Restricted Cubic Spline for Linearity Test & Continuous Variable Control

SAS HEALTH USER GROUP (HUG)
APRIL 11TH, 2014

JIMING FANG, PHD
CARDIOVASCULAR PROGRAM, ICES

Institute for Clinical Evaluative Sciences

Introduction – A Real Study Case at ICES

Compare 1-year mortality between heart failure patients with reduced ejection fraction (EF) versus those with preserved EF

	Cox Model -1	Cox Model-2
Systolic BP (SBP)	Adjusted as dichotomized variable (140+ vs. <140 mmHg)	Adjusted as continuous variable
Adjusted HR (Reduced EF vs. Preserved EF)	1.23 (95%CI: 1.03-1.47) p=0.03	1.13 (95%CI: 0.94-1.36) p=0.18
Conclusion	When adjusted for baseline characteristics, the survival of heart failure patients with preserved EF is slightly better than those with reduced EF	When adjusted for baseline characteristics, the survival of heart failure patients with preserved EF is similar to those with reduced EF

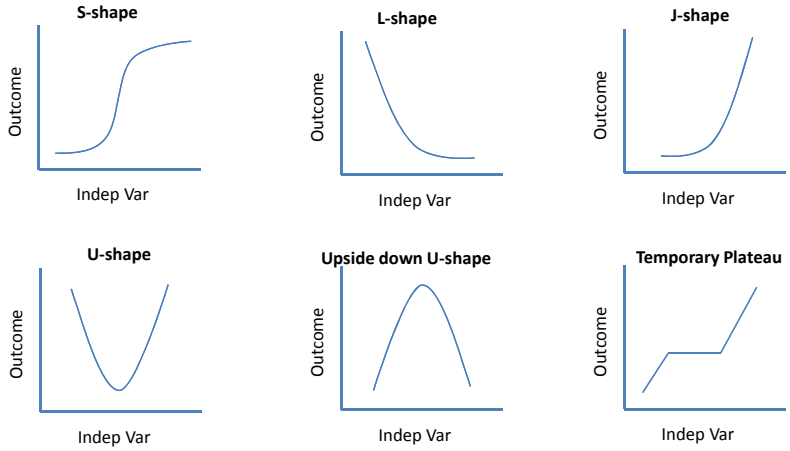
Introduction – Independent variables in multivariable regression

- **Dichotomous variables (e.g., Sex)**
  - 1 vs. 0
- **Nominal variables (e.g., Ethnicity)**
  - Dummy variables
- **Ordinal variables (e.g., Income Quintiles)**
  - Dummy variables
- **Continuous variables (e.g., Age, Weight, BP)**
  - Easy, just add them into model
  - Assume that a unit change anywhere on the scale of the interval variable will have an equal effect on the modeled outcome

Introduction – Linearity Assumption

- **Linear regression model (Outcome: continuous measurement)**
  - an equal size change will have an equal size change to the mean value of the outcome
- **Logistic regression mode (Outcome: event)**
  - an equal size change will have an equal size change to the logit of the outcome
- **Cox model (Outcome: time-to-event)**
  - an equal size change will have an equal size change to the logarithm of the relative hazard

Nonlinear Relationships in Real World



Don't Simply Divide Continuous Variable

- **Results in a step function relationship between the predictor and the dependent variable**
- **Reduce the predictive power of the variable in a predictive model**
- **Lead to more Type-I error**

Altman (1991) British J Cancer, 64: 975
Austin (2004) Statistics in Medicine, 23:1159-78

