場者の中でも特に調査に自主的に協力してくれた人々を対象しているということで有意調査であり、無作為抽出の標本調査ではないために分析結果の解釈には交響楽団とオペラの聴衆調査を実施した Kurabayashi and Matsuda[7]で行われているような慎重な吟味が必要であり、分析に時間がかかっているのが現状である。参考までに、平成12年度(有効回答数 4,499)と平成13年度(有効回答数 3,512)の調査について、表3に各質問項目の単純集計結果を、表4に過去1年に限定した実演芸術鑑賞経験と性別および年齢との間のクロス集計結果を示しておくが、これらの表からも容易にわかるように、本調査の回答者は性別では女性に、年齢では40歳代から60歳代の年齢層に偏っており、過去1年間に限定した実演芸術の鑑賞経験率(社会生活基本調査の行動者率、レジャー白書の参加率に相当)は、無作為標本による社会生活基本調査での行動者率やレジャー白書での参加率を大きく上回っている。

4. 今後の課題

本研究では、「モノの豊かさ」から「ココロの豊かさ」へと人々の価値観が大きく転換して芸術・文化への意識が都市や地方を問わず高まる環境のもと、兵庫県の県民芸術劇場来場者調査の分析結果をもとに、レジャー白書や社会生活基本調査とも対照しながら、多様化した余暇活動の中で、ライブ鑑賞やメディア鑑賞を通じて、芸術がどのように需要されているのかの一端を明らかにしようと試みた。しかし、分析が不十分であり、今後は、今回の分析を基礎に、さらに深く分析をしていく必要があるといえる。なお、平成14年度と平成15年度も、兵庫県県民生活部芸術文化課と財団法人兵庫県芸術文化協会の協力を得て、平成12年度ならびに平成13年度に実施した県民芸術劇場来場者調査と同内容の調査を継続しており、将来は時系列的な分析も行っていく予定である。また、本研究は兵庫県内の公立文化施設で開催されている兵庫県民劇場の来場者に限った来場者調査(聴衆調査)を試みているが、同様の内容の来場者調査は、民営か公営かを問わず全国の劇場・ホールで実施されており、調査票の様式と設問を共通化してこのような調査を全国規模で実施できれば、劇場・ホールの来場者のプロフィールを体系的に知ることができ、今後の地方での芸術・文化関連施策の立案や実施に資することが大となると期待できる。

謝辞

県民芸術劇場への来場者調査にご協力いただいた来場者の皆さん、ならびに来場者調査を実施するにあたって面倒な作業を担って下さった地元主催団体の担当者の方々、また来場者調査の実施にご理解を示された兵庫県県民生活部芸術文化課と財団法人兵庫県芸術文化協会の皆様に、この場を借りて感謝の意を表させていただきます。なお、本研究は、平成12年度神戸商科大学特別調整研究費と平成13年度神戸商科大学学術研究会研究助成金の支援を受けて行った研究であり、共同研究者である神戸商科大学附属情報処理教育センター教授周防節雄先生ならびに同講師古隅弘樹先生から貴重なアドバイスをいただくとともに、神戸商科大学大学院経営学研究科経営情報科学専攻博

士後期課程の小田真樹子氏からはデータクリーニング用のプログラム開発などで助力を得たことに対しても、ここに感謝の意を表します。

参考文献

- [1]有馬昌宏、「パソコン版 SAS システムによる大規模統計調査データの解析—『現代青年の芸術意識と芸術活動』調査の分析」、第 11 回日本 SAS ユーザー会総会および研究発表会論文集、pp.297-314、1992。
- [2]有馬昌宏、「無制限複数回答形式のアンケート調査データの入力と処理方法」、第 20 回日本 SAS ユーザー会総会および研究発表会論文集、pp.277-284、1998。
- [3]有馬昌宏、「社会生活基本調査による余暇活動の分析」、財団法人統計情報研究開発センター、『平成 10 年度総務庁統計局委託研究報告 標本データの提供に関する研究報告書』、pp.25-30、1999。
- [4]有馬昌宏、「文化経済学における実証研究の動向と課題」、文化経済学(文化経済学会<日本>)、第3巻第1号、pp.11-16、2002。
- [5]有馬昌宏・周防節雄、『消費実態から見た芸術・文化の需要構造に関する基礎的研究』、マイクロ統計データ活用研究会平成 13 年度分研究成果報告会、2002。
- [6]有馬昌宏、「地方における実演芸術の需要の実態—家計調査・社会生活基本調査・県民芸術劇場来場者調査から—」、商大論集(神戸商科大学)、第 54 巻第6号、pp.99-152、2003。
- [7]Kurabayashi Y. and Y. Matsuda, *Economic and Social Aspects of the Performing Arts in Japan: Symphony Orchestras and Opera*, Kinokuniya Co. Ltd., 1988.

青年期女性の自意識と完全主義傾向の関連

中村晃士* 牛島定信* 縣俊彦** 清水英佑**

* 東京慈恵会医科大学精神医学講座

** 東京慈恵会医科大学環境保健医学講座

The relationship between self-conscious and perfectionism of adolescent female
Koji Nakamura* Sadanobu Ushijima* Toshihiko Agata** Hidesuke Shimizu**

*Department of Psychiatry, Jikei University School of Medicine

**Department of Public Health and Environmental Medicine, Jikei University School of Medicine

要 旨

女子大学生に対し、MPS (Multidimensional Perfectionism Scale), 公的自意識尺度, 私的自意識尺度の3つの自己記入式質問紙を施行し, 自意識と完全主義傾向の関連について検討した. 公的自意識には, MPS の下位項目「ミスへの過度のとりわれ」「親の期待」が取り込まれ, 私的自意識には, 「自身の行動への疑い」が $p < 0.01$ で取り込まれ, それぞれの質的な違いが明らかとなった.

キーワード: 青年期女性, 完全主義傾向, 公的自意識, 私的自意識, 重回帰分析

1. はじめに

青年期には, 自我の発達上においても自我同一性の獲得の問題, 社会への適応など様々な問題を抱えており, 現代社会においては, そういった問題を解決する時期が延長し, 青年期自体も長い期間として捉えられるようになってきている. そして, 社会への不適応を起こす中で, 個人の性格の問題が浮き彫りとなるケースが少なくない. その不適応を起こしやすい性格の中で, 最近「完全主義」が注目されている. 自らの完全主義という性格から, 自分がミスをするを許せず, 自分に常に高い目標を設定して, あがき続け, 結果として不適応を起こすと考えられる.

また, 最近青年期女性において対人恐怖症例の増加が指摘されており, 一般的にも対人緊張を持つ人たちが増えている印象がある. そこには自意識との関連があり, 中でも公的自意識が高いと対人緊張を持ちやすいとされている.

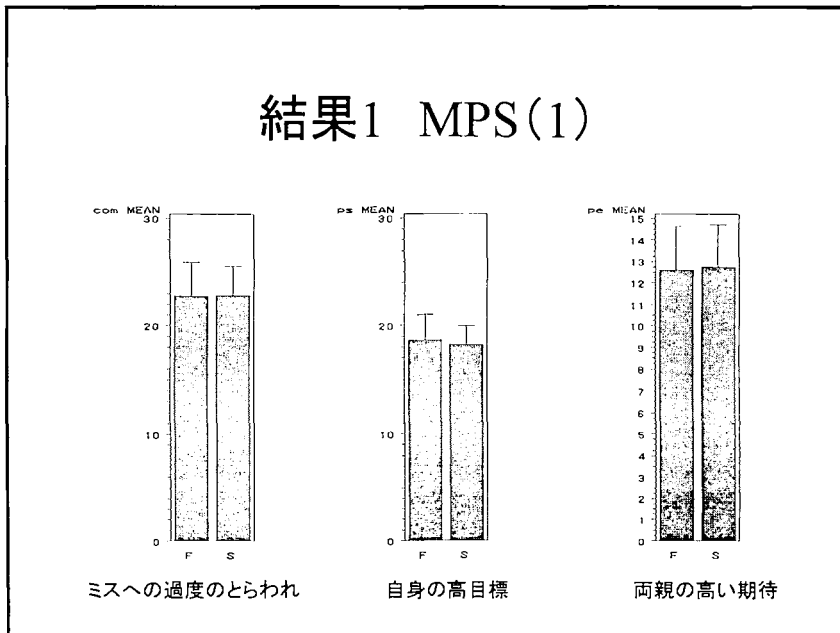
そこで今回我々は, 青年期女性の完全主義傾向, 公的自意識, 私的自意識のについて, その関連と合わせて調査し, 検討した.

2. 対象と方法

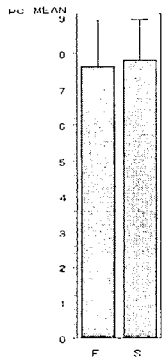
女子大学生 78 名に対し、MPS (Multidimensional Perfectionism Scale), 公的自意識尺度, 私的自意識尺度の 3 つの自己記入式質問紙を施行した。MPS は Frost ら(1990)が作成した完全主義傾向を多次的に把握することを目的とした評価尺度で、35 項目の質問からなり、下位項目は「ミスへの過度のとりわれ」「自身の高目標」「親からの高い期待」「親からの批判」「自身の行動への疑い」「整理整頓好き」の 6 項目から構成されている。各質問項目に対しては、「強く同意する」から「全く同意できない」の 5 段階リカレントスケールが用いられている。その邦訳版の MPS は田中ら(1999)が作成し、その信頼性と妥当性 (基準関連) が証明されている。また MPS はすでに国内外でも摂食障害患者を対象に用いられ、その有用性が指摘されている。また、自意識尺度日本語版は Fenigstein, Scheier, & Buss(1975) が作成したものをもとに、菅原 (1984) によって作成されたもので、公的自意識尺度 11 項目、私的自意識尺度 10 項目からなり、これも 7 段階のリカレントスケールが用いられている。背景因子として、年齢、長子か否かについても調査した。得られた結果は SAS を用いて統計学的検討を行った。

3. 結果

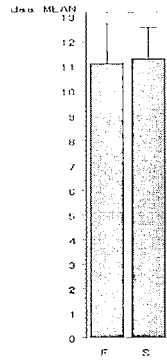
データの記入ミスがあったものを除いた有効回答数 (有効回答率) は、73 名 (93.6%) で、平均年齢は 21.2 ± 4.1 (Mean \pm SD) 歳であった。まず長子か否かで MPS および自意識尺度の結果を比較したのが、結果 1 から結果 3 までのグラフである。F が長子の群を表し、S が長子以外の群を表している。完全主義傾向の下位項目および公的自意識、私的自意識の長子か否かで、優位な差は認められなかった。



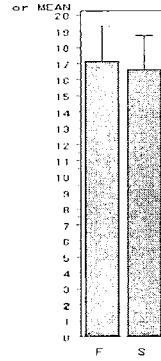
結果2 MPS(2)



両親からの批判

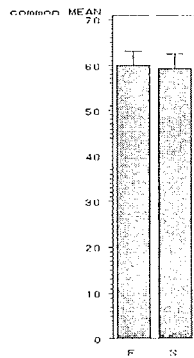


自身の行動への疑い

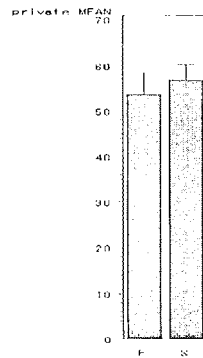


整理整頓好き

結果3 自意識



公的自意識



私的自意識

次に長子か否かの区別無く、全体として自意識と完全主義傾向の関連について調べた。公的自意識尺度と私的自意識尺度を、完全主義傾向の下位6項目でステップワイズ法により重回帰分析を行った。この際の取り込み基準は、0.15とした。結果は公的自意識は表4、私的自意識は表5に示した通りであった。

表4 公的自意識のステップワイズ法による重回帰分析の結果

Step	Variable Entered	Parameter Estimate	Standard Error	Partial R-Square	Model R-Square	Pr > F
1	COM	0.68853	0.13273	0.2462	0.2462	0.0001
2	PE	-0.53009	0.19645	0.0910	0.3373	0.009

表5 私的自意識のステップワイズ法による重回帰分析の結果

Step	Variable Entered	Parameter Estimate	Standard Error	Partial R-Square	Model R-Square	Pr > F
1	DA	0.98770	0.35076	0.1219	0.1219	0.0001

公的自意識には、MPSの下位項目「ミスへの過度のとりわれ (COM)」「親の期待 (PE)」が $p < 0.01$ で取り込まれた。私的自意識には、「自身の行動への疑い (DA)」が $p < 0.01$ で取り込まれた。

4. 考察

以前より、対人恐怖患者などは、「厳しくしつけられた完全主義傾向が高い長子に多い」といった報告があったが、今回の調査では、完全主義傾向、自意識ともに長子と長子以外に優位な差は認められなかった。このことから、昔の家長制度といった規範が現代社会においては崩れてきているのではないかと推察された。

