Modelling continuous risk variables: Introduction to fractional polynomial regression

Hao Duong[1], Devin Volding[2]

[1]*Centers for Disease Control and Prevention (CDC), U.S. Embassy, Hanoi, Vietnam*
[2]*Houston Methodist Hospital, Houston, Texas, USA*

Editor: Phuc Le, Center for Value-based Care Research, Medicine Institute, Cleveland Clinic, OH

* To whom correspondence should be addressed: Hao Duong. 02 Ngo Quyen, Hoan Kiem, Hanoi. Tel: 04-39352142.
Email: hduong@cdc.gov

Abstract: Linear regression analysisis used to examine the relationship between twocontinuous variableswith the assumption of a linear relationship between these variables. When this assumption is not met, alternative approaches such as data transformation, higher-order polynomial regression, piecewise/spline regression, and fractional polynomial regression are used. Of those, fractional polynomial regression appears to be more flexible and provides abetter fitto the observeddata.

Tóm tắt: Hồi quy tuyến tính được sử dụng để đánh giá mối liên quan giữa hai biến liên tục với điều kiện mối liên quan giữa chúng là một đường thẳng. Khi mối liên quan này không phải là một đường thẳng thì các phương pháp khác được dùng thay thế như là chuyển đổi dữ liệu, hồi quy đa thức bậc cao, hồi quy tuyến tính từng mảnh, hoặc hồi quy đa thức phân đoạn. Trong đó hồi quy đa thức phân đoạn linh động hơn và cho phép mô hình hóa số liệu chính xác hơn.

Keywords: Continuous variables, fractional polynomial, regression.

Introduction

Linear regression analysisis used to examine relationships among continuous variables, specifically the relationship between a dependent variableand one or moreindependent variables. Alternative approaches, including higher-order polynomial regression, piecewise/spline regression and fractional polynomial regression, have been developed to be compatible with examining different forms of associations among variables.

Traditional regression models

Continuous risk variables, used in linear regression models, are typically entered into the modelswith the underlying assumption of a linear relationship between risk variables and outcomes of interest. This assumption, unfortunately, is not always met, and therefore various alternative statistical approaches are used. The most common approach is data transformation (logX, sqrt(X), or 1/X – X is the risk variable of interest) which may solve the issueof non-linearity in some cases but in many cases more flexible approaches are needed. Using higher-order polynomial regression (quadratic, cubic) or piecewise models/splines may improve model fit. The morecomplex modelalmost alwaysfitsthe databetter thantheless complex model, i.e. linear model, but testing is needed to determine whether the improvement in model fit is significant (1).

Models and functions:
Linear model: $\beta_0 + \beta_1 X$ (straight line)
Other models used to improve the model fit:
Quadratic model: $\beta_0 + \beta_1 X + \beta_2 X^2$ (parabola curve)
Cubic model: $\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ (S-shaped curve)
Piecewise/spline model: $\beta_0 + \beta_1 X$ (if $X \in [x_1, x_2]$) $+ \beta_2 X$ (if $X \in [x_2, x_3]$) $+\ldots+ \beta_n X$ (if $X \in [x_{n-1}, x_n]$)
where X is the risk variable of interest, and $x_1<x_2<x_3<...<x_{n-1}<x_n$. Commonly, these knots can be predetermined or based on visual data inspection.