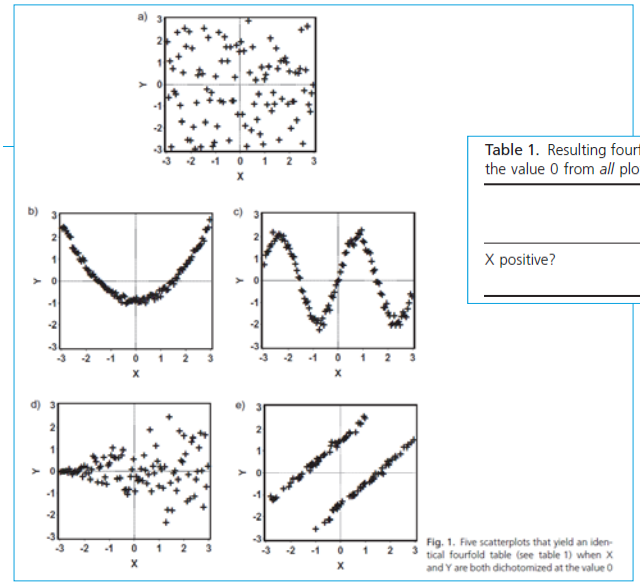


Fig. 1. Five scatterplots that yield an identical fourfold table (see table 1) when X and Y are both dichotomized at the value 0

Table 1. Resulting fourfold table from dichotomizing X and Y at the value 0 from *all* plots in figure 1

		Y positive?	
		Yes	No
X positive?	Yes	25	25
	No	25	25

Kuss (2013) Teaching Statistics, 35:78-79

Linearity Tests in Bivariate Analysis

- **Scatter plot of the outcome and the continuous variable**
  - OK for continuous outcome
  - Not OK for binary outcome or time-to-event outcome
- **Binary outcome or Time-to-Event**
  - First, categorize the continuous variable into multiple dichotomous variables of equal intervals (e.g., age: 21-30, 31-40, 41-50, etc.)
  - Second, compute the % of outcomes in each interval and create 2xn table. Run Proc Freq Trend test to see if it is significant or not.
  - Or enter the categorical variable into the logistic/Cox models. Graph the coefficients to see if there is a straight line (steadily increase or decrease)

Katz (2011) Multivariable Analysis (3rd Ed)

Linearity Tests in Multivariable Model

- **Easy test (in quality)**
  - Plot raw residuals against each independent variable and the estimated value of the outcome
    - If linear, the points will be symmetric above and below a straight line, with roughly equal spread along the line
    - In contrast, if residuals are particularly large at very high and/or low levels of one of the independent variables or of the outcome variable
  - Create multiple dichotomous variable of equal intervals for given continuous variable
    - If linear, the numeric difference between the coefficients of each successive group is approximately equal
- **Complex test (with p-value)**
  - Restricted Cubic Spline (Today's main objective)

Katz (2011) Multivariable Analysis (3rd Ed)

Spline – Concepts

- **Splines enable us to model complex relationships between continuous independent variables and outcomes**
- **Defined to be piecewise polynomials curve, which was constructed by using a different polynomial curve between each two different x-values.**
- **The points at which they are connected are called knots**

Smith (1979) The American Statistician, 33:57-62

Spline – Piecewise polynomials curve

- **Piecewise regression** $y = a_1 + b_1 x \quad for\ x \le c$; $y = a_2 + b_2 x \quad for\ x > c.$
- **Polynomials** $y = a + bx + cx^2 + dx^3 + \ldots$
- **Polynomials may be considered a special case of splines without knots**
- **Two key values for splines**
  - Number of knots
  - Number of degrees

Splines – Knots

- **Default knot locations are placed at the quantiles of the x variable given in the following table**
- **Five knots is sufficient to capture many non-linear pattern**
- **For smaller dataset, it is reasonable to use splines with 3 knots**

Number of knots k	Knot locations expressed in quantiles of the x variable						
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.03	0.183	0.342	0.5	0.658	0.817	0.98

Harrell (2001) Regression Modeling Strategies

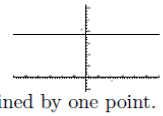
Splines – Degrees

$$y = a + bx + cx^2 + dx^3 + \ldots$$

- **Degree 0**
- **Degree 1**
- **Degree 2**
- **Degree 3**

degree 0: *Constant*, only a is non-zero.

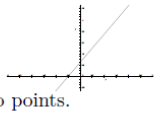
Example: $y = 3$



A constant, uniquely defined by one point.

degree 1: *Linear*, b is highest non-zero coefficient.