また公的自意識には「ミスへの過度のとりわれ」、「親の期待」の2項目が取り入れられたことから、公的自意識は、ミスへの過度のとりわれ、親の期待など、周囲に対する敏感さが関連していることが分かった。また私的自意識には、「自身の行動への疑い」が取り込まれたことから、内省的な自己への敏感さが私的自意識には関連していることが分かった。これは質問紙の内容から当然の結果とも言え、今後さらなる検討が必要と考えられた。

5. まとめ

完全主義傾向、自意識は、長子か否かでは差がみられなかった
公的自意識と私的自意識は、完全主義傾向と関連があり、さらには質的な違いが明らかとなった。

日本SASユーザー会（SUGI-J）

個人レベルの選好を基にしたクラスタリング

河崎 一益 * 松沢 利繁 **

株式会社日本アルトマーク 統計解析部 *

(株)インターナショナル・クリエイティブ・マーケティング プロダクショングループ **

Clustering based on selection of an individual level

Kazumasu kawasaki * Toshishige Matsuzawa **

Statistical Analysis Division, Ultmarc Inc. *

Production Group, International Creative Marketing **

要 旨

クラスタ分析は、マーケット・セグメンテーションのために広く行われてきた。多くの場合消費者の意識（製品に対する態度やライフスタイルを表わす意見項目など）をクラスタ化することが多かった。今回は、消費者の商品に対する選好を基にクラスタ分析を行うと同時に、従来のクラスタ分析との比較を行い、その有効性を探ることを目的とした。

キーワード： クラスタ分析、選好、補償型モデル、SAS/STAT

1. 問題の背景

クラスタ分析は、1970年代から1980年代にマーケット・セグメンテーションのために多用された。この時代のクラスタ分析は、ライフスタイル分析を基本においており、特に、VALS^(注)は、マズローの欲求5段階説をベースに消費行動に関する800項目の質問を組み合わせ、9つのセグメントに分けたものであった。しかし、消費社会の高度化、複雑化に伴い、ライフスタイル分析を基礎においたクラスタ分析でのマーケット・セグメンテーションは、「マーケティング活動にどのように活用するのか」といった具体的場面の有効な活動におけるサポートの難しさから徐々にその意義が薄れてきた。

今回、我々は「消費者が商品をどのように選択するのか」を基にクラスタ分析を行うことによって、購入のパターン別のセグメントを発見することを試みた。

^(注)Value And Lifestyle 購買決定プロセスにおける消費者の行動を予測するために心理学的理論と社会学的理論に従って消費者を類型化したスタンフォード研究所のライフスタイル分析

2. 調査の設計

分析に使用するデータを得るためにアンケート調査を行い、株式会社日本アルトマークの社員が分担して収集した。回収数は213票であったが、内容に不明・未記載等の不備のない187票(男:96票、女:91票)を有効票として分析を行った。

表1. 今回調査した商品

デジタルカメラ	缶入りお茶	プラズマテレビ	カップラーメン
パナソニック DMC-F1-S	伊藤園 お〜いお茶	シャープ PZ-43BD3	日清 カップヌードル
オリンパス ミュー10 DIGITAL	コココーラ まる茶	日立 W50-PDH3000	明星 ラーメン職人
キャノン PowerShot G3	ヤクルト お茶	ソニー KE-32TS2	エースコック わかめラーメン
ミノルタ DIMAGE F300S	キリン 生茶	パナソニック TH-42PX10	日清 ラ王

3. 分析の流れ

今回は各対象者に対して、デジタルカメラ・缶入りお茶・プラズマテレビ・カップラーメンの各4商品(計16商品)を提示して各商品の属性評価(デザイン・使いやすさ・機能(味)・メーカー(ブランド)・価格)と総合的な購入意向を聞いた。

消費者の選好モデルは補償型モデルと非補償型モデルに大別されるが、今回は補償型モデルを想定して、総合的な購入意向を従属変数、各属性評価を説明変数として個人別に重回帰分析を行い、各属性変数の偏回帰係数を算出した。R-squareの平均は0.7567とかなり高い値となった。

個人別に得られた偏回帰係数が各個人の選好を表すものと考え、算出した偏回帰係数を用いてクラスター分析を実施した。

表2. クラスター分析の結果

	全体	クラスター 1	クラスター 2	クラスター 3	クラスター 4	クラスター 5
サンプルサイズ	187	23	37	43	53	31
デザイン	0.1822	0.2702	0.1252	0.1795	0.1715	0.2067
使いやすさ	0.1712	0.1438	0.2191	0.1311	0.2241	0.0994
機能	0.2739	0.2910	0.3131	0.2253	0.2645	0.2981
価格	0.3196	0.2669	0.4077	0.3386	0.2945	0.2701
メーカー	0.2979	0.3048	0.3056	0.2616	0.3066	0.3194

クラスター分析の結果から、以下のように各クラスターの性格付けを行った。

表 3. クラスター別特徴

クラスター	特 徴
クラスター 1	デザインを特に重視するグループ
クラスター 2	価格コンシャスであり、機能も重視するグループ
クラスター 3	やや価格を重視するグループ
クラスター 4	使いやすさを重視するグループ
クラスター 5	メーカー、使いやすさを重視するグループ

各クラスターがうまく分かれているかどうかをみるために、16 商品の購入意向に対して、クラスターを要因として分散分析を実施した。その結果、16 商品中 7 商品が $p<0.05$ で有意となった。

表 4. 選好クラスターを要因とした場合の購入意向の有意性

有意となった商品	有意とならなかった商品
パナソニック DMC-F1-S	オリンパス ミュー10 DIGITAL
キャノン PowerShot G3	ミノルタ DIMAGE F300S
伊藤園 お〜いお茶	コカコーラ まろ茶
キリン 生茶	ヤクルト お茶
シャープ PZ-43BD	日立 W50-PDH3000
日清 カップヌードル	ソニー KE-32TS2
明星 ラーメン職人	パナソニック TH-42PX10
	エースコック わかめラーメン
	日清 ラ王

上記結果をみると、定番(よく知られた)商品でクラスター間に購入意向の差がみられる傾向にある。

同様の分析を意識項目についても実施した。25 項目のうち以下の 6 項目が $p<0.05$ で有意となった。こだわりと情報探索項目で各クラスター間での差がみられる。

表 5. 有意となった意識項目

どんなことにも関心を持ち、何でも自分で試してみたい
新しいファッションや流行を人より早く取り入れる方だ
特定の商品(時計など)にこだわりがある
インターネットで情報をよく検索する
おいしいものを求めてあちこち食べ歩いている
衝動買いをすることが多い

4. 従来型クラスターとの比較

従来型クラスターとの比較のため、25個の意識項目を5段階で回答してもらい、それを基に8つの因子を抽出した。説明率は64.5%であった。

抽出した因子を用いてクラスター分析を行い以下の5つのクラスターを得た。

表6. クラスター別因子得点

	クラスター1	クラスター2	クラスター3	クラスター4	クラスター5
サンプルサイズ	55	33	23	34	42
社交因子	0.3092	-0.4544	0.2924	-0.8685	0.4950
ファッションセンス因子	-0.0431	0.0639	-0.5163	0.0291	0.2654
買い物楽しみ因子	-0.0792	-1.0945	0.8048	0.0308	0.4980
本物志向因子	-0.4638	0.1922	-0.7846	-0.1258	0.9878
インターネット因子	0.3067	-0.4943	-0.1217	-0.1024	0.1364
情報探索因子	0.1115	-1.0532	-0.1114	1.0799	-0.1317
計画購買因子	0.2520	-0.1061	-0.9938	0.3810	-0.0108
機能重視因子	-0.9289	0.2964	1.0572	0.5166	-0.0135

この結果から、各クラスターに以下のようなネーミングを行った。

表7. クラスター別特徴

クラスター 1	ものにこだわらないグループ
クラスター 2	買い物を楽しみと感じていない・情報非探索グループ
クラスター 3	機能を重視する・買い物エンジョイグループ
クラスター 4	情報を個人で探索するグループ
クラスター 5	本物を志向する・買い物を楽しみとするグループ

消費者の選好でクラスター分析を行ったときと同様に、16商品の購入意向について分散分析を実施した。この結果、16商品中6商品が $p < 0.05$ で有意となった。消費者選好から作成したクラスターの場合と大きな違いはないが、有意差のみられた商品群がお茶・カップラーメンに偏っている。

表8. 従来型クラスターを要因とした場合の購入意向の有意性

有意となった商品	有意とならなかった商品	
パナソニック DMC-F1-S	オリンパス ミュー10 DIGITAL	日立 W50-PDH3000
コココーラ まろ茶	キャノン PowerShot G3	ソニー KE-32TS2
麒麟 生茶	ミノルタ DIMAGE F300S	パナソニック TH-42PX10
日清 カップヌードル	伊藤園 お〜いお茶	エースコック わかめラーメン
明星 ラーメン職人	ヤクルト お茶	
日清 ラ王	シャープ PZ-43BD3	

また、25個の意識項目についてはすべての項目において $p < 0.05$ で有意差があった。25個の意識項目をベースに因子分析を行い、それをもとにクラスター分析を実施したわけであるから当然の結果といえよう。

次に、佐々木(1984)の行った REC スケールを用いて、選好に基づいて作成したクラスターと従来型のクラスターとの比較を行った。

REC スケール(Rationality and Emotionality of Consumer)は以下の12の項目について、「そう思う」から「そう思わない」までの5段階での回答を求めるものである。12項目のうち、①、③、⑤、⑧、⑩、⑫が合理性に関する項目であり、②、④、⑥、⑦、⑨、⑪が情緒性に関する項目である。「そう思う」に5点を与え、以下4点3点2点、「そう思わない」を1点として、合理性、情緒性の各項目を合計したものである。

表 9. REC スケールの項目(佐々木, 1984より作成)

買い物時にはよくバーゲンを利用する	(合理性)
流行のものを良く買う	(情緒性)
どの店で買えば得かに行く前に良く調べてみる	(合理性)
そのもののムードや情緒を特に重視して買う	(情緒性)
買う物は必要最低限にとどめておく	(合理性)
買う時には店員がすすめるものにする	(情緒性)
買う時にはよく広告をしているブランドを買う	(情緒性)
実用性とか使いやすさを特に重視して買う	(合理性)
見た感じとか美しさを特に重視して買う	(情緒性)
できるだけ多くのおものを比較したうえで買う物を決める	(合理性)
新しいものが出た時は人より早く買う	(情緒性)
とにかく安くて経済的な物を買う	(合理性)

各クラスター別の平均値は以下ようになった。

表 10. クラスター別 REC スコア

	サンプル サイズ	合理性		情緒性		
		平均値	標準偏差	平均値	標準偏差	
全 体	187	19.567	4.019	16.529	3.649	
選好を基にした クラスター	クラスター 1	23	18.870	5.137	17.957	3.509
	クラスター 2	37	19.541	4.161	15.811	3.526
	クラスター 3	43	19.326	3.920	17.791	3.596
	クラスター 4	53	19.434	3.495	16.151	3.427
	クラスター 5	31	20.677	3.945	15.226	3.694
従来型の クラスター	クラスター 1	55	20.073	2.930	17.836	3.149
	クラスター 2	33	16.455	3.251	14.758	3.545
	クラスター 3	23	20.087	5.080	15.957	3.735
	クラスター 4	34	21.588	4.164	14.676	3.470
	クラスター 5	42	19.429	3.769	18.024	3.197

これを見ると、選好を基にしたクラスターは合理性では差がみられないが、情緒性では差がみられる。従来型のクラスターでは、合理性・情緒性とも差がみられる。分散分析を実施すると、選好を基にしたクラスターでは $p < 0.05$ で情緒性に有意差がみられるが、合理性では有意差がみられない。従来型のクラスターでは合理性・情緒性ともに $p < 0.05$ で有意差がみられた。

5. 今後の課題

①調査商品の選定

調査商品をどのように選定するかがクラスター分析に大きな影響を与える。特に多数の人がよく知っており、ある程度価格について知識のあることが条件となる。今回は多少欲張りすぎて商品カテゴリーを広く取ってしまったが、対象者を設定するには対象者がある程度価格感度を持った商品を選択することが重要である。

②消費者の商品カテゴリーに対する関与

今回は消費者の各商品カテゴリーに対する関与の問題を考慮することなくモデルを設定したが、消費者の関与の程度によって選好モデルが異なることが予想される。したがって消費者の商品カテゴリーに対する関与度の測定方法の確立と、関与度を考慮したモデルの構築を考えていく必要があると思われる。

③モデルの精緻化

今回は、消費者の選好が補償型モデルで行われることを前提に進めてきたが、商品カテゴリーや消費者のタイプによっては非補償型ルールが採用されるケースもある。したがって、消費者の選好ルールを把握して、それをモデルに生かしていくことが重要である。特に消費者の選好ルールは商品のカテゴリーや関与度といった様々な要因が絡まってくる。これらの要因を考慮したうえでモデルの精緻化を行っていきたいと思う。

