Assessing the suitability of fractional polynomial methods in health services research

A perspective on the categorisation epidemic

ABSTRACT

Objective: To show how fractional polynomial methods can usefully replace the practice of categorising data in epidemiology and health services research.

Methods: A health service setting is used to illustrate a structured and transparent way of representing non-linear data without arbitrary grouping.

Results: When age is a regressor its effects on an outcome will be interpreted differently depending upon the placing of cutpoints or the use of a polynomial transformation.

Conclusions: Although common practice categorisation comes at a cost. Information is lost, accuracy and statistical power reduced, leading to spurious statistical interpretation of the data. The fractional polynomial method is widely supported by statistical software programs, and deserves greater attention and use.

INTRODUCTION

In health services research continuous variables such as, for example, age, body mass index or blood pressure, are commonly described and analysed in categories or bins, with cutpoints at the boundaries. Categories are routinely used in regression models where they are widely perceived to simplify the statistical analysis [1]. An obvious benefit is ease of interpretation in that the effect of each category within a variable x can be expressed relative to a reference category.

Categorisation has the perceived advantage of achieving robust results when tests of linearity between the continuous covariate and the outcome are not met. However, the stepped function associated with categories is often a poor approximation of a continuous data distribution and the resulting loss of efficiency can be severe [2-6]. Unless the mean of the data within each category corresponds with the mean of that bin's cutpoints, the binned representation will be in error [7-9].

Percentiles are often used to split the data into equal-sized groups, but they do not always represent the data accurately. For example, when the distribution is heavily skewed, percentiles may represent a disproportionate spread of values [5, 10, 11]. These problems also arise when the median is used as a cutpoint to dichotomise continuous data [12].

Compared with a model that retains a continuous scale of measurement, the grouping of data reduces effect size and statistical power. While this situation can be improved by investigators choosing cutpoints that exhibit “convincing” effects, the practice of setting so-called “optimal” cutpoints in order to show statistical significance is inconsistent with best practice research [5].

While the aim of a regression analysis is to determine broad associations between independent and dependent variables, at the detailed level, the binning of continuous predictors implies that observations close to but on opposite sides of each cutpoint, are somehow different. For example, age cutpoints are often set at five or ten-year intervals when there is no reason to expect that a step-change occurs at these cutpoints.

Discrete representation of continuous data at some level is inevitable. However it is important that options to maximise data precision and minimise the likelihood of errors are selected at the survey design stage. For example, if a questionnaire requires respondents to select from pre-set categories that refer to continuous measures such as age, height, weight, annual income etc, not only is precision lost, but the possibility of misclassification error is increased [9].

There are of course many areas in which data cannot be easily described on a continuous scale and a categorical scale is apposite. For example, validated psychometric instruments require responses within categories, and survey answers are often best interpreted through frequency counts of binned data. While categories can be a sensible and practical alternative in some instances, this paper argues against the *unnecessary* use of categories and shows how, using familiar statistical software, the regression of continuous data using fractional polynomials improves accuracy and precision.

The epidemiology literature includes some published studies that use fractional polynomials, although they are the exception rather than the rule [9, 13-17]. Many researchers and practitioners are unaware of their value and ease of application believing that polynomials are mathematically complex and impractical. While the examples discussed here are straightforward, more complex applications can be found in the statistical and epidemiological literature [2, 3, 5, 9]. The aim of this paper is to show how fractional polynomial methods can usefully replace the practice of categorising data in epidemiology and health services research.

METHODS

The examples given here refer to a cohort study that examined equity and access in an Australian hospital outpatient cardiac rehabilitation (CR) program [17-19]. The population comprised only those who were eligible (by hospital discharge diagnosis) for invitation to the program. Of interest to managers was the extent to which patients' age was a predictor of invitation. Logistic regression was used to analyse the statistical association between age (in years) and the binary outcome, "invitation" [19].