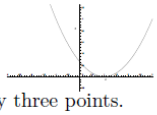
Example: $y = 1 + 2x$



A line, uniquely defined by two points.

degree 2: *Quadratic*, c is highest non-zero coefficient.

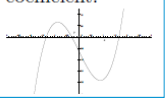
Example: $y = 1 - 2x + x^2$



A parabola, uniquely defined by three points.

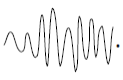
degree 3: *Cubic*, d is highest non-zero coefficient.

Example: $y = -1 - 7/2x + 3/2x^3$



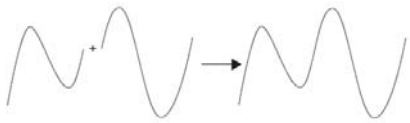
Splines – Cubic

- **Cubic Curve (i.e., degree 3 polynomial)**
- **Most typically chosen for constructing smooth curves in computer graphics, because**
  - it is the lowest degree polynomial that can support an inflection, so we can make interesting curves, and
  - it is very well behaved numerically that means that the curves will usually be smooth, and not jumpy



Splines – Piecewise Cubic Curve

- **The spline curve was constructed by using a different cubic polynomial curve between each knots. The spline will bend around these knots.**
- **In other words, a piecewise cubic curve is made of pieces of different cubic curves glued together. The pieces are so well matched where they are glued that the gluing is not obvious.**



Linearity Test via Restricted Cubic Splines – Piecewise regression

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+, \quad (2.17)$$

where

$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \le 0. \end{cases} \quad (2.18)$$

Linearity Test via Restricted Cubic Splines – Cubic splines

- **Cubic spline function is applied when not all pieces are linear**
- **A weakness of cubic spline is that they may not perform well at the tails (before the first knot and after the last knot)**

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \\ &+ \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3 + \beta_6 (X-c)_+^3 \\ &= X\beta \end{aligned} \tag{2.22}$$

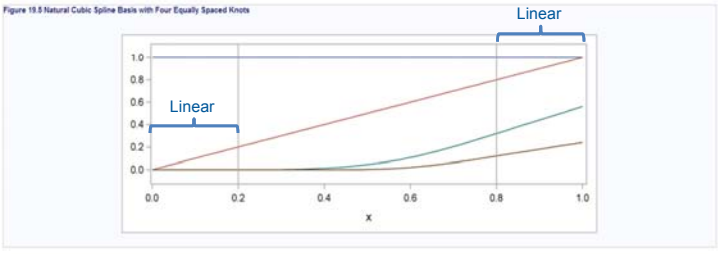
$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \le 0. \end{cases}$$

Linearity Test via Restricted Cubic Splines – Restricted cubic splines

- **Restricted: Constrains the function to be linear beyond the first and last knots (i.e., restricted to be linear in the tails)**

Natural Cubic Spline Basis

Natural cubic splines are cubic splines with the additional restriction that the splines are required to be linear beyond the extreme knots. Some authors use the terminology "restricted cubic splines" in preference to the terminology "natural cubic splines." The space of unrestricted cubic splines on n knots has dimension $n+4$. Imposing the restrictions that the cubic polynomials beyond the first and last knot reduce to linear polynomials reduces the number of degrees of freedom by 4, and so a basis for the natural cubic splines consists of n functions. Starting from the truncated power function basis for the unrestricted cubic splines, you can obtain a reduced basis by imposing linearity constraints. You can find details about this construction in Hastie, Tibshirani, and Friedman (2001). Figure 19.5 shows this natural cubic spline basis defined on $[0, 1]$ with four equally spaced internal knots at 0.2, 0.4, 0.6, and 0.8. Note that this basis consists of four basis functions that are all linear beyond the extreme knots at 0.2 and 0.8.