<参考文献>

- (1) 朝野熙彦(2000)「マーケティング・リサーチ工学」講談社
- (2) 片平秀貴(1987)「マーケティング・サイエンス」東京大学出版会
- (3) 片平秀貴(1991)「新しい消費者分析－LOGMAPの理論と応用－」東京大学出版会
- (4) 杉浦徹雄編著(1997)「消費者理解のための心理学」福村出版
- (5) 竹村和久編(2000)「消費者行動の社会心理学」北大路書房
- (6) 中西正雄編著(1998)「消費者選択行動のニュー・ディレクションズ」関西学院大学出版会
- (7) ピルヨ・ラークソネン(池尾・青木監訳)(1998)「消費者関与」千倉書房
- (8) Vithala R.Rao Joel H Steckel (1997)「Analysis for Strategic Marketing」

ADDISON-WESEY

日本SASユーザー会 (SUGI-J)

患者参加型医療情報交換システムのニーズ調査

○義澤 宣明* 船曳 淳* 小山博史**

*株式会社三菱総合研究所 安全科学研究本部

**東京大学大学院 医学系研究科 クリニカルバイオインフォマティクス研究ユニット

Investigation of Needs for a Patient-Oriented Interactive Health Communication System

Nobuaki Yoshizawa* Jun Funabiki* Hiroshi Oyama**

**Safety Science Research Division, Mitsubishi Research Institute, Inc.

Department of Clinical Bioinformatics, Graduate School of Medicine, The University of Tokyo

要 旨

患者参加型の医療情報交換システムに求められる機能を調べるために、1,684 名を対象とした医療情報交換に関するニーズ・意識調査を実施した。調査はインターネットアンケートを利用して実施した。調査内容は、①医療者からの情報開示や説明等に関する満足度や不満原因に関する基礎的調査、②病院内外を結ぶネットワークを利用する上でのニーズ調査、③高度医療ネットワーク及び先進的医療に関する意識調査、の3カテゴリから構成した。得られた回答について、SAS 及び JMP を用いて統計解析を実施した。本調査結果は、今後の医療情報交換システムの方向性について示唆にとむものであった。

キーワード： BaseSAS, SAS/STAT, JMP, インターネットアンケート, 医療情報システム

1. はじめに

インターネットを利用したインタラクティブ・ヘルス・コミュニケーション (Interactive Health Communication) は、今後急速に普及することが予想される。インターネット利用者数は平成 13 年末時点で 5,593 万人と推計されており、国民のほぼ半数を占めている[1]。特に、インターネットの世帯普及率は、平成 12 年の 34.0%から平成 13 年の 60.5%と 2 倍近くに急増している[1]。このような状況から、今後は患者参加型の IHC が医療情報システムにおいても重要な位置をしめるようになると思われる。これまでも、医療機関のホームページを中心に IHC に関する研究が報告されてきた[2,3,4]。本研究では約 1,600 名を対象としたインターネットアンケートによるニーズ調査の結果を報告する。調査内容は、①医療者からの情報開示や説明等に関する満足度や不満原因に関する基礎的調査、②病院内外を結ぶネットワークを利用する上でのニーズ調査、③高度医療ネットワーク及び先進的医療に関する意識調査、の3カテゴリで構成した。本研究では、①カテゴリを中心に調査結果の概要を報告する。

2. 調査方法と回答者属性

2.1 調査方法

goo リサーチのインターネットアンケートを利用した[5]。goo リサーチでは、あらかじめ登録されたモニターにアンケート参加を依頼し、web 画面で回答が入力される。郵送式のアンケートとは異なり、回答数が 1,684 件に達した時点で調査を打ち切った。

2.2 回答者属性

回答者の性別は、男性(50.1%)、女性(49.9%)であり、性別については偏りが無かった。年齢は、30～34 歳、35～39 歳、40～44 歳が、それぞれ約 20%程度で全体の 60%ほどであった。これに、10%程度である、25～29 歳、45～49 歳、55～59 歳を加えると回答者全体の 90%となる。図1に年齢・性別の分布を示す。年齢が上がるにつれて、男性の割合が多くなる。

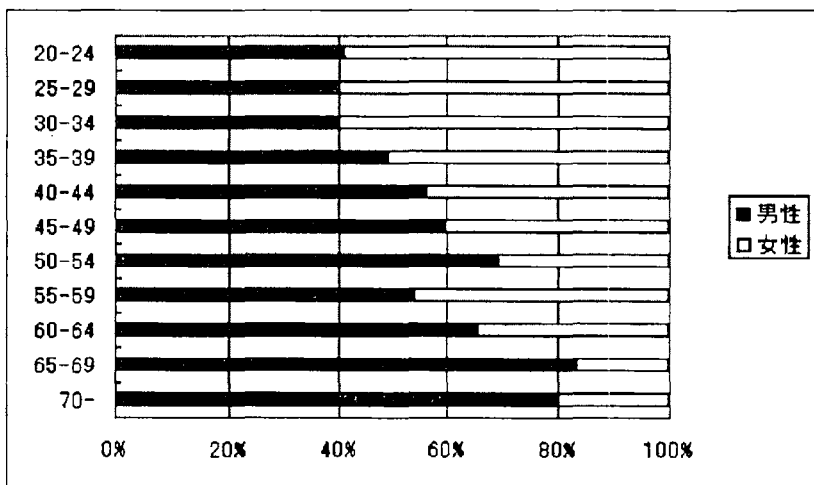


図1 回答者の年齢・性別分布 (N=1,684)

最も多かった回答者は「専業主婦」(20.9%)で、それについて「給与所得者(技術・専門職)」(17.4%)が多かった。居住地は、首都圏(埼玉, 千葉, 東京, 神奈川: 45.1%)、関西(京都, 大阪, 兵庫: 17.2%)で全体の 6 割以上となった。

3. 単純集計結果

3.1 診療経験

複数回答の結果は、短期通院(42.8%)が最も多かった。ここで、あらためて以下の方法で分

類しなおした。

- ・「入院」または「長期通院」経験者は、「入院・長期通院」に分類
- ・上記以外の「短期通院」経験者のみを、「短期通院」に分類
- ・「入院及び通院ともない」は、そのまま「なし」に分類

上記の結果、回答者の診療経験は以下の3つに分類された。

「入院・長期通院」	842名(50.0%)
「短期通院」	498名(29.6%)
「なし」	344名(20.4%)

医療情報に関する具体的な質問については「入院・長期通院」の842名のみが回答している。

3.2 診療記録等の閲覧希望

日本看護協会の調査[6]と比べると、本調査では“自分に関する記録の“閲覧希望が強かった(表1)。また、“十分説明されれば見なくてもよいと思う。”が日本看護協会の調査の半分ほどであった。この違いは、実際に入院している患者への調査とインターネットによる一般への調査の違いと考えられる。入院患者は相対的には医療スタッフからの説明に満足していることが示唆されている。今回の調査では、「長期通院」に比べて「入院」を経験した回答者が「自分に関する記録は全て見たいと思う。」割合が高かった。

表1 診療記録の閲覧希望(N=842)

	入院+ 長期通院	入院	長期通院 のみ	日本看護 協会調査[6]
自分に関する記録はすべて見たいと思う。	51.9%	54.3%	48.6%	33.6%
自分が見たいと思うものだけ見られればよいと思う。	26.1%	26.2%	26.0%	18.3%
十分説明されれば見なくてもよいと思う。	21.0%	18.4%	24.6%	41.3%
見なくてもよいと思う。	1.0%	1.0%	0.8%	4.3%
無回答・不明	—	—	—	2.4%

3.3 医師からの説明

医師からの説明について満足度を調査した。前述の診療記録の閲覧希望と満足度の関係を図2に示す。“非常に不満だった”回答者は31名と少ないが、医師の説明に対する不満が大きいほど、カルテ等の診療記録の開示要求が強くなっている。

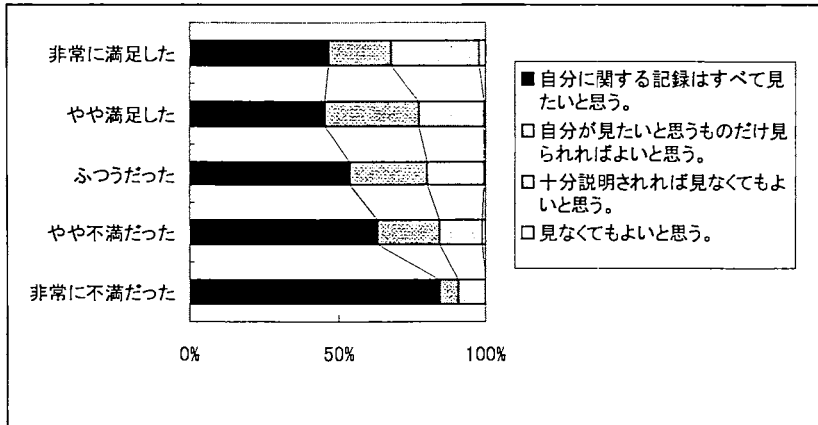


図 2 診療記録の閲覧希望と医師への説明の満足度の関係 (N=842)

3. 4 医師からの説明への不満

医師からの説明への不満の理由は図 3 に示すようになった。

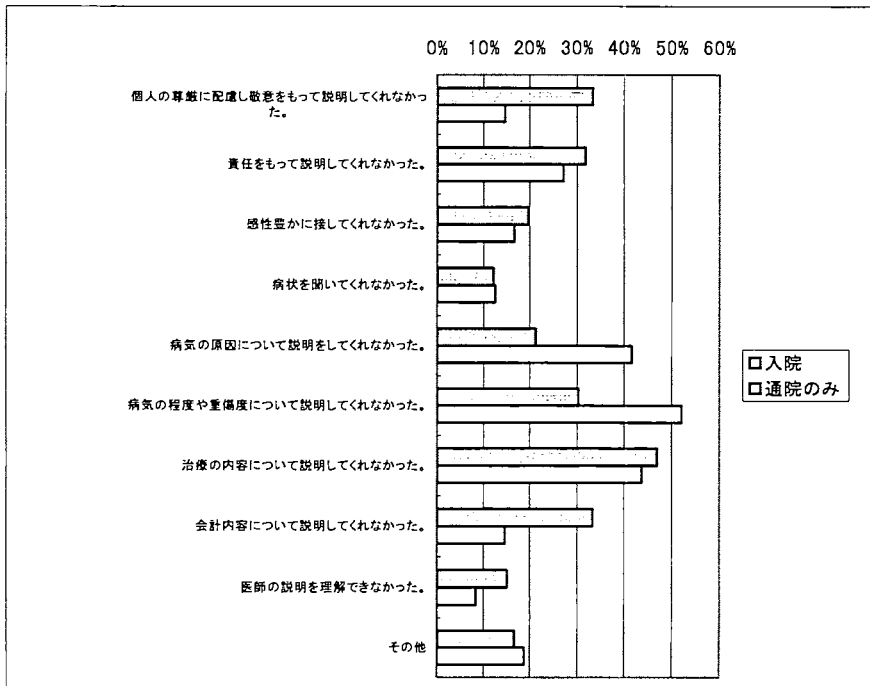


図 3 医師からの説明への不満の理由 (N=114)

「入院」と「長期通院のみ」で回答内容に違いが認められた。「入院」経験者では、「個人の尊厳に配慮し敬意をもって説明してくれなかった。」が「長期通院」の 2 倍以上であった。また、「長期通院のみ」の場合、「病気の原因」, 「病気の程度」に関する説明不足への不満が、「入院」経験者を上

回った。“治療の内容”説明の不足は、「入院」経験者及び「長期通院」の両方の半数程度が不満をもっていた。“会計内容”の説明については「入院」経験者の不満が、「長期通院」の2倍ほどとなっている。

3.5 医師・看護婦・その他の病院のスタッフに対して望むこと

「医師・看護婦・その他の病院のスタッフに対して望むこと」の各質問の回答について次のような得点付けを行い分析した。

“強く望む”=4点，“ある程度望む”=3，“どちらでもない”=2，

“あまり望まない”=1，“望まない”=0

各問の得点平均を以下に示す。

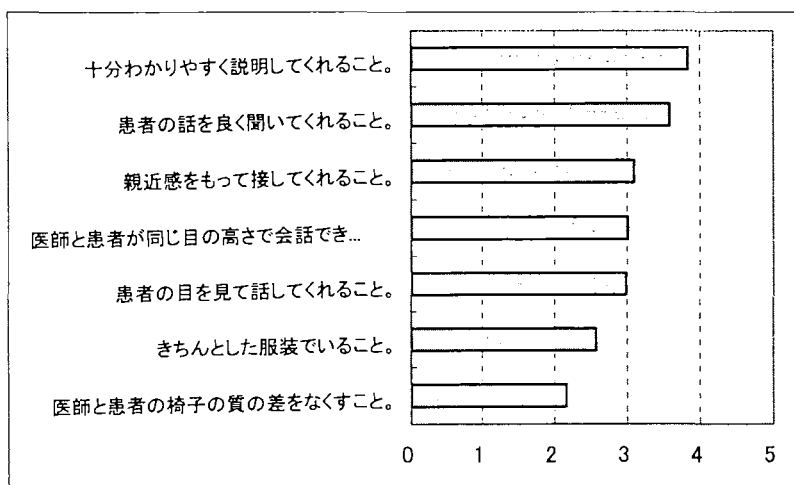


図4 医師・看護婦・その他の病院のスタッフに対して望むこと(N=842)

