CONTINUOUS VARIABLES: TO CATEGORIZE OR TO MODEL?

Willi Sauerbrei[1] and Patrick Royston[2]
[1]IMBI, University Medical Center Freiburg, Germany
[2]MRC Central Trials Unit, United Kingdom
wfs@imbi.uni-freiburg.de

Continuous variables are often encountered in life. We measure age, blood pressure and many other things. In medicine, such measurements are often used to assess risk or prognosis or to select a therapy. However, the question of how best to use information from continuous variables is relevant in many areas. To relate an outcome variable to a single continuous variable, a suitable regression model is required. A simple and popular approach is to assume a linear effect, but the linearity assumption may be violated. Alternatively, researchers typically apply cutpoints to categorize the variable, implying regression models with step functions. We illustrate problems caused by categorization and introduce fractional polynomials (FP) as a useful extension of polynomial regression. Investigating the effect of age as a prognostic factor for breast cancer, we show how conclusions depend critically on how the continuous variable is analyzed.

INTRODUCTION

Continuous variables are often encountered in life. We measure age, weight, blood pressure and many other things. In medicine, such measurements are often used to assess risk or prognosis or to select a therapy. However, the question of how best to use information from continuous variables is relevant in many areas. To relate an outcome variable to a single continuous variable, a suitable regression model is required.

A simple and popular approach is to assume a linear effect, but the linearity assumption may be questionable. To avoid this strong assumption, researchers often apply cutpoints to categorize the variable, implying regression models with step functions. This simplifies the analysis and interpretation of results. It seems that the usual approach in clinical and psychological research is to dichotomize continuous variables, whereas in epidemiological studies it is customary to create several categories, often four or five, allowing investigation of a possible dose–response relationship. However, categorization discards information and raises several critical issues such as how many cutpoints to use and where to place them (Altman et al., 1994; Royston et al., 2006).

Here, we illustrate problems caused by categorization and argue that the approach should be avoided when investigating the functional relationship between a continuous variable and the outcome. We introduce fractional polynomials (FP) as a useful extension of polynomial regression and as a sensible way to model the relationship (Royston and Sauerbrei 2008). Use of a suitable function selection procedure (FSP) gives a simple way to check whether a linear function (our default) is adequate or whether a non-linear FP function improves the fit of the data substantially. By investigating the effect of age as a prognostic factor for breast cancer, we illustrate how conclusions depend strongly on the manner in which the continuous variable age is analyzed. In real data, several variables influence the outcome and a multivariable model is required. The basic idea behind the multivariable fractional polynomial (MFP) approach is discussed.

From July 1984 to December 1989 the German Breast Cancer Study Group (GBSG) recruited 720 patients with primary node positive breast cancer into a trial. The dataset we use comprises recurrence-free survival (RFS) time of the 686 patients (with 299 events) who had complete data on the seven potential prognostic variables age, menopausal status, tumour size, tumour grade, number of positive lymph nodes, progesterone receptor and estrogen receptor. The values of age range from 21 to 80 years, the 10%, 25%, 50%, 75% and 90% centiles of the distribution being 40, 46, 53, 61 and 65 years, respectively. Further details and the website for downloading the data are given in Royston and Sauerbrei (2008) and the literature references there.

PROBLEMS CAUSED BY CATEGORIZATION

From a biological point of view, a cutpoint model is unrealistic, with individuals close to but on opposite sides of the cutpoint characterized as having different rather than similar outcomes. The underlying relationship with the outcome would be expected to be smooth but not necessarily linear. Use of two groups makes it impossible to detect a non-linear relationship.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