Polynomials

A general polynomial is a function of the form: $y = a_0 + a_1 x + a_2 x^2 + \cdots a_n x^n$. For regression purposes, Royston and Altman [11] proposed a constrained generalization which they called "fractional polynomials". Fractional polynomials take the form: $y = a_0 + a_1 x^{p_1} + a_2 x^{p_2} + \cdots a_n x^{p_n}$. The power p is chosen from a fixed set of possibilities (-2; -1; -0.5; 0; 0.5; 1; 2 or 3) representative of a number of different curve shapes. This set of possibilities is considered adequate because higher order polynomials can represent the data poorly [5].

Royston and Altman's algorithm [11] determines the fractional polynomial that describes the best-fit regression relationship between the predictor and the outcome. In practice, the algorithm selects first or second-degree fractional polynomials. First-degree fractional polynomials (FP1) involve a single term (regressor) raised to a power selected from either -2, -1, -0.5, 0, 0.5, 1, or 2. Second-degree fractional polynomials (FP2) occur when there are two terms for the regressors with different powers. The latter produce a wider variety of curve shapes [5].

For explanatory purposes this paper uses a first-degree polynomial $y = \beta_0 + \beta_1 x^p$ where p may be an integer or "fractional". Summary statistics are used to describe how well different models fit the data. Model deviances, which are the difference between data points and model predictors, are used to estimate goodness of fit when a pair of models are compared. The method uses a stepwise process, which looks at differences in statistical deviances when each model is compared with a linear model (i.e. with exponent of the regressor = 1) [5]. The maximum deviance difference is distributed approximately as x^2 with 1 degree of freedom. The significance of the deviance statistic is tested at a given probability. If, for example, the p-value is < 0.05, the hypothesis that the model is linear is rejected at 5% level of significance and the test is repeated. Where the significance of the deviation difference is not statistically significant, the simpler model is usually preferred. These tests apply to both univariable and multivariable models [20, 21].

Functions using the algorithm are available in SAS, Stata and R. This paper used Stata Version 9.0 (Stata Corp, College Station, Texas). Tests of statistical significance were at 5%.

RESULTS

Age was the regressor and invitation the outcome. Ten-year categories are common practice in epidemiological studies, but five-year categories (higher resolution) are also used in this example.

Table 1 includes estimates of odds ratios (with 95% confidence intervals) resulting from the logistic regression of age, in five-year categories, with the binary outcome "invitation". The odds of being invited to CR decreased with age. For patients aged >=55 & <60 years, the odds of being invited to CR were not statistically different from the reference group of patients aged less than 55 years. (The confidence interval includes 1). Odds ratios were statistically significant in all other age categories. In Table 2 the same analysis is shown when age is categorised in ten-year groups. This result shows a statistically significant association between age and invitation in each age group, but we know from Table 1 that this is not true for 55 to 60 years. At the very best, these categorical estimates offer an approximation of the relationship between age and invitation.

The analysis is repeated using the fractional polynomial algorithm that finds the best fit relationship between age and invitation. The algorithm tested the linear, the FP1 and the FP2 models. The first test, between the linear and FP1 models, resulted in a deviance difference of 16.930. A statistically significant relationship ($p < 0.001$) was evidence of non-linearity. In the second test, which compared FP1 and FP2 models, the deviance difference was 3.586. This relationship was not statistically significant ($p > 0.05$) and the FP1 model was preferred to the more complex FP2 model. In conjunction with finding an optimal curve shape based on a minimum deviance difference, Stata optimised a transformation of age such that age = x^3 -302.81 where x = *age in years/10.*

Scatter plots can inform data interpretation. Although it is common to assume that binary data are not easy to interpret in this way, this need not be the case. Scatter plots are important visual tools. They should be routinely used to qualitatively assess data distributions in health services research. In Figure 1, the data points are "jittered"; i.e. randomly nudged vertically to make the density clearer. This scatter (or jitter plot) shows that the number of patients invited to attend CR initially increased with age and then fell with increasing age. By comparing the density of the jitter plots one can estimate roughly whether more patients in the age band of interest were invited or *not* invited to CR. For older patients (>70years) the scatter is relatively denser in the *not* invited band and the reverse is true for younger patients. Although scatter plots are not normally used in health services research when an outcome variable is binary, they can be useful visual tools for managers and other decision-makers. (R software was used for Figure 1).