図4より、要望の強い順に、選択肢は以下の3つに分類することができる。

分類	質問
説明・コミュニケーション (インフォームド・コンセント)	“十分わかりやすく説明してくれること。” “患者の話を良く聞いてくれること。”
しぐさ・まなざし (間接的なコミュニケーション)	“親近感をもって接してくれること。” “医師と患者が同じ目の高さで会話できること。” “患者の目を見て話してくれること。”
医療スタッフの外見的な特徴	“きちんとした服装でいること。” “医師と患者の椅子の質の差をなくすこと。”

3.6 病院のスタッフに対する希望

病院のスタッフへの希望に対する回答を、SAS/STAT を用いて因子分析を行った。結果を下の表 2 に示す。

表 2 病院のスタッフへの希望についての因子分析結果

質問項目	コミュニケーション因子	外見因子	共通性
“十分わかりやすく説明してくれること。”	<u>0.75696</u>	-0.10097	0.58
“患者の話を良く聞いてくれること。”	<u>0.74200</u>	0.06181	0.55
“親近感をもって接してくれること。”	<u>0.63078</u>	0.29716	0.49
“患者の目を見て話してくれること。”	<u>0.58856</u>	0.54982	0.65
“医師と患者の椅子の質の差をなくすこと。”	0.01748	<u>0.81016</u>	0.66
“きちんとした服装でいること。”	-0.00343	<u>0.71721</u>	0.51
“医師と患者が同じ目の高さで会話できること。”	<u>0.45696</u>	0.64426	0.62
説明分散	2.08	1.99	

得点の平均点で求めた分類と、因子はおおむね上記のように対応するといえる。なお、“患者の目を見て話してくれること。”及び“医師と患者が同じ目の高さで会話できること。”は、両方の因子に同程度の寄与を及ぼしている。

4. 医療情報システムへのニーズ

“入院中に、パソコンや携帯端末から病院内の医療情報にアクセスできるとした場合、どのようなことができればよいと望まれますか。”という問への結果を図 5 に示す。なお、診療歴別に大きな違いは無かった。

この他に、“病院内から病院外へインターネットで接続可能な場合”及び“ご家族が入院されている場合、自宅や職場等からインターネットを利用してできること”についてもニーズ調査を行った。なお、いずれの質問でもセキュリティは保証されている前提とした。また、インターネットアンケートの特長である自由記述の処理の容易さを利用して、自由記述の分析も行った。なお、これらの調査結果をもとに”患者参加型医療情報交換システム“が開発された[7]。

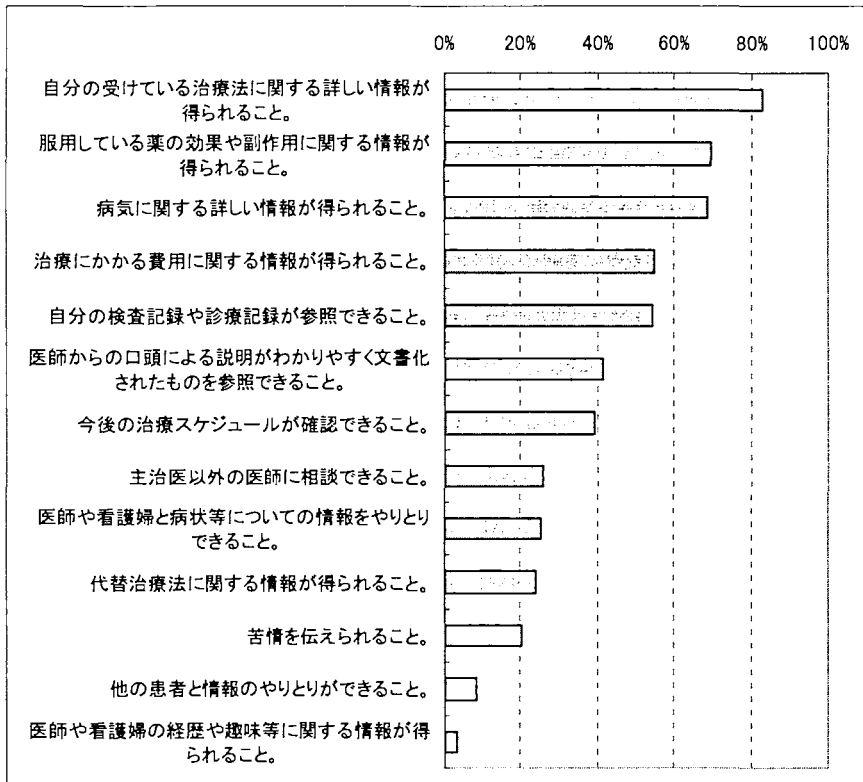


図 5 医療情報システムへのニーズ(N=1,684)

5. その他

セキュリティ技術の進展にともない、将来、自分自身の検査記録や診療記録を自己管理することが可能となった場合を想定して、情報の自己管理の希望を質問した。結果は、“非常に自己管理したい。”(28.0%)、“ある程度は自己管理したい。”(60.5%)との回答だった。したがって、9割程度の回答者が、環境が整備されれば検査記録や診療情報を“自己管理”する希望をもっていることが確認できた。また、インターネットや超高速ネットワークを利用した高度医療に関しても、“非常に期待する”(45.6%)または“やや期待する”(48.3%)との回答だった。

最後に、今後の医療に関して注目される 9 項目について、重要度を質問した。結果は図6に示す通りである。図の下に示した項目ほど、“わからない”の占める割合が増えている。この点に関しては、今後一般への分かりやすい情報の提供等が望まれる。

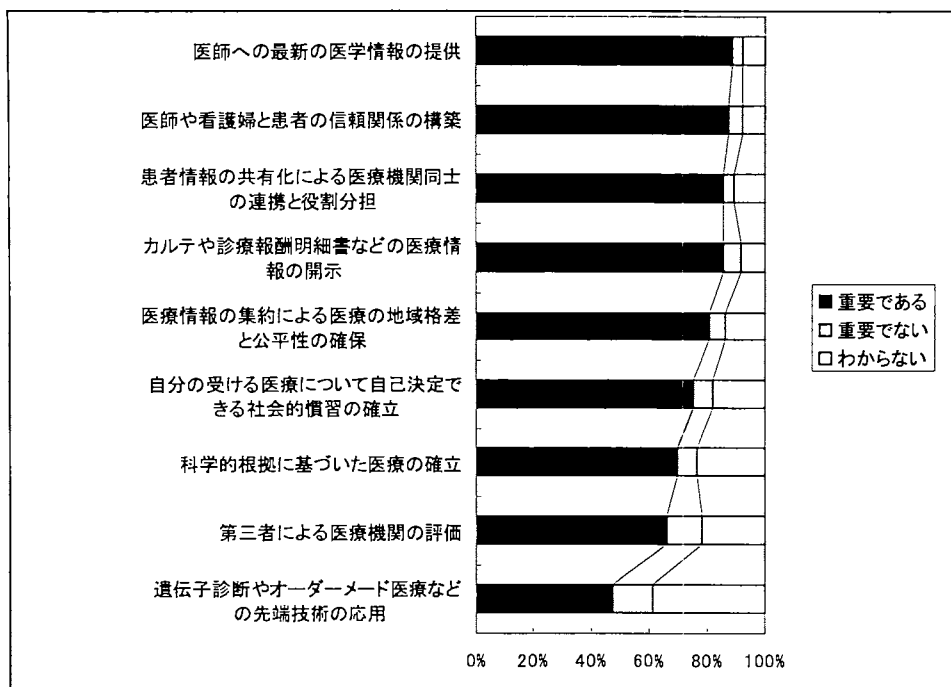


図6 医療に関連する様々な項目の重要度(N=1,684)

6. まとめと今後の課題

インターネットアンケートにより患者参加型医療情報交換システムのニーズ調査を行った。調査の結果から、インターネットを用いた情報提供や交換のニーズが高いものが確認できた。また、医療情報の自己管理に対するニーズが高いことも分かった。

なお、今回は直接の調査対象としなかったが、最近注目が高まっている医療の安全に関して、患者参加によるインシデントの低減などが指摘されている[8]。今後は、医療の安全の視点に立った患者参加型情報交換システムの重要性も高まると考えられる。

参考文献

- 1) 総務省編,「平成 14年版情報通信白書」,ぎょうせい,(2002).
- 2) 福田吉治ら,「インタラクティブ・ヘルスコミュニケーションの現状と効果に関する研究:医療機関のホームページに関する意識調査とその現状」,医療と社会,vol.11,No.3,pp.43-54(2002).
- 3) 橋本栄里子ら,「インターネット上の病院の情報発信内容に関する研究:病院のホームページは患者に何を伝えているのか」,vol.11,No.3,pp.69-87(2002).
- 4) インターネット医療協議会,「インターネット上の医療情報の提供と利用の実態に関する調査研究報告」<http://www.jima.or.jp/JISSEKI/kousei1999.html>(2003年5月現在).
- 5) gooリサーチ,<http://research.goo.ne.jp/business/top.html>(2002年5月現在).

- 6) 日本看護協会, 「2000 年患者への診療情報提供に関する調査」, 日本看護協会調査研究報告 No.61, (2001).
- 7) J. Funabiki, N. Nobuaki, H. Tsunoda, H. Oyama, "Development of a web-based system for patient-oriented interactive health communication", Japan Journal of Medical Informatics, Vol.22-Supplement, p.692(2002).
- 8) 村上陽一郎ら, 「リスクマネジメント 医療内外の提言と放射線部の実践」, 医療科学社,(2002).

謝辞

本研究は、平成 13, 14 年度科学技術振興調整費「高速ネットワーク環境下における高度医療アプリケーションの研究開発」の一環として実施された。

ポスターセッション
グラフィック・統計教育

日本SASユーザー会 (SUGI-J)

SAS/GRAPH 入門 ～社内における教育研修事例～

○林 行和 畑中 雄介 小出 起美雅 山口 孝一
株式会社 CRC ソリューションズ/CRO 業務部 DM・統計解析チーム

Introductory SAS/GRAPH ～An Introduction of In-house Training Course～

Yukikazu Hayashi Yusuke Hatanaka Kiminori Koide Koichi Yamaguchi
CRC Solutions Corp.
CRO Department Data Management & Biostatistics Section

要 旨

市販の表計算ソフトなどによるグラフ作成と違い、SAS/GRAPH によるグラフの作成は細部までを思うように作り込むことができ、非常に有用なツールである。そこで弊社における SAS/GRAPH 初心者のための教育研修を紹介する。

キーワード： SAS/GRAPH、社内教育、Windows 版 SAS System 8、ANNOTATE MACRO

1. はじめに

弊社のような CRO 業務を行っていると、クライアントの希望により細部まで指示のあるグラフを作成する必要性が出てくる。SAS/GRAPH は ANNOTATE MACRO 等を使うことによりその仕様に沿ったグラフを作成することができる有用なツールである。しかしながら、ANNOTATE MACRO までを説明した初心者のための入門書になるようなものがなかなか見当たらない。弊社でも結局、業務で必要性が出た時に随時、グラフ作成経験者が教える状況であり、無駄が多く、情報の共有化の観点からも問題がある。そこで、SAS/GRAPH 初心者を対象とした SAS/GRAPH の教育研修を紹介する。

2. まずはグラフを描いてみよう！

SAS/GRAPH の代表的なプロシジャに GPLOT プロシジャというものがあります。まずは

GPLOT プロシジャを使ってどんなグラフが描けるのか試してみましょう。

次のような3群(KEY)5例ずつ計15例に対し、0,2,4の3時点で測定値が存在するデータセット WRK を使います。GPLOT プロシジャは PLOT というだけあって、最も簡単に描けるグラフは散布図です。縦軸に時点0の測定値(VAL0)、横軸に時点2の測定値(VAL2)となるような散布図を描いてみましょう。

```
/* データ作成 */
data WRK;
  input KEY PATNO VAL0 VAL2 VAL4;
  cards;
    1 1 156 155 156
    1 2 157 154 155
    1 3 154 153 153
    1 4 155 153 153
    1 5 152 153 156
    2 6 156 154 150
    2 7 154 155 153
    2 8 156 154 148
    2 9 153 154 150
    2 10 157 156 152
    3 11 151 150 147
    3 12 157 151 145
    3 13 156 148 142
    3 14 158 152 143
    3 15 155 147 145
  ;
run;

*---- 散布図 ----;
proc gplot data=WRK;
  plot VAL0*VAL2;          *- 縦軸 × 横軸 の順で -;
run;
quit;
```