Fractional polynomial (FP) models

Transforming data or using higher-order polynomialmodelsmay provide a significantly better fit than a linear regression modelalone, but these options may notprovide for the best fit to the data. Royston and Altman developed modeling frameworks– fractional polynomial (FP) models that aremore flexible on parameterization and offer a variety of curve shapes (2). These frameworks include transformations that are power functions X^p or $X^{p1} + X^{p2}$ for different values of powers (p, p^1 and p^2), taking from a predefined set S = {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3}.
They are presented as follows:
FP *degree 1* with one power p: FP1 = $\beta_0 + \beta_1 X^p$; when p=0, FP1= $\beta_0 + \beta_1 \ln(X)$
FP *degree 2* with one pair of powers (p_1, p_2):FP2 = $\beta_0 + \beta_1 X^{p1} + \beta_2 X^{p2}$; when p1=p2, FP2 = $\beta_0 + \beta_1 X^{p1} + \beta_2 X^{p2} \ln(X)$
FP1 has 8 models with 8 different power values, and FP2 has 36 different models, including 28 combinations of those 8 values and 8 repeated ones. Table 1 presents FP modelswith degrees 1 and 2. These two sets of models provide a very wide range of curves shapes (Figures 1, 2) and cover many types of continuous functions encountered in the health sciences and elsewhere (2).

Model selection strategy

Simple function, i.e. linear function is preferred over the complex functionexcept when fit of the best-fitting alternative model (power $p \neq 1$) is significantly improved compared to that of the linear model (power p=1)(2,3). The best fitting FP degree 1 model is the one with the smallest deviance from among the 8 models with different power values, taking fromthe set {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3} (Table 1). The best fitting FP degree 2 model is the one with the smallest deviance from among the 36 models with different pairs of powers, taking from the set {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3} (Table 1).
Model selection stepsare as follows (2, 3):

Step 1: Test whether there is any effect of X – risk variable of interest on the outcome by comparing the fit of the best fitting FP2 withthat of the null model, using the chi-square test with 4 degrees of freedom. If the test is not significant, stop and conclude that there is no effect of X on the outcome. Otherwise continue.

Step 2: Test whether the relationshipbetween X and the outcome is nonlinear by comparing the fit of the best fitting FP2 with that of the linear model, using the chi-square test with 3 degrees of freedom.If the test is not significant, stop and conclude that there is straight relationship between X and the outcome. Otherwise continue.

Step 3: Test whether the FP 2 fits better than the FP 1 using chi-square test with 2 degrees of freedom. If the test is not significant, the final model is FP1, otherwise the final model is FP2. Table 2 illustrates one example of model selection.

Conclusion

FP is computationally intensive and complicated when modelling multiple continuous variables. Interpreting FP results is also more difficult than other models. In contrast, FP allows more flexibility for modelling and potentially provides a better fit to the observed data.