From a methodological point of view, loss of information is an important drawback of categorizing continuous variables, key issues being the number of cutpoints and where to place them (Altman et al., 1994, Royston et al., 2006). In prognostic research and when investigating interactions between a continuous covariate and treatment in a randomized trial, a common approach is to create two groups by using one cutpoint, which may increase the probability of false positive results (Altman et al. 1994, Royston and Sauerbrei 2008). It becomes extreme when 'optimal' cutpoints are used. Every possible cutpoint on x is considered and the value of x which minimizes the P-value is chosen. The cutpoint actually selected is to some extent due to chance. (Altman et al. 1994) called this procedure the 'minimum P-value' approach. Multiple testing increases the type I error probability from a nominal 0.05 to around 0.4. Although a correction to the P-value for multiple testing is available (Altman et al. 1994) the chosen cutpoint has a wide confidence interval and is rarely clinically meaningful. Critically, the difference in outcome between the two groups is over-estimated and its confidence interval is too narrow.

The methods for determining a cutpoint described below still incur information loss, but at least not an inflated type I error probability. Possibilities include recognized cutpoints (e.g. > 25 kg/m^2 to define 'overweight' based on body mass index), 'round number' such as multiple of five or ten, the upper limit of a reference interval in healthy individuals or cutpoints used in previous studies. In the absence of a prior cutpoint, the most common choice is the sample median. However, different studies have different cutpoints, so that their results can be neither easily compared nor summarized. For example, assessing the prognostic value of S-Phase fraction in breast cancer patients, 19 cutpoints were identified (Altman et al 1994). Several of them were the result of an 'optimal' cutpoint analysis.

Age as prognostic factor in breast cancer patients

Figure 1 illustrates the results of several analyses using different cutpoints. We show estimates of RFS probabilities for subgroups created by the cutpoints. Figure 1(a) shows a large difference between two groups created by using the optimal cutpoint of 37 years. Younger patients have much lower survival probabilities than patients above the cutpoint. The P-value is 0.004 (after adjustment for multiple testing, P = 0.1). The corresponding hazard ratio estimate for older patients from a Cox model is 0.54 (95% CI 0.37, 0.80). The difference between the two age groups is much reduced if the cutpoint is taken at the median (53 years), see Fig 1(b). The P-value is 0.4 and the estimated hazard ratio is 1.1. In Figure 1(c) we give the results for 3 groups created by using the cutpoints 45 and 60. These two cutpoints were predefined Differences between RFS are small and the test of an effect is not significant at a conventional level (P-value 0.15). Compared with the youngest group (age ≤ 45) the estimated hazard rates are 0.75 (middle group) and 0.82 (age > 60). Figure 1(d) gives RFS probabilities in five 10 year age groups, starting at 40 years. Some of the groups are small and this analysis is presented for illustrative purposes only. Patients younger then 40 have lower survival probabilities whereas differences between other groups are negligible. This result indicates that the prognostic effect of age cannot sensibly be described by a linear function.

FRACTIONAL POLYNOMIALS TO CHECK FOR NON-LINEARITY

In many cases, linearity reasonably well describes the functional relationship between a continuous variable and an outcome. This simple function has significant advantages when interpreting and presenting a model, for certain an important reason for its popularity. Nevertheless, there are numerous cases where linearity is not a reasonable assumption and results in a bad fit. For example, the prognostic effect of age in 10 year categories indicates that a linear function for age is probably unsuitable. In such cases the need for more complex functional forms is obvious. In the following we will propose the fractional polynomial (FP) approach, an extension of polynomial regression and therefore still simple, explainable and understandable. Nevertheless, the approach is quite flexible and can describe more complex relationships.