散布図を描くには PLOT ステートメントを使用します。

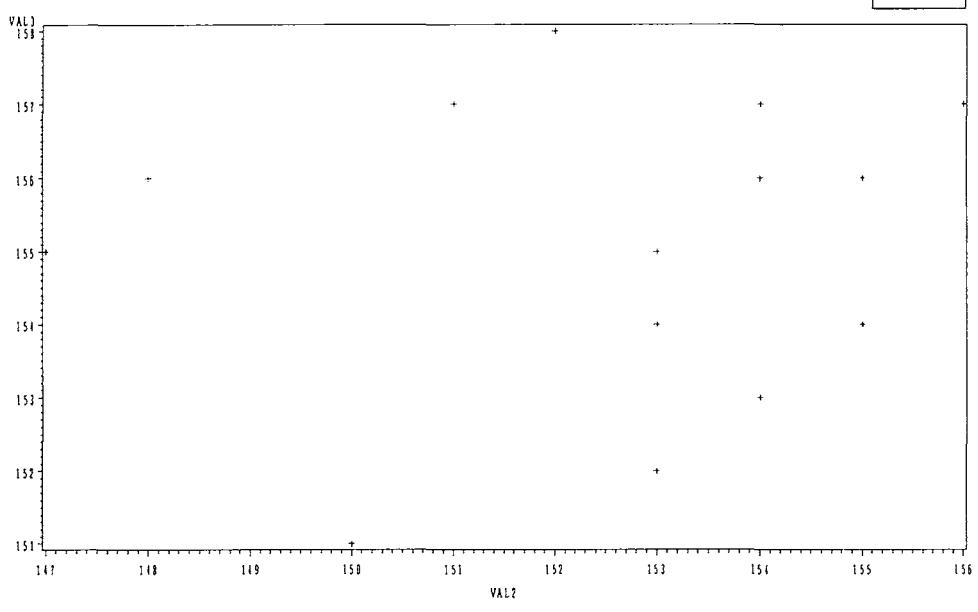
PLOT [縦軸の変数]*[横軸の変数] ;

の順で指定しないと縦横が入れ替わってしまいます。そこだけ注意しましょう。

以下のようなグラフが描けたでしょうか？(Fig1)

このグラフから分かるように、OUTPUT そのままを縮小して貼り付けを行なうと、ラベルの文字が小さすぎて潰れてしまったり、枠ギリギリにプロットがあり見にくかったりします。その見栄えを整えるのが最終目標です。

Fig1



3. 推移図を描いてみよう！

見栄えを整える前に GPGLOT プロシジャでできることをいろいろ見てみましょう。そこで、散布図の次に推移図を描いてみましょう。縦軸に測定値、横軸に時点となるような推移図を考えます。そのような推移図を作成するには、データセットの構造を変えた方が便利です。

```

*- データを縦型に変換 --;
proc sort data=WRK;
  by KEY PATNO;
run;
proc transpose data=WRK out=WRK1;
  by KEY PATNO;
  var VAL0 VAL2 VAL4;
run;
data WRK1; set WRK1;
  if _NAME_='VAL0' then VISIT=0;
  if _NAME_='VAL2' then VISIT=2;
  if _NAME_='VAL4' then VISIT=4;
  drop _NAME_;
  rename COL1=VAL;
run;

*-- 推移図 ----;
proc gplot data=WRK1;
  plot VAL*VISIT=PATNO;          /* Y軸変数*X軸変数=紐付け変数 の順 */

```



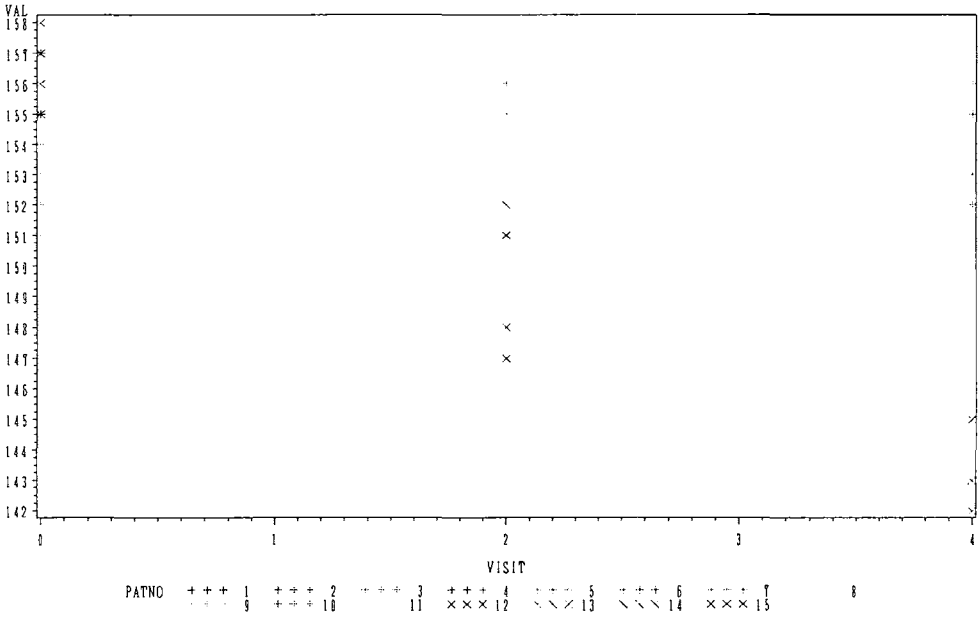
```
run;
quit;
```

推移図を描くには、縦軸と横軸、そして横軸の推移を紐付ける変数が必要になります。PLOTステートメントを使い、

```
PLOT [縦軸の変数]*[横軸の変数]=[紐付け変数] ;
```

と指定します。すると次のようなグラフが描けます。

Fig2



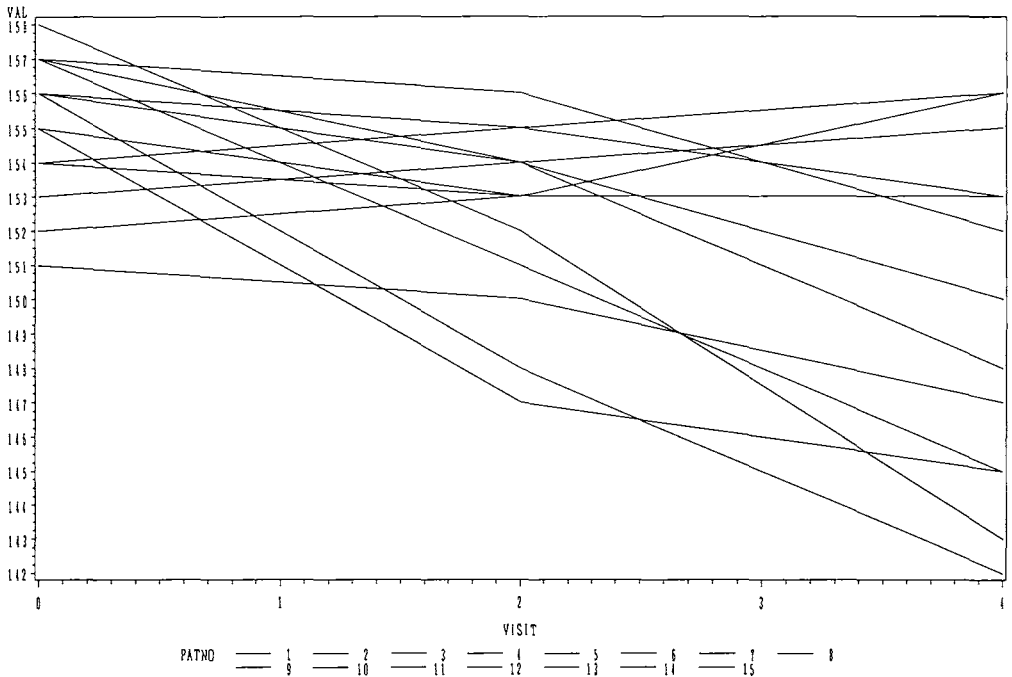
色やシンボルの形を変えることで紐付けされていますが、白黒出力だと、どの値が対になっているかわかりません。推移図ですから推移を線で結びましょう。SYMBOLステートメントを使います。

```
proc gplot data=WRK1;
  plot VAL*VISIT=PATNO;
  symbol v=none c=black i=join l=1 r=12;
run;
quit;
```

ここで使用した SYMBOL ステートメントのオプションは、

- v= :プロットの種類、
- c= :線やプロットの色、
- i= :補間線
- l= :補間線の線種、
- r= :指定した symbol の繰り返し数

Fig3



4. 体裁を整えよう！

さて、それでは、他の文書に貼り付けても見栄えがよくなるように調整していきましょう。縦軸横軸共にラベルを大きくした方が見やすそうです。目盛も工夫してみましょう。

```

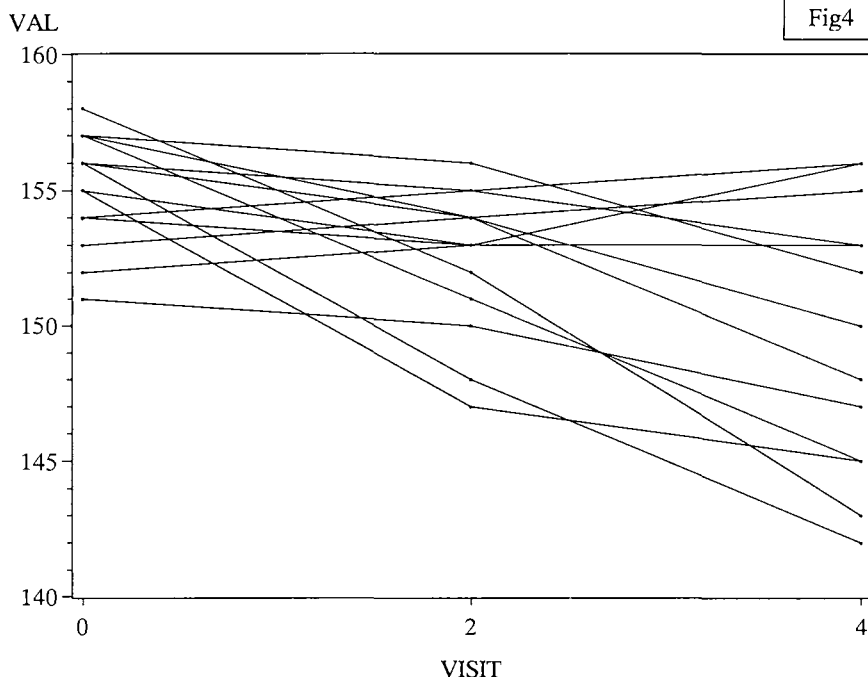
goptions gunit=pct;

proc gplot data=WRK1;
  plot VAL*VISIT=PATNO / nolegend vaxis=axis1 haxis=axis2;
  symbol v=dot h=0.5 c=black i=join l=1 r=15;
  axis1 label=(font='Times New Roman' h=4.0 offset=(0, 0)
    minor=(n=4 h=0.5) major=(w=1 h=1.0)
    length=75pct width=2 order=(140 to 160 by 5)
    value=(font='Times New Roman' h=4.0) origin=(15,15);
  axis2 label=(font='Times New Roman' h=4.0 offset=(1,1)
    minor=none major=(w=1 h=1.0)
    length=80pct order=(0 to 4 by 2) width=2
    value=(font='Times New Roman' h=4.0) origin=(15,15);
run;
quit;

```

AXIS ステートメントを使い、縦軸、横軸のラベル等を細かに調整することができます。これらのオプションによって規定する単位は[GOPTIONS UNIT=]で指定することができます。複数の端末でプログラムを流す可能性があることを考えると、pct を指定しておいた方がよいでしょう。

LABEL …変数のラベルの表示、 OFFSET …軸のどこから目盛をスタートするか、
 MINOR,MAJOR …目盛の線の設定、 LENGTH …軸の長さ、 WIDTH …軸の太さ、
 ORDER …軸上の目盛の飛び幅、 VALUE …目盛の値の設定、 ORIGIN …始点



随分と見栄えがよくなったんじゃないでしょうか？どの症例も同じ SYMBOL で設定したので凡例を外しました。PLOT ステートメントの NOLEGEND がその設定です。

5. 体裁を整えよう！ part2(Mean ± SD)

次に群毎の平均値 ± 標準偏差をグラフにしてみましょう。± 標準偏差の部分のバーなどは ANNOTATE MACRO を使用することによって自分の好きなように調整できます。まずは平均値と標準偏差の計算。

```
proc univariate data=WRK1 noprint;
  var VAL;
  by KEY VISIT;
  output out=WRK2 mean=MEAN std=SD;
run;
```

± 標準偏差の縦線とバーは ANNOTATE MACRO の %LINE を使用します。

```
data ANNO1; set WRK2;
  %dclanno;
  %system(2, 2, 3);
  if SD>0 then do;
```

```

%line(VISIT, MEAN-SD, VISIT, MEAN+SD, black, 1, 0.2); *- SD縦線 -;
%line(VISIT+0.2, MEAN-SD, VISIT+0.2, MEAN-SD, black, 1, 0.2); *-上横線 -;
%line(VISIT-0.2, MEAN+SD, VISIT-0.2, MEAN+SD, black, 1, 0.2); *-下横線 -;
end;
if SD<=0 then delete;
run;

```

座標(x1, y1)と(x2, y2)を結びたい時に

%LINE(x1, y1, x2, y2, 線の色, 線の種類, 線の太さ);

