Published: September 2011

Use of fractional polynomials in medical research

Edmore Marinda - Senior Biostatistician, University of the Witwatersrand, Medical School, Johannesburg

Statistical models which are data driven and analyst derived are fundamentally different from models arising from physical sciences that 'explain' natural phenomena using mathematical models.

Most multivariable models in clinical and epidemiology research consider predictor variables as linear terms or as dummy variables after categorization of continuous variables. Clinically it may be desirable to classify patients into different prognosis groups e.g. poor, moderate and good, or diagnosis groups e.g. benign or malignant tumour. Categorization of continuous variables assumes homogeneity of the trait under consideration within each specified category. This may however be unrealistic especially when few categories are used. Although individuals close to but on opposite sides of the cutpoint are expected to be very similar, categorizing assumes they have very different outcomes. It is unlikely that there will be consensus on the choice of cutpoints when creating categorical groups (1). Distributional measures such as median, upper or lower quartiles, and rounded cutpoints as is usually done in categorized age groups (e.g. 20 - 24,25 -29,…) are often used. Categorization may result in overparameterized models and there is usually loss of efficiency (2). Important relevant predictor variables are sometimes missed in prognostic or diagnostic models because the true functional form of a predictor variable may be non-linear. Medical knowledge may dictate that the relationship between an outcome and a predictor variable is monotonic or there is some levelling off (asymptote) at high or lower values of the predictor. Categorizing confounding variables may result in residual confounding; where the bias due to confounding is not substantially removed. The percentage of bias removed by categorizing continuous confounders under certain distributional assumptions and for monotonic relations has been previous estimated at 64%, 79%, 86%, 90%, and 92% for 2, 3, 4, 5 and 6 cut-off categories respectively (3).

Investigate functional forms of continuous predictor variables

It may be important to differentiate between predictors of main interest and confounders depending on the aim of the study. Fractional polynomials (FP) have been proposed in epidemiological studies to investigate functional forms of continuous predictor variables (4). Final FP models usually have fewer parameters compared to step function models especially when many confounders are considered.

The general form of a FP model is

$$\eta_i(X;\boldsymbol{\beta};\boldsymbol{p}) = \sum_{j=0}^{d} \beta_j g_j(X)$$

for $j = 0$ to d, $g_j(X)$ is X^{pj} *for* $p_j \neq p_{j-1}$ and $g_{j-1}(X)$ is $X^p ln(X)$ for $p_j = p_{j-1}$ and X^p denote the Box Tidwell transformation defined as X^p if $p_j \neq p_{j-1}$ and $ln(X)$ if $p_j = 0$ [4]. The coefficients and powers of the model are contained in the vectors $\boldsymbol{\beta} = (\beta_1, \ldots,\beta_d)$ and $\boldsymbol{p} = (p_1, \ldots,p_d)$ such that $p_1 < \ldots < p_d$. The power list which is usually restricted to a determined set of integers and non-integers: (-3, -2, -1,-0.5,0,0.5,1,2,3) includes the reciprocal, logarithm, square root, square and repeated-powers transformations (4). $\boldsymbol{\eta}$ is an appropriate link function. A wide range of regression commands that cover a number of link functions are available in widely used statistical software such as SAS, STATA and R. These include linear, logistic, cox-regression, conditional logistic, generalized linear models, ordinal logistic, ordinal probit, poisson, probit, quintile regression, parametric regression and generalized estimating equations (5). A FP model of degree d has $2d$ degrees of freedom (DF) (excluding β_0); 1 DF for each β and 1 degree for each power. The best first-degree FP for X is that with the smallest deviance (minus twice the maximum log-likelihood) among models with one predictor variables. Similarly the best second-degree FP is the model with the smallest deviance among those with all possible pairs of powers from the power list. The second degree FP with minimal deviance is preferred at the α % level to the best first-degree FP if the deviance difference exceeds the (100-α) percentile of χ^2 with 2 DF. Otherwise, the first-degree FP is preferred to a linear term model if the corresponding deviance difference exceeds the (100 – α) percentile of χ^2 on 1 DF.

Applications of FP modelling have been demonstrated in a number of studies (5-7). FP provide solutions to serious problems associated with arbitrary categorization. These models provide sufficient power to detect strong non-linearities which should be accommodated in clinical and epidemiologic research. FP models are also relatively simple, a requisite for statistically stable models. Second or lower degree models reduce the likelihood of overfitting. Some biological relationships require that functions are monotonic or that there is an asymptote. This type of medical knowledge ought to be incorporated in the statistical analysis of these studies. Incorporating this knowledge can easily be achieved by using 1st order FP (simple power and log-transformations), or applying transformations such as the negative exponential transformation before applying FPs as described by Sabuerbrei et. al. (8).