Linearity Test via Restricted Cubic Splines – Model and SAS Codes

Given x and k knots, a restricted cubic spline can be defined by

$$y = \alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1}$$

where

$$x_1 = x$$

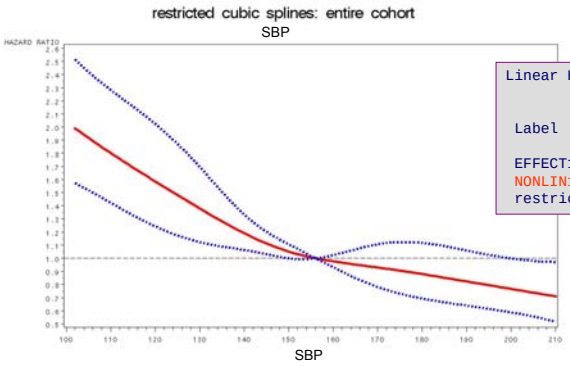
$$x_j = (x - t_{j-1})_+^3 - \frac{(x - t_{k-1})_+^3 (t_k - t_{j-1})}{(t_k - t_{k-1})} + \frac{(x - t_k)_+^3 (t_{k-1} - t_{j-1})}{(t_k - t_{k-1})}$$

$$\text{for } j = 2, \ldots, k-1$$

$$(u)_+ = \begin{cases} u: & u > 0 \\ 0: & u \le 0 \end{cases}$$

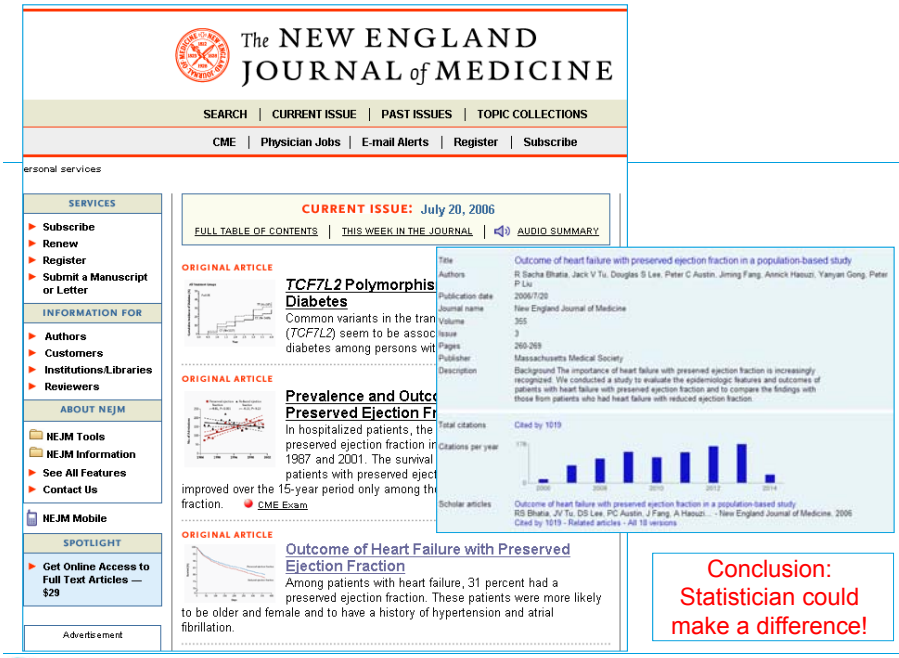
```
Proc PhReg Data=CHF;
  Model surv_1yr*mort_1yr(0)=SBP
                             SBP_1 SBP_2 SBP_3
              CHF_Type Age PVD Cancer ...../RL;

**** spline modelling of fixed covariate SBP;
**** with 5 knots located at;
**** 102 130 150 170 210 (k=5);

SBP_1= ((SBP-102)**3)*(SBP>102)
       -((SBP-170)**3)*(SBP>170)*(210-102)/(210-170)
       +((SBP-210)**3)*(SBP>210)*(170-102)/(210-170);

SBP_2= ((SBP-130)**3)*(SBP>130)
       -((SBP-170)**3)*(SBP>170)*(210-130)/(210-170)
       +((SBP-210)**3)*(SBP>210)*(170-130)/(210-170);

SBP_3= ((SBP-150)**3)*(SBP>150)
       -((SBP-170)**3)*(SBP>170)*(210-150)/(210-170)
       +((SBP-210)**3)*(SBP>210)*(170-150)/(210-170);

*--------- Testing variable: SBP ---------*;
EFFECT1: TEST  SBP, SBP_1, SBP_2, SBP_3;
NONLIN1: TEST      SBP_1, SBP_2, SBP_3;

RUN;
```