Fractional polynomials

As a starting point, we use the straight line model, $\beta_1 X$, for simplicity suppressing the constant term, β_0. A simple extension is a power transformation model, $\beta_1 X^p$. The latter model has often been used by practitioners in an *ad hoc* way, utilizing different choices of p. (Royston and Altman 1994) formalise the model slightly by calling it a first-degree fractional polynomial or FP1 function. The power p is chosen from a pragmatically restricted set $S = \{-2,-1,-0.5,0,0.5,1,2,3\}$, where X^0 denotes $\log X$ .As with polynomial regression, extension from one-term FP1 functions to the more complex and flexible two-term FP2 functions follows immediately. Instead of $\beta_1 X^1 + \beta_2 X^2$, FP2 functions with powers (p_1, p_2) are defined as $\beta_1 X^{p_1} + \beta_2 X^{p_2}$ with p_1 and p_2 taken from S. If $p_1 = p_2$ we use $\beta_1 X^{p_1} + \beta_2 X^{p_1} \log X$, a so-called repeated-powers FP2 model. For a more formal definition, see (Royston and Sauerbrei 2008). With the set S of powers as just given, there are 8 FP1 transformations, 28 FP2 transformations with distinct powers $(p_1 \neq p_2)$ and 8 FP2 transformations with equal powers $(p_1 = p_2)$. The best fit among the combinations of powers from S is defined as that with the highest likelihood. The class of FP2 curves is very flexible, despite the small number of models; see for example Figures 4.4 and 4.5 of Royston and Sauerbrei (2008).

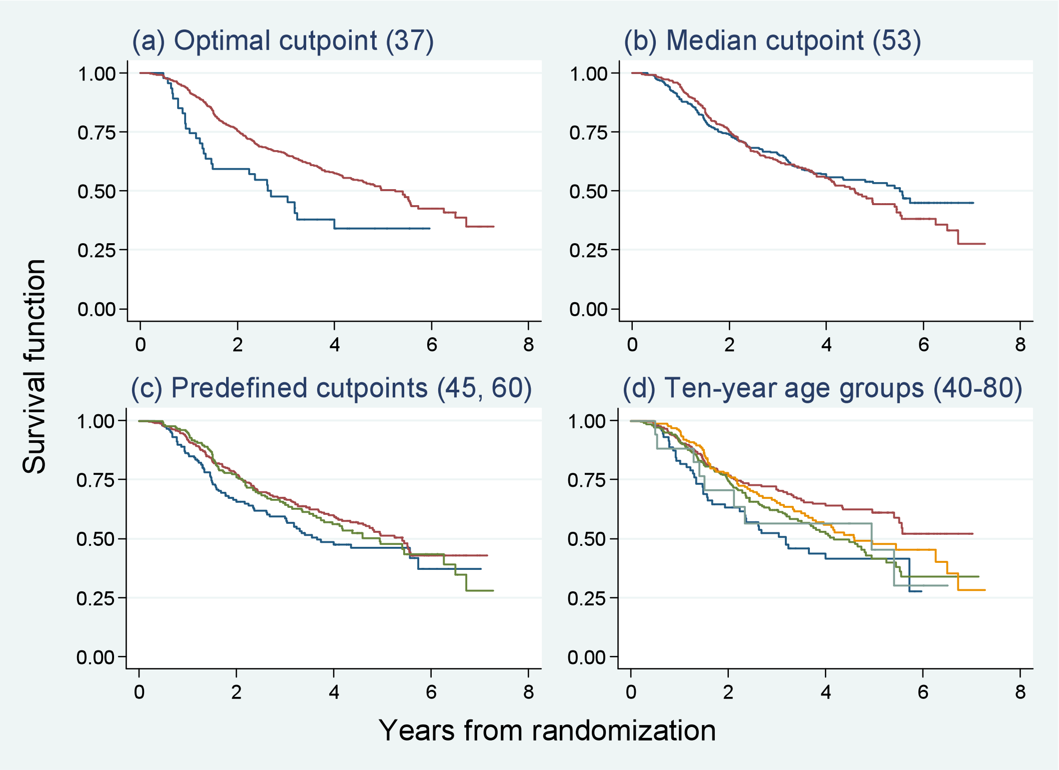

Figure 1. Four Kaplan-Meier Plots using cutpoints
The youngest group is always in blue. (a) ('Optimal' (37 years); (b) median (53 years); (c) predefined from earlier analyses (45, 60 years); (d) popular (10-year groups)

Function selection procedure (FSP)

Choosing the best FP1 or FP2 function by mininizing the deviance (minus twice the maximized log likelihood) is straightforward. As we prefer simple models, we consider the linear function to be a sensible default. Therefore, unless the data support a more complex FP function, a straight line model is chosen. As a strategy for selecting a 'best' FP model the following function selection procedure (FSP) is proposed (Royston and Sauerbrei 2008).