It is important to extract as much information as possible from predictors and responses, while trying to avoid overfitting. Retaining continuous variables in this form in models is essential to avoid loss of information. Predictors that are not significant in univariate analysis should not be excluded before their functional form with the outcome and other important variables in models are fully explored (9). It is preferable to keep variables continuous rather than categorize them since much more predictive information is retained. Royston et. al. believe that dichotomization of continuous data is unnecessary for statistical analysis and in particular should not be applied to explanatory variables in regression models (10).

Control for continuous confounders

In order to appropriately control for confounding effects, the functional form of confounders need to be investigated. FP are preferred to step functions (categorizing) because they better control for confounder effect within strata as well as across strata. Greenland advises that epidemiological analysis of dose response and trends as well as methods for controlling of continuous confounders should be expanded beyond simple categorical and linear (single coefficient) approaches to include flexible curves that make use of intra-category information (11). In an editorial Weinberg stated that approaches based on FP or regression splines merit a greater role in epidemiology, and should have a lasting influence on epidemiological practice (12). Given the amount of 'noise' which typically obscures risk relationships in epidemiology, 1st and 2nd order FP will provide sufficiently accurate approximations to unknown realities for most purposes (11). The determination of functional form is particularly important in studies where a variable may have a dual role as a confounder and as a risk factor.

Fitted multivariable models should be as simple as possible. Thus for continuous variables, the starting point should be to consider linear models. FP models (preferably of degree 1 or 2) should be considered if there is sufficient evidence of non-linearity within the data or when there is prior medical knowledge suggesting non-linear functional forms. Care should be taken when examining the shape and location of FP curves since they can strongly be influenced by one or a few data points. In particular fitted values for a point can be strongly influenced by data that are far away on the graph. It is important to use accurate models developed by using appropriate data sets in developing clinical prognostic and diagnostic indices. The model should easily be communicated mathematically, fit the data well, be parsimonious and consistent with medical knowledge. Bootstrap replication can be used to investigate the stability of the functional form of selected models.

Alternatives to FP models include local regression models such as splines and kernel methods. Generalized Additive Models (GAM) have been used to check that important features of data are not missed by parametric models such as FP models (8). Royston (2000) proposes getting a sense of the functional form of predictor variables on outcome variables using non-paramatric methods before fitting parametric models whose fitted values agree adequately with those from the non-parametric models (13). These non-parametric models are usually flexible and their confidence intervals are generally wider and probably more realistic than those of parametric models. They may be used in exploratory data analysis and in helping one to select appropriate parametric models (4). It is however not easy to report the mathematical models for these non-parametric curves because they are very complex, and reporting of results is by extensive tabulation and graphs. Data dependence on the final model is more marked in non-parametric models than for parametric models.

As in all multivariable modelling, issues such as missing data have to be considered. Missing data are usually directly or indirectly related to disease characteristics, including the outcome under study. Thus exclusion of all individuals with a missing value leads not only to loss of statistical power of the model but often to incorrect estimates of the predictive power of the model and specified predictors (14).

Edmore Marinda, Senior Biostatistician, University of the Witwatersrand, Medical School, Johannesburg. Areas of interest: Non linear regression models in HIV research, survival analysis, and nutritional epidemiology. *Edmore.Marinda@wits.ac.za*

References:

1. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. J Natl Cancer Inst. 1994;86(11):829-835.
2. Lagakos SW. Effects of Mismodelling and Mismeasuring explanatory variables on tests of their

association with a response variable. Stat Med. 1988;7:257-274.
3. Cochran WG. The effectiveness of adjustment by subclassication in removing bias in observational studies. Biometrics. 1968;24:295-313.
4. Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. J R Stat Soc. Series C (Appl Stat). 1994;43(3):429-467.
5. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. Computational Statistics & Data Analysis. 2006;50(12):3464-3485.
6. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol. 1999;28(5):964-974.
7. Gray L, Cortina-Borja M, Newell ML. Modelling HIV-RNA viral load in vertically infected children. Stat Med. 2004;23:769-781.
8. Sauerbrei W, Royston P. Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials. J R Stat Soc. Series A (Stat Soc). 1999;162(1):71-94.
9. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-387.
10. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-141.
11. Greenland S. Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. Epidemiology. 1995;6(4):356-365.
12. Weinberg C. How bad is categorization? (Editorial). Epidemiology. 1995;6(4):345-347.
13. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med. 2000;19(14):1831-1847.
14. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-1091.