Linearity Test via Restricted Cubic Splines – Plot and Wald Chi-square test




```
Linear Hypotheses Testing Results

                    Wald
 Label      Chi-Square      DF    Pr > ChiSq

 EFFECT1       60.0518       4        <.0001
 NONLIN1        2.4782       3        0.4792
 restricted cubic splines: entire cohort
```

NonLin1 test is a test for the null hypothesis that the effect of SBP on survival is linear. P-value of 0.4792 indicated a linear association.



SAS Macros for Linearity Tests

Author	Year	Reference	Country
Heinzl	1996	Statistics in Medicine, 15:2589–2601	Austria
Harrell Howe	2001 2011	Regression Modeling Strategies, page: 20–23 Epidemiology 22:874-875	USA
Spiegelman	2007	Statistics in Medicine, 26:3735–3752	USA
Gregory	2008	Computer methods and Programs in Biomedicine, 92:109–114	Germany
Desquilbet	2010	Statistics in Medicine, 29:1037–1057	France

Comparison between SAS macros for Linearity Tests

Author	SAS Macro Name	Cox model	Logistic model	GLM	GEE	Define reference	Spline	Y-axis of Graph	SAS IML	Adjust other spline
Heinzl (1996)	%rcs	Yes	No	No	No	Middle value between min knot & max knot value of predictor	RCS	Log(HR) HR	Yes	No
Harrell (2001) Howe (2011)	%psplinet %rqspline	Yes	Yes	No	No	Not applicable	RCS	Log(Odds) Log(Hazard)	Not Need	No
Spiegelman (2007)	%lgtphcurv9	Yes	Yes	No	No	Free define	RCS	OR HR	Yes	No
Gregory (2008)	%regspline %regspline_plot	No	Yes	No	No	Yes	B-spline	OR	Yes	No
Desquibet (2010)	%rcs_reg	Yes	Yes	Yes	Yes	Free define Default: Median	RCS	Log(OR) Log(HR)	Yes	Yes

General Comments on RCS

- **To visually check the assumption of linearity, the Y-axis must be Ln(Odds) or Ln(Hazard), instead of OR or HR**
- **Do NOT use RCS to select the cutoff points**
  - The shape of RCS curve can be influenced by the values and numbers of knots
- **RCS has become common statistical method in modeling**

register.

Serum retinol levels were evaluated both as a continuous variable and as a categorical variable, in quintiles. Separate analyses were performed for fractures specifically designated as osteoporotic (i.e., fractures of the hip, pelvis, spine, distal forearm, and proximal humerus).[15] The results were similar whether or not we included the seven cases of fracture due to suspected high-impact trauma, and these cases were therefore retained in the analyses. The nonlinear risk in the highest quintile of the retinol level was determined by inclusion of retinol as a quadratic term in the model together with retinol as a continuous variable. We then estimated the trend in the risk of fracture by a restricted cubic-spline Cox regression analysis[16] with eight "knots" (serum retinol percentiles 1, 5, 20, 40, 60, 80, 95, and 99), which enabled us to investigate extreme retinol values. The results of this analysis are presented as smoothed plots with 95 percent confidence intervals for both the overall risk of fracture and the risk of hip fracture.

(With Reference)

(Without Reference)

STATISTICAL ANALYSIS

To better represent long-term diet and to minimize any effects of within-person variation, we calculated the cumulative average of nut consumption. Because participants may alter dietary patterns after the diagnosis of a major illness, we suspended further updating of all dietary variables when participants reported a diagnosis of stroke, heart disease, angina, or cancer, although follow-up continued until death or the end of the study period.