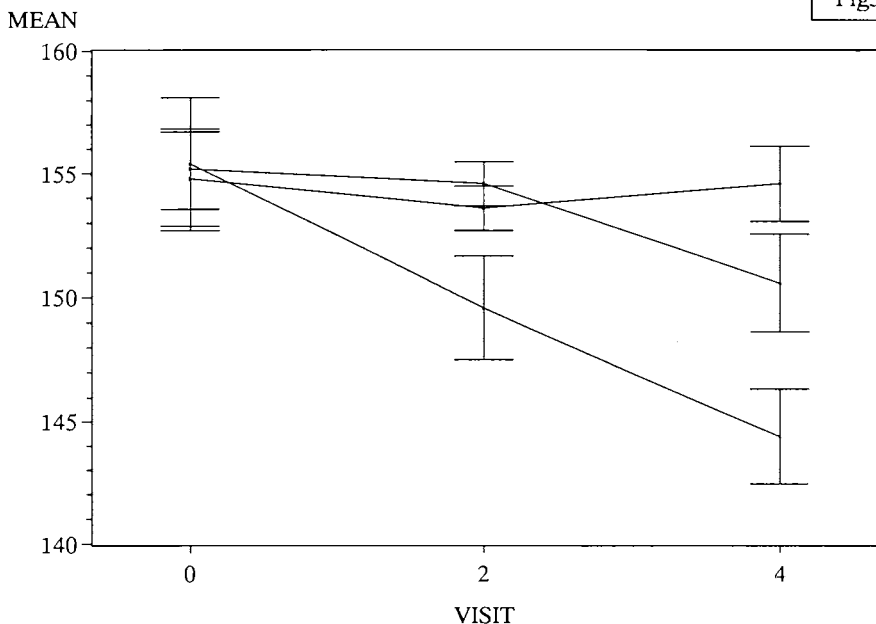
という指定の仕方をします。ANNOTATE MACRO の x 値、y 値が座標内の値であれば、%system(xs, ys, hs); の xs, ys の部分を2、画面内の%値にしたい場合は3を指定します。ANNOTATE MACRO で作成したデータセットは GPGLOT プロシジャの PROC GPGLOT ステートメントのオプションか、もしくは、PLOT ステートメントのオプションで指定することができます。

```

goptions gunit=pct;

proc gplot data=WRK2;
plot MEAN*VISIT=KEY / nolegend vaxis=axis1 haxis=axis2 anno=ANN01;
symbol v=dot h=0.5 c=black i=join l=1 r=3;
axis1 label=(font='Times New Roman' h=4.0) offset=(0, 0) minor=(n=4 h=0.5)
major=(w=1 h=1.0) length=70pct width=2 order=(140 to 160 by 5)
value=(font='Times New Roman' h=4.0) origin=(15,15);
axis2 label=(font='Times New Roman' h=4.0) offset=(10,10) minor=none
major=(w=1 h=1.0) length=80pct order=(0 to 4 by 2) width=2
value=(font='Times New Roman' h=4.0) origin=(15,15);
run;
quit;

```



ここでは、PLOT ステートメントのオプションとして[ANNO=]の部分で、ANNOTATE MACRO により作られたデータセットを指定しています。MEAN±SD の上と下のバーに長さが出てきたため、症例毎の推移図の X 軸より余白を持たす必要があります。それを X 軸側の設定である AXIS2 ステートメントの OFFSET オプションで指定しています。そうすると次のような MEAN±SD のグラフが出来上がります。(Fig5)

このグラフを見てどうでしょう？3 群が重なって見づらいですね？3 群を少しずつずらし打ち出したり、SYMBOL を変えたり、線種を変えたりで群を見やすく工夫しましょう。さらにラベルやタイトルを整えて見栄えをよくしてみましょう。それを最後の章で実践します。

6. 見栄えを完成させよう！

ANNOTATE MACRO を使用し、見栄えの最終調整です。ANNOTATE MACRO を使用するには%ANNOMAC; の一文を実行します。またカタログファイルへの保存方法、更に、他のファイルに貼り付けるのに有用な拡張メタファイル(EMF)への変換方法を紹介します。

```
%ANNOMAC;

*- VISIT ずらし --;
data WRK3; set WRK2;
  if KEY=1 then VISIT=VISIT-0.3;
  if KEY=3 then VISIT=VISIT+0.3;
run;

*- SDバー ---;
data ANNO1; set WRK3;
  %dclanno;
  %system(2, 2, 3);
  if SD>0 then do;
    *- SD縦線 -;
    %line(VISIT, MEAN-SD, VISIT, MEAN+SD, black, 1, 0.2);
    *- MEAN-SD横線 -;
    %line(VISIT-0.1, MEAN-SD, VISIT+0.1, MEAN-SD, black, 1, 0.2);
    *- MEAN+SD横線 -;
    %line(VISIT-0.1, MEAN+SD, VISIT+0.1, MEAN+SD, black, 1, 0.2);
  end;
  if SD<=0 then delete;
run;

*- タイトル、縦軸、横軸表記 --;
data ANNO2;
  length TEXT $80;
  %dclanno;
  %system(3, 3, 3);
  *- タイトル -;
  %label(50, 95, "推移図(Mean±SD)", black, 0, 0, 5.0, 'MS ゴシック', 5);
  *- Y軸ラベル -;
```

```

%label(5, 50, "測定値", black, -89.99, 90, 4.0, 'MS ゴシック', 5);
*- X軸ラベル -;
%label(50, 5, "時点", black, 0, 0, 4.0, 'MS ゴシック', 5);
*- 凡例 -;
%line(80, 11, 90, 11, black, 1, 0.2);
%line(80, 7, 90, 7, black, 2, 0.2);
%line(80, 3, 90, 3, black, 14, 0.2);
%label(85, 11, "●", black, 0, 0, 2.0, 'MS ゴシック', +);
%label(85, 7, "○", black, 0, 0, 2.0, 'MS ゴシック', +);
%label(85, 3, "□", black, 0, 0, 2.0, 'MS ゴシック', +);
%label(91, 11, ": KEY1", black, 0, 0, 3.0, 'MS ゴシック', >>);
%label(91, 7, ": KEY2", black, 0, 0, 3.0, 'MS ゴシック', >>);
%label(91, 3, ": KEY3", black, 0, 0, 3.0, 'MS ゴシック', >>);
run;
*- X軸目盛 --;
data ANNO3;
  length TEXT $80;
  %dclanno;
  %system(2, 3, 3);
  *- 目盛 -;
  %line(0, 20, 0, 19, black, 1, 0.2);
  %line(2, 20, 2, 19, black, 1, 0.2);
  %line(4, 20, 4, 19, black, 1, 0.2);
  *- 目盛値 --;
  %label(0, 17.5, "0", black, 0, 0, 4.0, 'Times New Roman', 5);
  %label(2, 17.5, "2", black, 0, 0, 4.0, 'Times New Roman', 5);
  %label(4, 17.5, "4", black, 0, 0, 4.0, 'Times New Roman', 5);
run;
*-----;
data ANNO2; set ANNO2 ANNO3;
run;

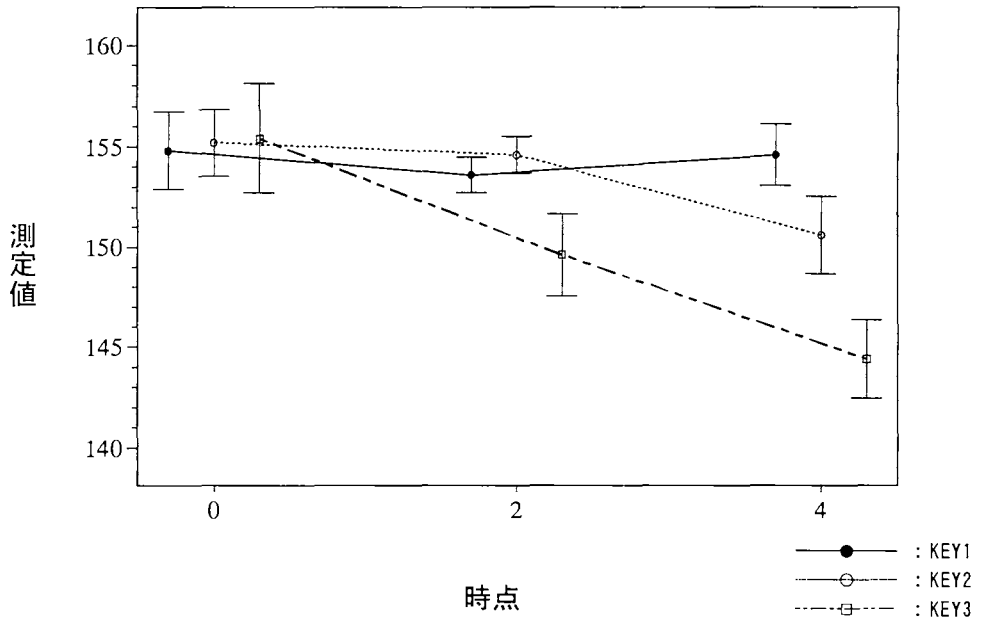
*-- EMFファイルへ ---;
FILENAME FIG "C:%SAS_FIG_EMF%FIG.emf";
GOPTIONS RESET=ALL GUNIT=PCT TARGET=LIPS3A4 DEVICE=EMF GSFNAME=FIG GSFMODE=REPLACE
          ROTATE=LANDSCAPE;

goptions gunit=pct;
proc gplot data=WRK3 anno=ANNO2;
  plot MEAN*VISIT=KEY / nolegend vaxis=axis1 haxis=axis2 anno=ANNO1;
  symbol1 v=dot h=1.8 c=black i=join l=1 r=1 w=2;
  symbol2 v=circle h=1.8 c=black i=join l=2 r=1 w=2;
  symbol3 v=square h=1.8 c=black i=join l=14 r=1 w=2;
  axis1 label=none offset=(5, 5)
        minor=(n=4 h=0.5) major=(w=1 h=1.0)
        length=65pct width=2 order=(140 to 160 by 5)
        value=(font='Times New Roman' h=4.0) origin=(15,20);
  axis2 label=none offset=(0,0)
        minor=none major=none
        length=75pct order=(-0.5 to 4.5) width=2
        value=none origin=(15,20);
run;

```

推移図 (Mean ± SD)

Fig6



SAS System 8より、更に様々なタイプの True Fontが使用できるようになりました。しかし、ファイルの保存形式によってはサポートされない Font もあり、他文書への貼り付け時の見え等を考慮すると、EMF ファイルが汎用的であるようです。

[参考文献]

- (1)「SAS/GRAPH リファレンスガイド Release6.03 Edition」SAS Institute Japan 株式会社
- (2)「SAS プロシジャリファレンス Version6,First Edition」SAS Institute Japan 株式会社

ポスターセッション
グラフィック・レポーティング

SAS グラフによる動く万華鏡の作成

岸本 容司

神戸商科大学

経営学研究科 経営情報科学専攻

Moving Kaleidoscopes by SAS/GRAPH

Yoji Kishimoto

Graduate School of Business Administration, Kobe University of Commerce

要 旨

昔からある万華鏡という玩具のカラフルな絵模様を、SAS/GRAPH を使って描画し、GIF アニメーションとして出力することを試みた。当初は、従来の万華鏡で見えるような絵模様の静止画像を想定していたが、開発を進めるうちに、動きのある方がよりダイナミックに見えるので、絵模様が回転したり、移動したり、色が微妙に変化していくといった多様な機能を付加していった。その結果、通常の万華鏡のイメージとは少し異なるものとなった。種々のマクロ変数を SAS プログラムの冒頭部で設定することにより、多彩なパターン絵模様が表示できるようにした。

キーワード： SAS/GRAPH、GIF アニメーション、万華鏡、マクロ変数

1. はじめに

万華鏡とは、三枚の鏡板を組んだ三角柱の中に種々の色ガラスや色紙の小片を入れたもので、回しながらのぞいて模様の変化を見て楽しむ一種の玩具である。その万華鏡の中に見える刻々と変化する模様をディスプレイ上で表現しようと考え、SAS の GIF アニメーションを作成する機能と乱数を用いてプログラムの作成を試みた。当初は、従来の万華鏡で見えるような絵模様の静止画像を想定していたが、開発を進めるうちに、動きのある方がよりダイナミックに見えるので、絵模様が回転したり、移動したり、色が微妙に変化していくといった多様な機能を付加していった。その結果、通常の万華鏡のイメージとは少し異なるものとなったが、本論文では便宜上、作成される絵模様を万華鏡と呼ぶことにする。

2. プログラムの概要と流れ

万華鏡の模様を表現するために、このプログラムでは直線を用いている。そのために、まず、

その直線を描くために必要な座標点を求める (図1の①)。また、GIFアニメーションは複数の静止画を連続して表示することによって、動いているように見せている。そのため、静止画を複数枚作成する (図1の②)。このときの静止画1枚1枚のことをここではフレームと呼ぶことにする。また、このプログラムでは、1つ目の万華鏡のデータをフレーム枚数分作成してから、2つ目、3つ目、・・・と万華鏡のデータを追加することで複数の万華鏡を同時に表示することもできる (図1の③)。その全てのデータを graphbase という SAS データセットに一旦保存をした後に、GIFアニメーション用の GIF ファイルに一度に書き出している。こうすれば、出力が1回で済むため、1フレーム毎に GIF ファイルへ書き出すより、プログラムの実行時間が節約できる。

