The Stata Journal (2010)
10, Number 2, pp. 297–302

Review of Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables, by Royston and Sauerbrei

William D. Dupont
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN
william.dupont@vanderbilt.edu

Abstract. This article reviews *Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables*, by Patrick Royston and Willi Sauerbrei.

Keywords: gn0050, applied statistics, nonlinear regression, fractional polynomials

1 Introduction

Royston and Sauerbrei (2008) provide an excellent introduction to building nonlinear regression models. Their preferred approach is to use fractional polynomial models, a technique that they have largely developed and on which they are undisputed authorities (Royston and Altman 1994; Sauerbrei and Royston 1999). This technique is applicable to models in which the response variable is continuous or dichotomous, to survival models, and to any generalized linear model.

The text is full of practical examples that Royston and Sauerbrei use to illustrate their model-building approach. They make extensive use of smoothed residual plots to evaluate their models and to guide model selection. Their approach is suitable for epidemiological studies in which the number of observations is at least an order of magnitude larger than the number of model covariates. For such data, the problems of multiple comparisons and the overfitting of models are not an overwhelming concern. The problem of model-building for genomic or other studies in which the number of covariates greatly exceeds the number of study subjects is not addressed in the text.

Fractional polynomial models are a subset of generalized linear models in which various powers of the covariates of interest are entered into the linear predictor. A fractional polynomial regression model of order 1 (FP1) is one in which the linear predictor takes the form

$$\beta_0 + \beta_1 x^p$$

where p takes one of the values in $S = \{-2,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 2,\ 3\}$ and $x > 0$. For example, the linear regression model

$$y = \beta_0 + \beta_1/\sqrt{x}$$

is an example of an FP1 model, as is the logistic regression model

$$\text{logit}(\pi[x]) = \beta_0 + \beta_1 x^2$$

A fractional polynomial model of order 2 (FP2) is one in which the linear predictor takes the form

$$\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$$

or

$$\beta_0 + \beta_1 x^p + \beta_2 x^p \log x$$

for p_1, p_2, and p in S and $x > 0$. This definition generalizes in a natural way: in an mth order model (FPm), the linear predictor equals

$$\beta_0 + \sum\nolimits_{j=1}^{m} \beta_j h_j\,(x)$$

where

$$h_j\,(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1} \\ h_{j-1}\,(x) \log x, & p_j = p_{j-1} \end{cases}$$

and $p_1, p_2, \ldots, p_m$ are chosen from S.

The fact that these models all involve linear combinations of the model parameters means that the extensive theory of generalized linear models applies and that many sophisticated programs for building and evaluating such models are already available.

2 Contents

Royston and Sauerbrei's text (2008) starts with an introductory chapter that explains the need for robust ways to fit nonlinear regression models. The authors consider polynomial regression models and find them wanting for many situations. Then they introduce fractional polynomial models and illustrate the use of smoothed residual plots to evaluate model fit. They recommend using the simplest model that fits the data well. Royston and Sauerbrei also illustrate the application of the fractional polynomial approach to Cox proportional hazards regression analysis. They contrast models using age at entry as a continuous covariate with two models with categorical age intervals and with FP1 and FP2 models. They make a convincing argument in favor of the optimal FP2 model over these other alternatives.

Royston and Sauerbrei briefly discuss other modeling approaches, and they make an important distinction between global influence models, such as fractional polynomial models, and local influence models, such as those using restricted cubic splines. They

discuss different types of residuals along with those residuals' roles in guiding model selection.

Chapter 2 introduces the authors' approach to multivariable model selection. Their overall preference is to use backward elimination based on repeated significance tests. They stress the importance of assessing the stability of the selected model by bootstrapping. They distinguish between the modeling goals of prediction and explanation. They place considerable emphasis on reducing the often daunting complexity of observational data and finding comparatively simple models that identify prognostically and diagnostically important variables. Royston and Sauerbrei discuss the pros and cons of different stepwise approaches to model selection, and they describe Akaike's and the Bayesian information criteria (AIC and BIC, respectively). They also discuss shrinkage methods to reduce the effects of selection bias.

Chapter 3 details how to handle categorical and continuous predictors. It contains good advice that is applicable to any exploratory multivariable regression analysis. The multiple-comparisons problems associated with choosing an optimal cutpoint for a continuous covariate are nicely illustrated. Royston and Sauerbrei provide an interesting discussion of the pros and cons of local-influence models, such as lowess regression or cubic splines, and global models, such as those using fractional polynomials.

Chapters 4 and 5 describe in detail the use of fractional polynomials for one variable. Royston and Sauerbrei give the shapes of FP1 and FP2 curves along with their justification of the set of powers, S, that they consider. They explain both naïve and bootstrapped confidence intervals for the linear predictor; the latter adjust for overfitting because of the model-selection algorithm. They discuss methods of graphical and tabular presentation of results from fractional polynomial models along with a worked example on the relationship between systolic blood pressure and all-cause mortality. The authors also describe transformations to improve the robustness of fractional polynomial models.