1. Test the best FP2 model for X at the α level against the null model using 4 d.f. If the test is not significant, stop, concluding that the effect of X is not significant. Otherwise continue.
2. Test the best FP2 for X against a straight line at the α level using 3 d.f. If the test is not significant, stop, the final model being a straight line. Otherwise continue.

3. Test the best FP2 for X against the best FP1 at the α level using 2 d.f. If the test is not significant, the final model is FP1, otherwise the final model is FP2. End of procedure.
   The test at step 1 is of overall association, step 2 examines the evidence for non-linearity.

The test at step 3 chooses between a simpler or more complex non-linear model. Before applying the procedure, the analyst must decide on the nominal P-value (α) and on the degree (m) of the most complex FP model allowed. Typical choices are $\alpha = 0.05$ and FP2 ($m = 2$).

Age as a prognostic factor in breast cancer patients

We extend the example investigating the effect of age on RFS in patients with breast cancer. In Table 1 we give the χ^2 values for the influence of age for the 8 FP1 models and the 36 FP2 models. Assuming a straight line model ($p_1 = 1$), age hardly affects RFS. χ^2 is 0.58, the corresponding critical value of χ^2 (1 d.f.) at the 5% level is 3.84. However, considering several non-linear models reveals that age has a non-linear effect. The best FP1 model ($p_1 = -2$) and the best FP2 model ($p_1 = -2, p_2 = -0.5$) both offer substantial improvements in model fit compared with the straight line model. The first step of the FSP compares the χ^2 difference between the best FP2 and the null model (χ^2=17.61, 4 df, $P < 0.001$). This is highly significant. The difference from the linear model (17.61 − 0.58=17.03, 3 d.f., $P < 0.001$) is considered in the second step. As this test gives also a highly significant result a final step has to decide whether an FP2 model is required or whether the best FP1 model is sufficient. The difference 11.20 ($17.61 - 6.41$) is also highly significant (2 d.f., $P < 0.001$) and an FP2 model with power terms ($-2,-0.5$) is selected. Figure 2 (left) shows the age functions for the three models, i.e. assuming a linear effect, selecting the FP2 function, or categorizing age (cutpoints 45 and 60) and estimating the resulting step functions. The functional forms are very different, and assessments based on significance testing give different results. Whereas age has no influence if linearity is assumed or if the two cutpoints 45 and 60 are chosen, a strong influence is indicated when the FP approach is used. As mentioned above, the ‘optimal’ cutpoint for age is 37 years. This cutpoint results in a function with a single large step (not shown).

The right part of Figure 2 shows the results of similar analyses for the number of positive lymph nodes, a well established strong prognostic factor.Although the effect is highly significant with all three approaches, the functional forms indicate a major difference. The step functions (the two cutpoints 3 and 10 have been used in clinical decision making for many years) give a rough approximation to the FP function, whereas the linear function underestimates the risk for a very small number of positive nodes and overestimates it for many positive nodes.

MULTIVARIABLE FRACTIONAL POLYNOMIALS

In most observational studies, several predictors are available and must be considered in the analysis. A suitable multivariable model is required. The aim is often to capture the important features of the data: the stronger predictors are included, predictors with weak or no effect are excluded and plausible functional forms are found for continuous variables.

Assuming linearity for continuous variables, backward elimination (BE) is a popular approach to determine which variables should be included and which can be excluded. BE starts with a model including all variables (often a mixture of binary, categorical and continuous variables) and uses significance testing to decide whether the variable with the largest P-value can be excluded from the model without harming the fit ‘seriously’. This procedure is done repeatedly until the largest P-value of the remaining variables is smaller than a pre-specified nominal significance level. Then BE terminates and the model is selected. Variations involving re-inclusion and re-exclusion of variables are sometimes used.