Table 1. Fractional polynomial (FP) models

	FP degree 1 with 8 different powers, taking from {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3}	
1.	FP1: Power (-2)	$\beta_0 + \beta_1 (1/X^2)$
2.	FP1: Power (-1)	$\beta_0 + \beta_1 (1/X)$
3.	FP1: Power (-0.5)	$\beta_0 + \beta_1 (1/\text{sqrt}(X))$
4.	FP1: Power (0)	$\beta_0 + \beta_1 (\ln X)$
5.	FP1:Power (0.5)	$\beta_0 + \beta_1 (\text{sqrt}(X))$
6.	FP1:Power (1)	$\beta_0 + \beta_1 X$
7.	FP1:Power (2)	$\beta_0 + \beta_1 X^2$
8.	FP1:Power (3)	$\beta_0 + \beta_1 X^3$
	FP *degree 2 with 36 different pairs of powers, taking from {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3}*	
1.	FP2:Powers (-2, -2)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (1/X^2) (\ln X)$
2.	FP2:Powers (-1, -1)	$\beta_0 + \beta_1 (1/X) + \beta_2 (1/X) (\ln X)$
3.	FP2:Powers (-0.5, -0.5)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (1/\text{sqrt}(X)) (\ln X)$
4.	FP2:Powers (0, 0)	$\beta_0 + \beta_1 \ln X + \beta_2 (\ln X) (\ln X)$
5.	FP2:Powers (0.5, 0.5)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (1/\text{sqrt}(X)) (\ln X)$
6.	FP2:Powers (1, 1)	$\beta_0 + \beta_1 X + \beta_2 X (\ln X)$
7.	FP2:Powers (2, 2)	$\beta_0 + \beta_1 X^2 + \beta_2 X^2 (\ln X)$
8.	FP2: Powers (3, 3)	$\beta_0 + \beta_1 X^3 + \beta_2 X^3 (\ln X)$
9.	FP2:Powers (-2, -1)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (1/X)$
10.	FP2:Powers (-2, -0.5)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (1/\text{sqrt}(X))$
11.	FP2:Powers (-2, 0)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (\ln X)$
12.	FP2:Powers (-2, 0.5)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (\text{sqrt}(X))$
13.	FP2:Powers (-2, 1)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (X)$
14.	FP2:Powers (-2, 2)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (X^2)$
15.	FP2:Powers (-2, 3)	$\beta_0 + \beta_1 (1/X^2) + \beta_2 (X^3)$
16.	FP2:Powers (-1, -0.5)	$\beta_0 + \beta_1 (1/X) + \beta_2 (1/\text{sqrt}(X))$
17.	FP2:Powers (-1, 0)	$\beta_0 + \beta_1 (1/X) + \beta_2 (\ln X)$
18.	FP2:Powers (-1, 0.5)	$\beta_0 + \beta_1 (1/X) + \beta_2 (\text{sqrt}(X))$
19.	FP2:Powers (-1, 1)	$\beta_0 + \beta_1 (1/X) + \beta_2 (X)$
20.	FP2:Powers (-1, 2)	$\beta_0 + \beta_1 (1/X) + \beta_2 (X^2)$
21.	FP2:Powers (-1, 3)	$\beta_0 + \beta_1 (1/X) + \beta_2 (X^3)$
22.	FP2:Powers (-0.5, 0)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (\ln X)$
23.	FP2:Powers (-0.5, 0.5)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (\text{sqrt}(X))$
24.	FP2:Powers (-0.5, 1)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (X)$
25.	FP2:Powers (-0.5, 2)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (X^2)$

26.	FP2:Powers (-0.5, 3)	$\beta_0 + \beta_1 (1/\text{sqrt}(X)) + \beta_2 (X^3)$
27.	FP2:Powers (0, 0.5)	$\beta_0 + \beta_1 (\ln X) + \beta_2 (\text{sqrt}(X))$
28.	FP2:Powers (0, 1)	$\beta_0 + \beta_1 (\ln X) + \beta_2 (X)$
29.	FP2:Powers (0, 2)	$\beta_0 + \beta_1 (\text{sqrt}(X)) + \beta_2 (X^2)$
30.	FP2:Powers (0, 3)	$\beta_0 + \beta_1 (\text{sqrt}(X)) + \beta_2 (X^3)$
31.	FP2:Powers (0.5, 1)	$\beta_0 + \beta_1 (\text{sqrt}(X)) + \beta_2 (X)$
32.	FP2:Powers (0.5, 2)	$\beta_0 + \beta_1 (\text{sqrt}(X)) + \beta_2 (X^2)$
33.	FP2:Powers (0.5, 3)	$\beta_0 + \beta_1 (\text{sqrt}(X)) + \beta_2 (X^3)$
34.	FP2:Powers (1, 2)	$\beta_0 + \beta_1 (X) + \beta_2 (X^2)$
35.	FP2:Powers (1, 3)	$\beta_0 + \beta_1 (X) + \beta_2 (X^3)$
36.	FP2:Powers (2, 3)	$\beta_0 + \beta_1 (X^2) + \beta_2 (X^3)$

Table 2. An example of model selection

Model	Deviance (D)	Power	Equation	Comparison	D Difference (χ^2 distribution)	*P-value*
FP2*	285882	1, 1	$\beta_0 + \beta_1 X + \beta_2 X(\ln X)$	Step 1: FP2 vs null	21306	<0.05
FP1**	287083	0.5	$\beta_0 + \beta_1(\text{sqrt}(X))$	Step 2: FP2 vs linear	1411	<0.05
Linear	287293	1	$\beta_0 + \beta_1 X$	Step 3: FP2 vs FP1	21306	<0.05
Null	307188	-	β_0			