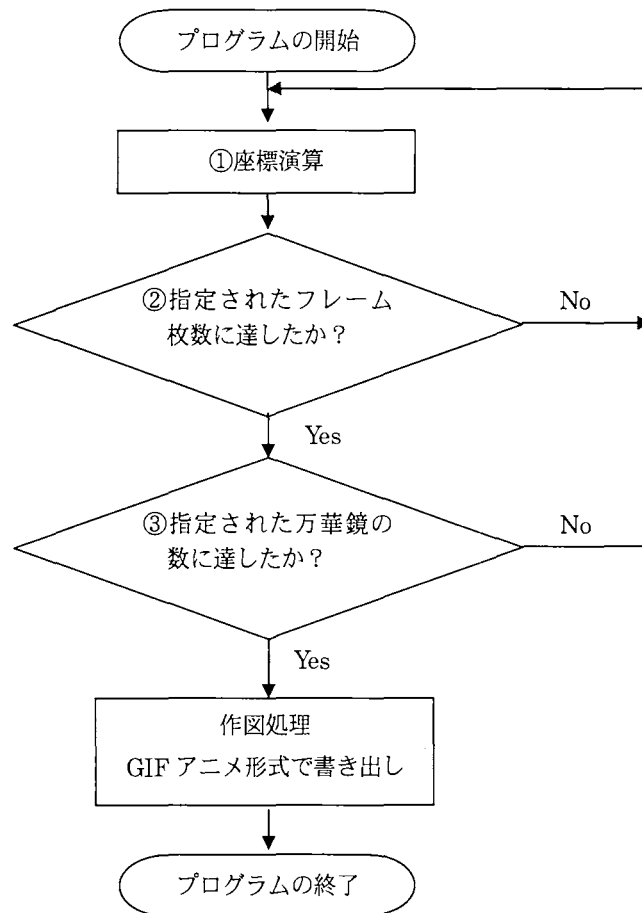


図1 プログラムの流れ

3. 万華鏡の表示パターンを設定するパラメータ

多彩な模様を描くために、多様なパラメータを用意した。それらはプログラムの冒頭部ですべてのマクロ変数を設定することで実現している。従って、プログラムを実行する際に、マクロ変数の設定を変えるだけで、まったく異なる万華鏡が作成できる。マクロ変数で指定できる

主な機能は以下の通りである。

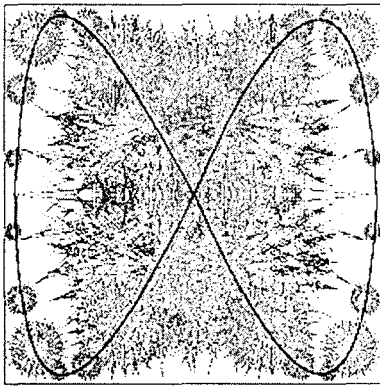
●GIF アニメーション

- ・フレーム枚数
- ・1枚分のフレームがディスプレイ上に表示される時間

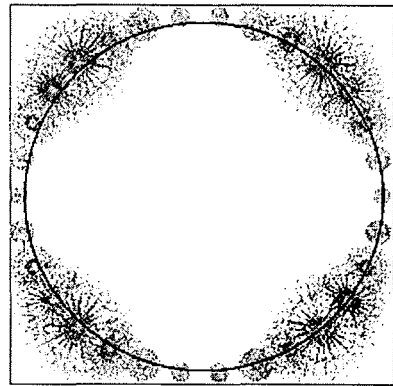
●万華鏡の5種類の移動パターン

- ①常に中心にあって不動、②∞状に移動、③枠に沿って移動(1)、④枠に沿って移動(2)、⑤ランダムに移動

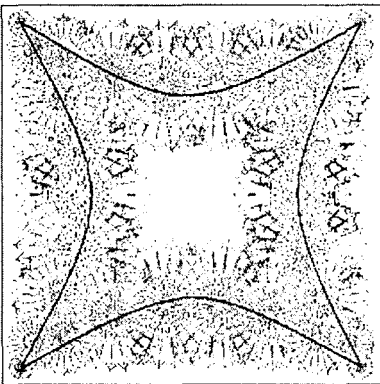
図2のaからdに②～⑤の移動パターンを示す。



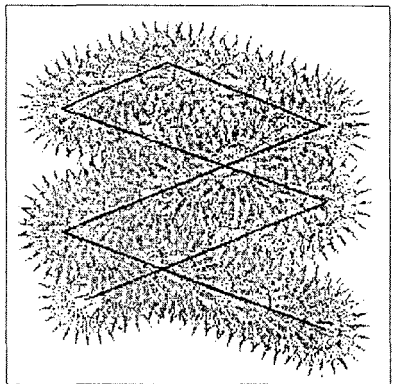
a ∞状に移動



b 枠に沿って移動(1)



c 枠に沿って移動(2)

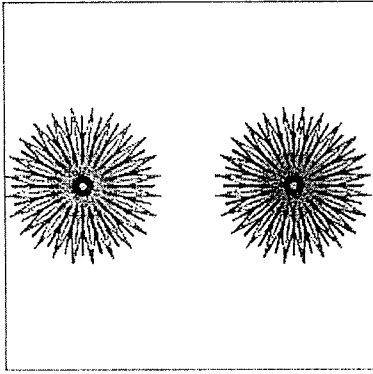


d ランダムに移動

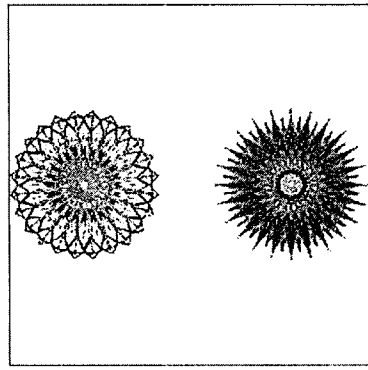
図2 移動パターン

●万華鏡の数

- ・個数
- ・複数個あるとき、①軌跡が1つ、②軌跡が万華鏡の数と同じ数
- ・複数個あるとき、①同じ絵模様、②別の絵模様



a 同じ絵模様

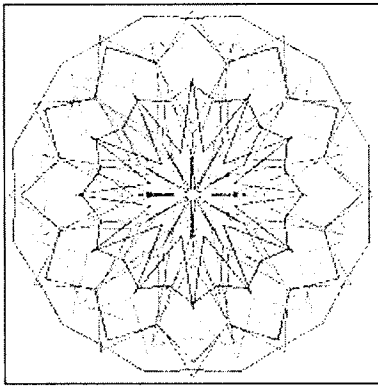


b 別の絵模様

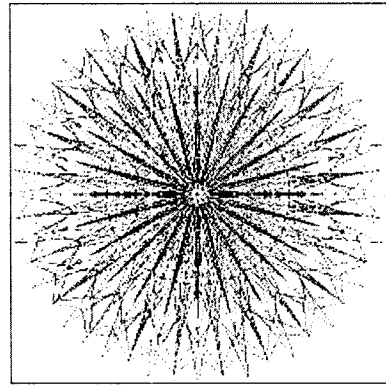
図3 万華鏡が複数個あるときの絵模様のパターン

●万華鏡の形状

- ・万華鏡を構成する基本の扇形の個数
- ・基本の扇形の中にある線の本数



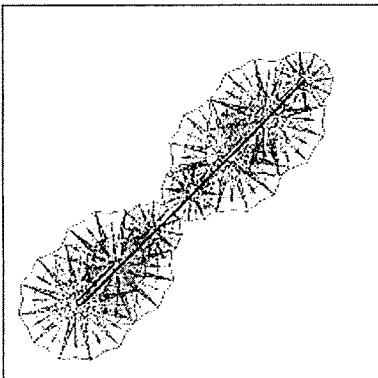
a 扇形の数 24、線の数 15



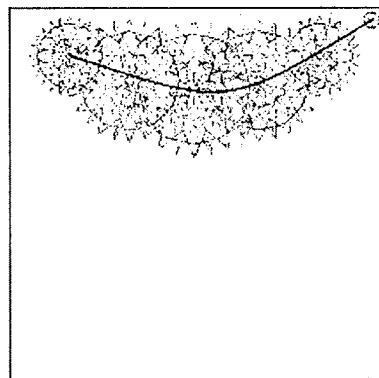
b 扇形の数 48、線の数 100

図4 万華鏡の基本の扇形の数とそこにある線分の本数

- ・万華鏡の半径 ①固定、②大小の繰り返し、③外枠までの最短距離



a 大小の繰り返し



b 外枠までの最短距離

図5 万華鏡の半径のパターン

- ・移動範囲
- ・回転速度
- ・万華鏡の模様
 - ①すべてのフレームが同じパターン、
 - ②アニメーション中にパターンが1回変化、
 - ③1フレームごとにパターンが変化

●万華鏡を構成する線

- ・太さ

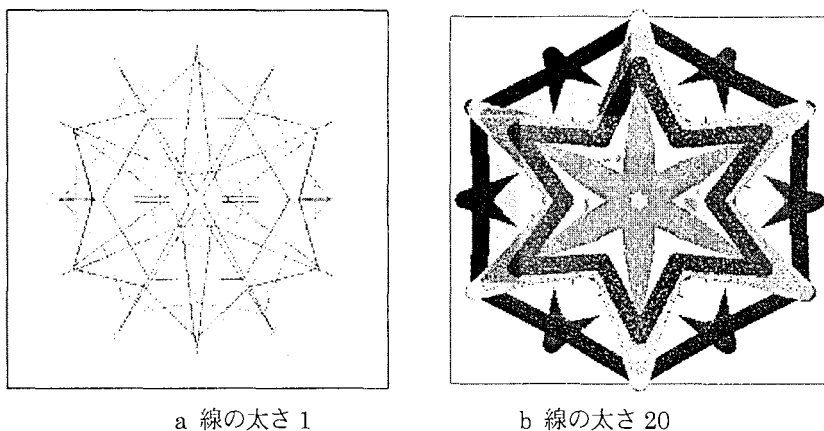


図6 線の太さ

- ・色もしくは、赤、緑、青の比率を指定
 - ・アニメーション時のグラデーションの有無
- 万華鏡の表示される範囲を示す枠
- ・枠の形 ①正方形、②横長の長方形
 - ・枠、タイトルの有無
- (このパラメータ「なし」に設定すると、万華鏡のみの GIF アニメが作成できる。)

これらマクロ変数の値の詳細についてはプログラム中のコメントを参照されたい。

4.まとめ

本論文で解説した万華鏡作成プログラムから出力される GIF ファイルは動きを伴うものである。そのため、本論文上には実際のディスプレイ上で見えるカラフルなアニメーションは掲載できない。SAS プログラムおよびサンプル画像は無料でダウンロードできるようにしている。<http://mighty.kobeuc.ac.jp/sugi-j/2003.html> ただ、将来、プログラムの機能のアップなどで本論文に掲載しているプログラムと若干異なる場合もあり得る。今回は直線を使って、絵模様を作成したが、将来、任意の図形を組み合わせる万華鏡のパターンを作成するプログラムを試作したい。

【謝辞】

本論文を作成するに当たり、神戸商科大学の周防節雄教授と古隅弘樹講師から様々なコメントやアドバイスを頂いた。また、SAS Institute Japan 株式会社の漁智一氏からは貴重な情報を頂戴した。ここに記して謝意に替えたい。

【参考文献】

- [1]長谷川 要(2001)「SAS/GRAPH ソフトウェアを用いたフラクタル図形の作成」『第20回 SAS ユーザー会総会および研究発表会論文集』,pp322-323
- [2]長谷川 要(2002)「スピログラフを再現しよう—GIFANIM Device Driver を用いたアニメーション図形の作成—」『第21回 SAS ユーザー会総会および研究発表会論文集』,pp587-592
- [3]SAS/GRAPH Sample Programs http://support.sas.com/techsup/sample/sample_graph.html