Figure 2 includes the observed data points (not jittered this time) and shows a curve of the predicted probability of invitation by age (with 95% confidence limits) from the best fit polynomial model. Compared with Tables 1 and 2, this presentation allows a more accurate interpretation of the marginal change in the probability of invitation for increments of age.

The practical importance of results such as shown here would obviously depend upon the focus of the study. For example, if the purpose was to inform managers about patterns of age discrimination in their program, the polynomial shows exactly where this occurs. What is important is that this approach allows accurate interpretation of the relationship for all possible values of the continuous predictor variable. While the example here is univariate, the same procedures apply when there are multiple predictors.

In the example given investigators may also be interested in whether there is an interaction between age and sex. When age is categorised, using either five or ten-year groups, the interaction is not interpreted as statistically significant ($p > 0.05$). However when age is transformed using the fractional polynomial algorithm, the age*sex interaction is statistically significant (χ^2=8.50, df=2, p=0.0143). Figures 3, 4 and 5 show predicted probabilities of invitation by age for males and females separately and together. Until age 60, older women were more likely to be invited than younger women, but women over 60 were less likely to be invited. Older men were less likely to be invited than younger men, and more likely to be invited than women. The age*sex interaction is significant above age 65 years (see divergent confidence intervals). The use of curves improves interpretation of the interaction between covariates.

DISCUSSION

In the health services research and across the social sciences, continuously measured variables are commonly grouped into categories. However, the simplicity achieved through the categorisation of continuous data is at a cost [9].

Cutpoints are usually based on largely untested assumptions. For any given area of research interest, there is typically a wide variation in cutpoints reported in the literature. Although a common practice has been to base cutpoints on previous studies, this can lead to incompatible and possibly biased results. The categorisation of data involves decisions regarding cutpoints the positioning of which influences estimates of effect size and statistical significance [22]. Additionally, when investigators stratify to control for confounding or to analyse interactions, the choice of categories can impact upon interpretation of the data [3].

Categories are often imposed at the data collection stage, but this can impede accurate interpretation of the distribution of values. While there are practical advantages in collecting data in pre-defined categories, it is important to acknowledge that this practice can compromise the results [23].

A well fitting model is not the same as a correct model. The categorisation of continuous variables can suggest that the model is correct in the central range of the data, but categories may not accurately represent the data for all possible values. Additionally, where distributions are skewed, or where values are under-represented or absent from the observed data, the validity of the study can be compromised. This paper uses simple examples, with graphs of curved relationships, to show the advantages of the fractional polynomial method for health services researchers when compared with the more usual practice of categorisation.

Fractional polynomials do not introduce investigator bias and they give interpretable curves. Although fractional polynomials can appear mathematically and computationally complex, the algorithm is performed efficiently by a number of accessible statistical software packages [5]. The resulting curves can show the best fit relationships for the observed data for given statistical confidence. This can provide insights not evident when data are binned.

This paper did not discuss the use of spline regression and generalised additive models (GAMs) for describing and fitting continuous covariates because these methods have been extensively documented in statistical texts and other publications [24, 25].

While polynomial methods have clear advantages, there is also a danger that the curved relationships will be "over interpreted by creative investigators". [23] Fractional polynomials can lack flexibility and lead to a poor fit of the data when, for example, the predictor variable has extreme values. Spline and kernel methods can be used to overcome such problems [5]. It is also important to keep in mind that evidence of non-linearity may result from the study design and implementation processes. Ultimately the methods chosen should supply a model that most adequately and suitably describes the data and addresses the study objectives.