*FP1 model (best fitting) with the smallest deviance among 8 FP1 models
**FP2 model (best fitting) with the smallest deviance among 36 FP2 models

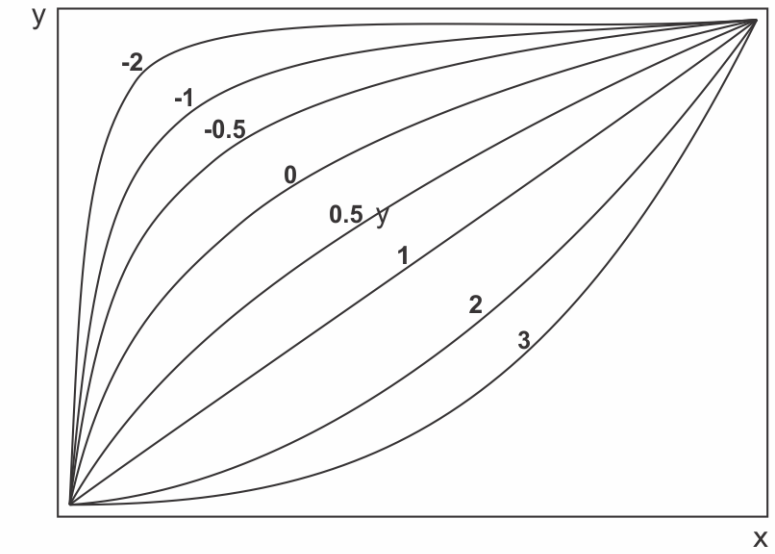

Figure 1.Schematic diagram of eight FP1 curve shapes. Numbers indicate the power p
Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Patrick Royston and Willi Sauerbrei.

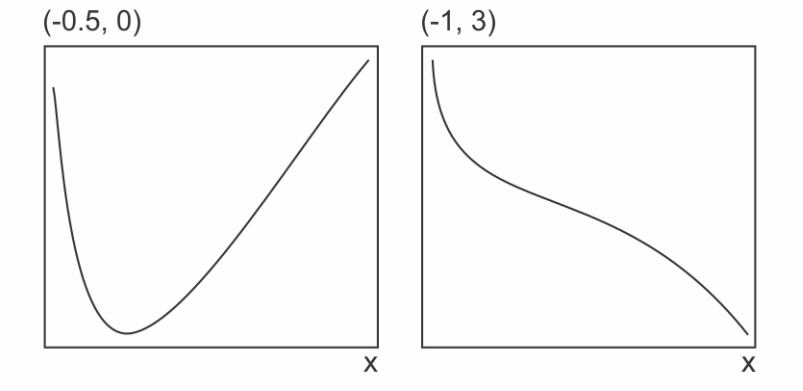

Figure 2. Schematic examples of FP2 curves
Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Patrick Royston and Willi Sauerbrei

References
1. Anonymous (2014) Likelihood-ratio test after estimation (http://www.stata.com/manuals13/rlrtest.pdf).
2. Patrick Royston and Willi Sauerbrei (2008) Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables (John Wiley & Sons).
3. Royston,P; Ambler,G; Sauerbrei,W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *Int. J. Epidemiol*.28(5): 964-974.

About the author: Dr. Hao Duong received her MD in Hue Medical School, and DrPH from theUniversity of Texas Health Science CenteratHouston, School of Public Health. She pursued her postdoctoral research in theDepartmentof Epidemiologyinvestigating risk factorsassociatedwithbirth defects, and currently works as a statistician at the Centers for Centers for Disease Control and Prevention (CDC), U.S. Embassy, Hanoi, Vietnam.