付録 プログラムリスト

```
/* mangekyo.sas */ /* 万華鏡の模様をGIFアニメーションにして書き出すプログラム */
options nosource nonotes nomprint;

/* GIFアニメーションについて */
%let pitch=100; /* GIFアニメーションのフレーム数 推奨値 50~300 */
/* 10未満の時、万華鏡は常に中心にあり、大きさ不変 */
/* 推奨値 50~200 (1の時、静止画像) */
%let frame_time=10; /* GIFアニメの1フレームの表示時間(単位0.01秒) 推奨値 5~20 */

/* 万華鏡の移動について */
%let move=0; /* 0=常に中心、1=∞状に動く、2=枠に沿って動く(1)、
3=枠に沿って動く(2)、4=ランダム軌道 */
%let move_speed=10; /* 万華鏡の移動速度 move=4のときのみ有効 (単位pct) */

/* 万華鏡の数について */
%let howmany=1; /* 万華鏡の数 */
%let shape=0; /* 複数個ある万華鏡の形状のパターン
0=別々 1=同じ(howmany>1のときのみ有効) */
%let shift=0; /* 複数個ある万華鏡の移動パターン
0=同じ 1=別々(howmany>1かつmove=1,2,3のときのみ有効) */

/* 万華鏡の形について */
%let divide=48; /* 万華鏡の分割数 偶数のみ 推奨値 12~128 */
%let line_no=15; /* 万華鏡を構成する線の数 推奨値 20~100(line_width=1)
5~25(line_width=3) */
%let radius=0; /* 万華鏡の半径の変化パターン
0=変化なし、1=小さくなってから元の大きさに戻る、
2=中心点から枠までの最短を半径とする */
/* 万華鏡が移動し、この値が0または1のとき最大値 100=move_range */
%let move_range=95; /* 移動時の万華鏡の移動範囲(単位pct) 推奨値 30~95 */
%let shape_pattern=1; /* 万華鏡の形状の変更パターン
1=変化なし、2=二種類、0=フレームごとに別のパターン */
%let round_speed=-360; /* 万華鏡の回転速度(単位°) ("360"でアニメーション中に一回転) */

/* 万華鏡を構成する線について */
```

```

%let line_width=2;          /* 線の太さ 推奨値 1~5 (単位pct) */
%let color_pattern=1;      /* 線の色指定
                           0=一色を指定、1=輪ごとに変化、2=全ての線が別の色 */
%let line_color=cxff00ff;  /* color_pattern=0のときのみ有効(色の指定はRGBまたはnameで指定) */

/* 色の濃淡(color_pattern=1または2のときのみ有効) */
%let r_id=1;               /* 0=赤色のグラデーションなし、1=あり */
%let g_id=1;               /* 0=緑色のグラデーションなし、1=あり */
%let b_id=1;               /* 0=青色のグラデーションなし、1=あり */
%let r_dense=0;            /* 赤色の濃度 2=赤の濃淡のみ、1=赤色が濃い、0=普通 */
%let g_dense=0;            /* 緑色の濃度 2=緑の濃淡のみ、1=緑色の濃い、0=普通 */
%let b_dense=0;            /* 青色の濃度 2=青の濃淡のみ、1=青色の濃い、0=普通 */
/* 上記の3つのうち2つ以上が1のとき全体的に暗色 */
/* 上記の3つのうち2つ以上が2のとき黒一色 */

%let back_color=cxf0ffff; /* 背景の色(色の指定はRGBまたはnameで指定) */

%let framework=0;         /* 枠の有無 0=あり、1=なし */
%let framework_pattern=0; /* 枠の形状 0=正方形、1=表示限界の長方形 */

%let anime=c:%mygif%mangekyo.gif; /* 作成するGIFファイルの出力先 */

/* ----- */
/* 万華鏡を作るためのプログラム */

filename anime "&anime";
%let time=round(time()); /* 乱数のシード */
%let pi=constant("pi"); /* 円周率 */
%let rate=(100+71*&framework_pattern)/100; /* 枠の縦と横の比率 */

/* 枠が長方形の時のデータ */
%macro oblong;
  axis1 length=7.7in order=(-172 to 172 by 172) color="&back_color";
  axis2 length=4.5in order=(-101 to 101 by 101) color="&back_color";
%mend;

/* 枠が正方形の時のデータ */
%macro square;
  axis1 length=4.5in order=(-101 to 101 by 101) color="&back_color";
  axis2 length=4.5in order=(-101 to 101 by 101) color="&back_color";
%mend;

/* 枠とタイトルがあるときの設定 */
%macro exist;
  symbol1 i=j;
  title1 h=5 pct f=mincho "動く万華鏡";
%mend;

/* 枠とタイトルがないときの設定 */
%macro noexist;

```

```

symbol1 i=none;
%mend;

%macro make_random_no;
  data dummy_make_seed;      /* 乱数発生シード作成 */
    dummy_seed_first=&time;
    dummy_seed_second=&time+123;
    seed_step=&time+1234;
  run;
%mend make_random_no;

%macro pointing;
  proc datasets lib=work nolist kill; /* 既存のSASテンポラリーデータセットの削除 */
  run ;
  quit ;

  data framework;          /* 枠のデータ */
    length key $ 8 ;
    do i=0 to %eval(&pitch-1) ;
      key=put(i,z8.) ;
      x=-100*&rate+1; y=-101; output;
      x=-100*&rate+1; y= 101; output;
      x= 100*&rate+1; y= 101; output;
      x= 100*&rate+1; y=-101; output;
      x=-100*&rate+1; y=-101; output;
    end;
  run;

  %make_random_no;

  %do order=0 %to %eval(&howmany-1); /* %do(1)の始まり */
    %if &shape=0 %then %make_random_no;

    data make_seed;          /* ランダム軌道時の始点と移動方向 */
      set dummy_make_seed;
      dummy_direction        =ranuni(seed_step+&order)*2*&pi;
      dummy_x_start_position=ranuni(seed_step+&order)*&move_range*&rate;
      dummy_y_start_position=ranuni(seed_step+&order)*&move_range;
    run;

    %do size=0 %to %eval(&pitch-1); /* %do(2)の始まり */
      data base;
        set make_seed;
        drop dummy_seed_first dummy_seed_second dummy_direction
              dummy_x_start_position dummy_y_start_position i;

        array dummy_var {5} dummy_seed_first dummy_seed_second dummy_direction
                          dummy_x_start_position dummy_y_start_position;

        array dummy_retain {5} dummy_retain1-dummy_retain5;

```



```

retain                dummy_retain1-dummy_retain5;

retain dummy 0;

if _n_=1 then do i=1 to 5; dummy_retain{i}=dummy_var{i}; end;

do while(dummy < &line_no*(&divide+1));
  dummy=dummy+1;
  direction=dummy_retain3;
  x_start_position=dummy_retain4;
  y_start_position=dummy_retain5;

  select(&shape_pattern);
    when(0) seed=dummy_retain1+&size;
    when(1) seed=dummy_retain1;
    when(2) do; if &size < round(&pitch/2) then seed=dummy_retain1;
              else seed=dummy_retain2;
            end;
  end;
  output;
end;
run;

data plot_KaleidoScope;
  set base;

  retain segment 0 vertex_no 0 first_random second_random;

  drop change_size angle x_center y_center angle_move angle_circle angle_add
        direction x_start_position y_start_position first_random second_random
        real_radius_id real_move random_no angle_shift;

  if mod(dummy,&divide+1)=1 then do;
    segment=segment+1;
    vertex_no=0;
    first_random =100*ranuni(seed);
    second_random=100*ranuni(seed);
  end;

  vertex_no=vertex_no+1;
  frame_order=&size;
  circle_order=&order;

  select;
    when(&pitch < 10) do; real_radius_id=0;
                        real_move=0;
                        end;
    otherwise do; real_radius_id=&radius;
                 real_move=&move;
                 end;
  end;

```

```

end;

/* 万華鏡の中心点の座標 */

angle_move=2*&pi*&size/&pitch;
angle_circle=2*&pi*&order/&howmany;
angle_add=angle_move+angle_circle;
angle_shift=&shift*angle_circle;

select(real_move);
  when(0) do; x_center=0; y_center=0; end;
  when(1) do; x_center=&move_range*sin(angle_add)*&rate;
             y_center=&move_range*sin(2*angle_add-angle_shift);
             end;
  when(2) do; x_center=&move_range*cos(angle_add)*&rate;
             y_center=&move_range*sin(angle_add-0.5*angle_shift);
             end;
  when(3) do; x_center=&rate*&move_range
             *(abs(mod(1.4*cos(angle_add)+2,2)-1)-1);
             if angle_add < 1/2*&pi OR
                3/2*&pi < angle_add < 5/2*&pi OR
                7/2*&pi < angle_add then x_center=abs(x_center);

             y_center=&move_range
             *(abs(mod(1.4*sin(angle_add-0.5*angle_shift)+2,2)-1)-1);
             if angle_add-0.5*angle_shift < &pi OR
                2*&pi < angle_add-0.5*angle_shift < 3*&pi then
             y_center=abs(y_center);
             end;
  when(4) do; x_center=(abs(mod(abs(&move_speed*&size*cos(direction)
             +&move_range+x_start_position+angle_circle),4*&move_range)
             -2*&move_range)-&move_range)*&rate;
             y_center=abs(mod(abs(&move_speed*&size*sin(direction)
             +&move_range+y_start_position+angle_circle),4*&move_range)
             -2*&move_range)-&move_range;
             end;
end;

/* 万華鏡の半径の係数 */
select;
  when(0 <= real_radius_id <=1)
  do; change_size=(abs(&pitch-mod(real_radius_id*&pitch+4*real_radius_id
             *&size,2*&pitch))+&pitch/100)/&pitch;
             if real_move > 0 then
             change_size=change_size*(100-&move_range)/100;
             end;
  when(real_radius_id=2)
  change_size=min(100*&rate-abs(x_center),100-abs(y_center))/100;
end;

```

```

/* 万華鏡の中心点と万華鏡を構成する座標を結んだ直線の角度 */
angle=2*&pi*vertex_no/&divide+2*&pi*&size/&pitch*&round_speed/360;

if mod(vertex_no,2)=0 then random_no=first_random;
    else random_no=second_random;

/* 万華鏡を構成する線分の座標 */
x=change_size*random_no*cos(angle)+x_center;
y=change_size*random_no*sin(angle)+y_center;
run;

/* 座標データの出力設定 */
data make_KaleidoScope;
    length key $8 function $8 color $8;

    retain count 0 color;
    retain xsys "2" ysys "2" when "a";

    drop bg_dense rb_dense rg_dense count seed_step dummy_retain1-dummy_retain5 seed;

    set plot_KaleidoScope;
    key = put(&size, z8.);

/* 座標点を結ぶ線分の設定 */
if vertex_no=1 then do; if &color_pattern=2 then function="move";
    else function="poly";
        color="&back_color";
        size=&line_width;
    end;
else do;
    select(&color_pattern);
        when(0) color="&line_color";
        otherwise
            if (&color_pattern=1 and vertex_no=2) OR &color_pattern=2 then
                do; bg_dense=max(&b_dense,&g_dense);
                    rb_dense=max(&r_dense,&b_dense);
                    rg_dense=max(&r_dense,&g_dense);

                    color_r=put(abs
                        (mod((512-256*bg_dense)*ranuni(seed)+10*&size*&r_id
                            , (512-255*bg_dense))-(256-128*bg_dense)), hex2.);
                    color_g=put(abs
                        (mod((512-256*rb_dense)*ranuni(seed)+10*&size*&g_id
                            , (512-255*rb_dense))-(256-128*rb_dense)), hex2.);
                    color_b=put(abs
                        (mod((512-256*rg_dense)*ranuni(seed)+10*&size*&b_id
                            , (512-255*rg_dense))-(256-128*rg_dense)), hex2.);
                    color="cx" || color_r || color_g || color_b;
                    count=count+1;
                    if &howmany*count>255 then do;

```

```

        put //
        "***** 警告 *****"/
        "色の設定は256色までですが、現在この限界値を超えました。"/
        "色の設定に関わるマクロ変数の値を変えてください。"/
        "howmany, line_no, shape, color_pattern等が該当します。"/
        "*****"//;
    abort; /* プログラムの実行中止 */
end;
end;
end;
size=&line_width;
if &color_pattern=2 then function="draw";
                    else function="polycont";
end;
output;
run;

/* グラフを描くための元データを溜め込む */
proc append base=graphbase data=make_KaleidoScope;
run;
%end; /* %do(2)の終わり */
%end; /* %do(1)の終わり */

/* データをアニメーションの順番に並べ替える */
proc sort data=graphbase;
    by frame_order circle_order segment vertex_no;
run ;

/* GIFアニメーションの設定 */
goptions reset=all cback="&back_color" border;
goptions device=gifanim gsfname=anime gsfmode=replace gepilog="3B"x delay=&frame_time;
options nobyline ;

/* GIFアニメーションの作成 */
proc gplot data=framework;
    by key; /* GIFアニメーションのフレームを変数keyに対応させる */
    plot y*x / annotate=graphbase haxis=axis1 vaxis=axis2;
    %if &framework_pattern=0 %then %square;
                    %else %oblong;
    %if &framework=0 %then %exist;
                    %else %noexist;

run;
quit;

%mend pointing;

%pointing;

```

◆ 日本 SAS ユーザー会世話人会

代表世話人	東京大学	大橋 靖雄
副代表世話人	麒麟ビール株式会社	本川 裕
世話人	成蹊大学	岩崎 学 (2003 年 年次総会チェアマン)
	三菱証券株式会社	青沼 君明
	株式会社 UFJ 銀行	小野 潔
	神戸商科大学	周防 節雄
	株式会社ベルシステム24	西 次男
	持田製薬株式会社	舟喜 光一
	株式会社竹中工務店	八木 章

◆ 日本 SAS ユーザー会事務局

SAS Institute Japan 株式会社内
〒104-0054 東京都中央区勝どき 1-13-1 イヌイビル・カチドキ 8F
TEL : 03-3533-6936 FAX : 03-3533-1613
E-mail : jpnaswg@sas.com
<http://www.sas.com/japan/>

第 22 回 日本 SAS ユーザー会総会および研究発表会 論文集

2003 年 7 月 31 日
発行

初版第 1 刷発行
日本 SAS ユーザー会
SAS Institute Japan 株式会社