We used Cox proportional-hazards models to estimate hazard ratios and 95% confidence intervals. Multivariate models were adjusted for known or suspected predictors of death. P values for trend were calculated with the use of the Wald test of a score variable based on the median number of servings of nuts consumed per day for each category of nut consumption. We also used restricted-cubic-spline regression to flexibly model the association.

Jan. 23rd, 2003

Nov. 21st, 2013

If Linearity Assumption Does Not Meets – What to do?

- **Add RCS terms into model**
  - Hard to interpret the results clinically
- **Create multiple dichotomous variables**
  - Advantage: No need to have linearity assumption
  - Limitation: Increase the number of variables in model
- **Create multiple dichotomous variables for primary predictor, and add RCS terms of other continuous predictors**

If Linearity Assumption Does Not Meets – How to select breaking points?

- **Main challenge: to determine the cut-offs**
  - Unfortunately, RCS is not to allow one to select break points
  - In general, it is best to use the cut-offs that reflect a natural, clinically relevant standard
- **Clinically (unequal sample sizes)**
  - SBP/DBP: 130mmHg/90mmHg
  - Serum Hemoglobin
    - Low Men <140 or Women <120
    - Normal Men 140-180 or Women 120-160
    - High Men 180+ or Women 160+
- **Statistically (equal sample sizes)**
  - Quintiles or Tertiles

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Serum Retinol Levels and the Risk of Fracture

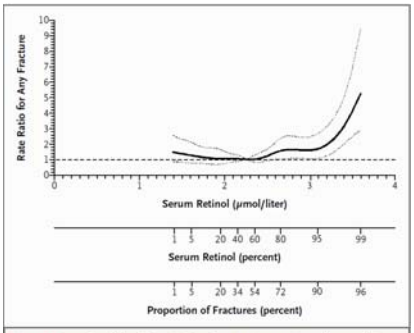


Figure 1. Smoothed Plot of Rate Ratios for Any Fracture According to the Serum Retinol Level.
The rate ratios (solid line) and 95 percent confidence intervals (dotted lines) were estimated by restricted cubic-spline Cox regression analysis, with the median serum retinol level, 2.26 μmol per liter, as the reference value. To convert the values for retinol to micrograms per deciliter, divide by 0.03491.

Table 3. Rate Ratio for Any Fracture and for Hip Fracture, According to the Base-Line Serum Retino

Retinol Quintile	Median Retinol Level	Any Fracture		
		No. of Men	Univariate RR (95% CI)	Multivariate RR (95% CI)†
	μmol/liter			
1 (<1.95 μmol/liter)	1.78	48	1.08 (0.72–1.62)	0.93 (0.62–1.41)
2 (1.95–2.16 μmol/liter)	2.07	33	0.80 (0.51–1.26)	0.78 (0.50–1.23)
3 (2.17–2.36 μmol/liter)‡	2.26	45	1.00	1.00
4 (2.37–2.64 μmol/liter)	2.48	47	0.96 (0.64–1.45)	0.91 (0.60–1.38)
5 (>2.64 μmol/liter)	2.88	68	1.72 (1.18–2.51)	1.64 (1.12–2.41)

Michaëlsson (2003) NEJM, 348:287-94

Options for Dealing with Continuous Variable in Multivariable Regression Model

Procedure	Characteristics	Recommendations
Dichotomization	Simple, easy interpretation	Bad idea
Linear	Simple	Reasonable as a start
Transformations	Log, square root, inverse, exponent, etc.	May provide robust summaries of non-linearity
RCS	Flexible functions with robust behavior at the tails of predictor distribution	Flexible descriptions of non-linearity
More categories	Categories capture prognostic information, better but are not smooth, sensitive to choice of cut-points and hence instable	Primarily for illustration (via percentiles)

Steyerberg (2009) Clinical Prediction Models

Thank You!

Qs & As