CONCLUSIONS

The categorisation of continuous variables in regression models can be improved by fitting fractional polynomials to make use of information that is lost when cutpoints are introduced. These methods allow a fuller representation of non-linear relationships between predictor and outcome variables, and avoid the many statistical compromises that are made when data are arbitrarily aggregated and grouped. Ultimately the strength of any statistical method rests upon valid and accurate representation of data.

REFERENCES

[1] Royston P, Altman DG, Sauerbrei W. Dichotomising continuous predictors in multiple regression: a bad idea. Statistics in Medicine. 2006;25:127-41.

[2] Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. Commentary. Journal of National Cancer Institute. 1994;86(11):829-35.

[3] Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology. 1995;6(4):450-4.

[4] Brenner H, Blettner M. Controlling for continuous confounders in epidemiological research. Epidemiology. 1997;8(4):429-34.

[5] Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. International Journal of Epidemiology. 1999;28:964-74.

[6] Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology with an emphasis on fractional polynomials. Methods Inf Med. 2005;44:561-71.

[7] Scott DW. On optimal and data-based histograms. Biometrika. 1979;66:3.

[8] Wand MP. Data-based choice of histogram bin width. American Statistician. 1997;51(1):59-64.

[9] Wainer H. Picturing the uncertain world. How to understand, communicate, and control uncertainty through graphical display. Princeton: Princeton University Press 2009.

[10] Greenland S. Dose-response and trend analysis in Epidemiology: alternatives to categorical analysis. Epidemiology. 1995;6(4):356-65.

[11] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Applied Statistics. 1994;43(3):429-67.

[12] MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods. 2002;7(1):19-40.

[13] Oddy WH, Sherriff JL, de Klerk NH, Kendall GE, Sly PD, Beilin LJ, et al. The relation of breastfeeding and Body Mass Index to asthma and atopy in children: A prospective cohort study to age 6 Years. American Journal of Public Health. 2004 Sept;94(9):1531-7.

[14] Hyndman JCG, Holman CDJ, P DV. Effect of distance and social disadvantage on the response to invitations to attend mammography screening. J Med Screen. 2000;7:141-5.

[15] Brameld KJ, Holman CD, Lawrence DM, Hobbs MST. Improved methods for estimating incidence from linked hospital morbidity data. International Journal of Epidemiology. 2003;32:617-24.

[16] Holman CD, Wisniewski JB, Semmens JB, Rouse IL, Bass AJ. Mortality and prostate cancer risk in 19 598 men after surgery for benign prostatic hyperplasia. BJU International. 1999;84:37-42.

[17] Stewart Williams JA. Using non linear decomposition to explain the discriminatory effects of male-female differentials in access to care. A cardiac rehabilitation case study. Social Science & Medicine. 2009 Oct 2009;69(7):1072-9.

[18] Stewart Williams JA. Evidence and Equity in Healthcare. School of Medicine and Public Health, Faculty of Health, University of Newcastle; 2009 22 Oct; Newcastle; 2009.

[19] Stewart Williams JA, Byles JE, Inder K. Equity of access to cardiac rehabilitation: the role of system factors International Journal of Equity in Health. 2010 21 January 2010;9(2):7pp.

[20] Royston P, Ambler G. Multivariable fractional polynomials: update. Stata Technical Bulletin. 1999 May;49:17-23.

[21] 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;50:3464-85.

[22] Robertson C, Boyle P, Hsieh C, Macfarlane GJ, Maisonneuve P. Some statistical considerations in the analysis of case-control studies when exposure variables are continuous measurements. Epidemiology. 1994;5(2):164-70.

[23] Weinberg CR. How bad is categorisation? (Editorial). Epidemiology. 1995;6:345-7.

[24] Rothman KJ, Greenland S. Modern epidemiology. Second edition ed. Philadelphia: Lippincott Williams & Wilkins 1998.

[25] Ahrens W, Pigeot I, eds. Handbook of epidemiology. Bremen: Springer 2005.