Chapter 6 introduces multivariable model-building with fractional polynomials—what Royston and Sauerbrei describe as the heart of their text. They present their algorithm for selecting multivariable models, which they illustrate with an example. In essence, they use backward elimination while allowing significant covariates to be fit with an optimal fractional polynomial model. They use functional plots to describe the effect of a covariate on the response variable adjusted for other variables in the model. They consider graphical analysis of residuals from multivariable models. And they use an R^2-like statistic to evaluate the contribution of individual variables.

Chapter 7 describes adding interaction terms to multivariable fractional polynomial models. Royston and Sauerbrei urge the use of graphical checks, sensitivity, and stability analyses as well as a cautious interpretation of the results of such models. Chapter 8 describes the use of bootstrap analyses to assess the stability of complex models.

Chapter 9 compares multivariable fractional polynomial models with spline models. Restricted cubic spline models are a major alternative to fractional polynomial models. They require the specification of three or more knots and fit curves that are cubic

polynomials between adjacent knots, are straight lines before the first and after the last knot, and have the property that the splines and their first and second derivatives are continuous at all knots. This latter property makes restricted cubic splines fairly insensitive to the precise location of their knots. Royston and Sauerbrei present an algorithm for fitting multivariate models with restricted cubic splines. They analyze several datasets with this algorithm and find that it gives models that are roughly comparable with those obtained using multivariable fractional polynomial models.

Chapter 10 provides further guidance on fitting multivariable fractional polynomial models. Chapter 11 describes hazard regression models with time-varying hazard ratios and other topics. Chapter 12, an epilogue, summarizes Royston and Sauerbrei's major recommendations as to how to build useful multivariable models with fractional polynomials.

3 Strengths and weaknesses

This book's greatest strength is its lucid writing style and practical guidance on how to build complex multivariable models. It contains many examples with models that are extensively evaluated using smoothed residual plots and partial predictor plots. I was also intrigued by Royston and Sauerbrei's approach to automated model fitting. Many statisticians have a negative attitude toward such methods because of the risk of overfitting and because of multiple-comparisons concerns (Harrell 2001). I found Royston and Sauerbrei's emphasis on bootstrap analyses to assess model stability to be reassuring. Software to implement their methods is available in Stata, either as part of Stata 11 or as user-contributed programs that can be downloaded over the Internet.

Weaknesses are few. Given that restricted cubic splines are, perhaps, the major competitor to fractional polynomial models, I would have preferred that the authors give a more thorough evaluation of this approach, particularly in regard to fitting univariate models.

4 Fractional polynomials versus restricted cubic splines

I must start this section with a disclaimer: I have been an advocate of restricted cubic splines for several years (Dupont 2009), while my knowledge of fractional polynomial models was limited prior to reading this book. This perhaps gives me a bias in favor of the former over the latter technique.

An argument that the authors make in favor of fractional polynomial models is their simplicity. FP1 models are simpler than restricted cubic spline models, and for this reason, I would recommend an FP1 model whenever it fits the data well. For FP2 models, I am not really sure. My best guess is that most statisticians, and virtually all medical scientists, will lack a visual image of what, say, a linear predictor of the form $\beta_0 + \beta_1 x^2 + \beta_2/\sqrt{x}$ looks like without drawing it. A restricted cubic spline with three knots takes the form $\beta_0 + \beta_1 x + \beta_2 f_2(x)$. Now I must admit the function $f_2(x)$ is neither

pretty nor edifying, but it is readily calculated by any computer. Just as some sausages are best appreciated by eating them without too much knowledge of their contents, a restricted cubic spline can be best understood by drawing it, without paying too much attention to the form of $f_2(x)$. Thus I would argue that FP2 models and three-knot restricted cubic spline models have similar complexity. For data that can be fit well by an FP2 model, my sense is that the two approaches give roughly comparable results. For more complicated data, I believe that restricted cubic splines may have an edge. In particular, my experiments fitting restricted cubic splines to bimodal data have worked well, while fractional polynomial models have either missed the bimodal nature of the data or have provided poor fits to very high or very low values of x.

Another advantage of restricted cubic splines is that the linear model is always nested within more complex models. This means that it is always possible to conduct a Wald or likelihood-ratio test of whether the linear predictor is, in fact, linear in x.

5 Conclusions

This is a very well-written book that provides a thoughtful approach to fitting nonlinear models. The authors are very experienced biostatisticians who have worked extensively in observational and experimental medical science. They make a convincing argument that fractional polynomials can be a valuable tool for building such models in many situations. Their many examples and excellent illustrations make their book accessible to a broad audience within statistical science. Their software will make this book particularly useful to the Stata community. I highly recommend this text.

6 References

Dupont, W. D. 2009. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. 2nd ed. Cambridge: Cambridge University Press.

Harrell Jr., F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Royston, P., and D. G. Altman. 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics* 43: 429–467.

Royston, P., and W. Sauerbrei. 2008. *Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, UK: Wiley.

Sauerbrei, W., and P. Royston. 1999. Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 162: 71–94.

About the author

William D. Dupont is a professor of biostatistics and preventive medicine at Vanderbilt University School of Medicine. His interests include the epidemiology of benign breast disease, power and sample-size calculations, statistical graphics, and teaching intermediate-level biostatistics to physician scientists. He is the author of *Statistical Modeling for Biomedical Researchers* (Cambridge University Press, 2009), which uses Stata to teach biostatistics to this audience.