As noted above for age, linearity may be seriously violated and BE may erroneously exclude variables with a non-linear effect. As a pragmatic strategy for building models, a systematic search for possible non-linearity (provided by the FSP) is added to the BE procedure. The extension is feasible with any type of regression model to which BE is applicable. (Sauerbrei and Royston 1999) called it the multivariable fractional polynomial (MFP) procedure.Using MFP successfully requires only general knowledge of how to build a regression model. The nominal significance level for dropping variables and simplifying FP functions is the main tuning parameter. Therefore, reporting of MFP models can easily be done. The importance to report sufficient details of the analysis strategy is stressed in recent guidelines (McShane et al 2005).

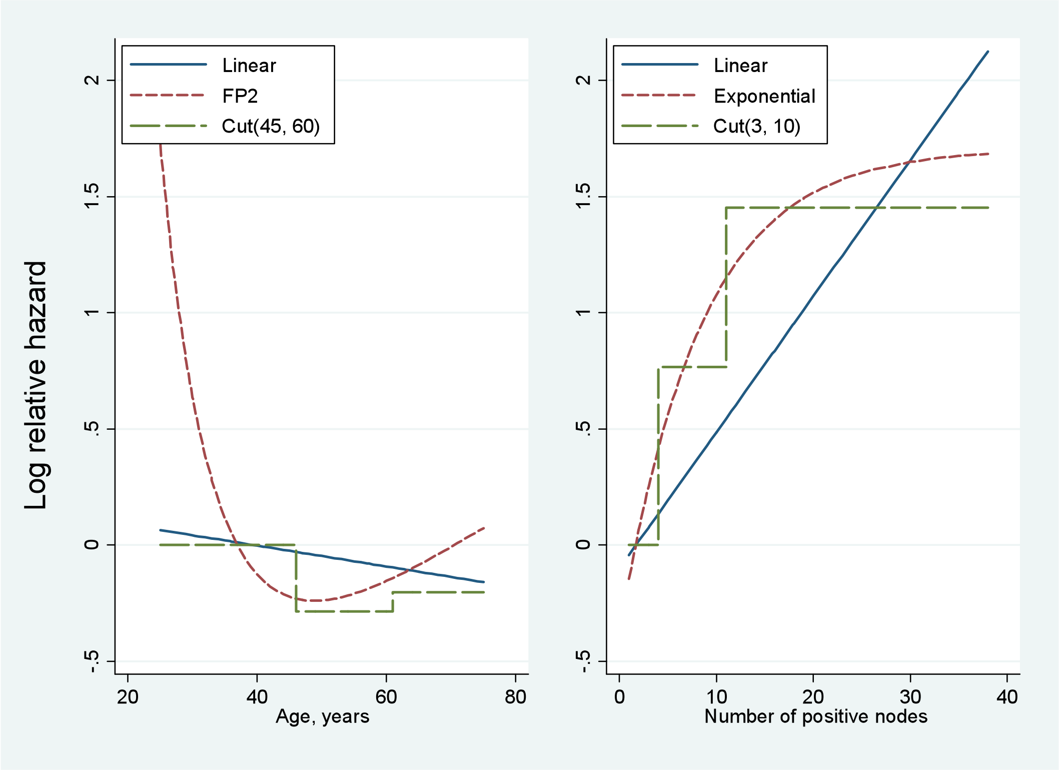

Figure 2. Linear, FP and step functions for age (left) and number of positive nodes (right)

Table 1. Deviance differences (compared to null model) of fractional polynomial models for age

Fractional polynomials										
First-degree		Second-degree								
Power	Model	Powers		Model	Powers		Model	Powers		Model
p	χ^2	p_1	p_2	χ^2	p_1	p_2	χ^2	p_1	p_2	χ^2
-2	6.41	-2	-2	17.09	-1	1	15.56	0	2	11.45
-1	3.39	-2	-1	17.57	-1	2	13.99	0	3	9.61
-0.5	2.32	-2	-0.5	17.61	-1	3	12.37	0.5		13.37
									0.5	
0	1.53	-2	0	17.52	-0.5	-0.5	16.82	0.5	1	12.29
0.5	0.97	-2	0.5	17.30	-0.5	0	16.18	0.5	2	10.19
1	0.58	-2	1	16.97	-0.5	0.5	15.41	0.5	3	8.32
2	0.17	-2	2	16.04	-0.5	1	14.55	1	1	11.14
3	0.03	-2	3	14.91	-0.5	2	12.74	1	2	8.99
		-1	-1	17.58	-0.5	3	10.98	1	3	7.15
		-1	-0.5	17.30	0	0	15.36	2	2	6.87
		-1	0	16.85	0	0.5	14.43	2	3	5.17
		-1	0.5	16.25	0	1	13.44	3	3	3.67

Seven prognostic factors were investigated in the breast cancer example. The final MFP model included the number of positive nodes and progesterone receptor, both with a non-linear function, grading as a binary variable and age as an FP2 function, very similar to the function from a univariate model shown in Figure 2. If we were to include in the model menopausal status, which is strongly correlated with age, the functional form for age would differ substantially for patients older than about 50 years. (see Figure 2 in Sauerbrei & Royston 1999). For further details of the multivariable model, see (Sauerbrei et al., 1999).

TEACHING EXPERIENCES

The material has been presented to audiences with big differences in statistical background and in lectures to students advanced in statistical modeling. We have also presented issues in handling continuous variables, fractional polynomials and MFP in two-day short courses on two occasions: once to an audience with a stronger medical background (clinical epidemiologists and researchers in general practice), and once to statisticians working in medical research. Besides the basic issues presented here, the course included extensive background to multivariable model-building and some extensions of FPs, including interactions with continuous predictors. The mode of presentation was traditional, comprising lectures with opportunities for the students to discuss issues among themselves and raise them with us. We did not include computer practicals.

We got very positive feedback from both groups of learners at the end of the course. However, we were aware that the non-statisticians struggled with some of the more complex aspects, particularly of the extensions, necessitating careful and sometimes lengthy explanation. Nevertheless, we feel that the concepts and practice of multivariable model-building with MFP are well within the grasp of researchers with some statistical knowledge and of masters-level students with a basic understanding of regression techniques. We encourage readers to consider putting together one-semester courses on the topic. Some specimen talks on aspects of MFP are available on the website of our book [http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/].

CONCLUSIONS

Continuous variables play a key role in many areas of research and in real life. In an example we have discussed and illustrated several critical issues when categorizing the data or assuming a linear relationship in a regression model, the two most popular ways to use continuous variables in data analysis. Most of the discussion has been restricted to a univariate analysis; in a multivariable analysis, the difficulties increase. We have presented the key aspects of the fractional polynomial approach and its extension for multivariable model building (MFP). We consider it as a suitable approach to handling continuous variables in many types of regression models. As the basic ideas require only the understanding of polynomial regression and of backward elimination, the approaches are potentially acceptable and useful in many application areas. Presentation of a FP model is simple and can be done in categories (for a detailed example see section 4.13 of Royston and Sauerbrei 2008). Such a presentation allows one to use the result of an MFP analysis in medical decision making and many other areas requiring categorization of data.

REFERENCES

Altman, D. G., Lausen, B., Sauerbrei, W., & Schumacher, M., (1994). Dangers of using “Optimal” cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute, 86*, 829-835.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., & Clark, G.M. for the Statistics Subcommittee of the NCI-EORTC Working on Cancer Diagnostics (2005). REporting recommendations for tumor MARKer prognostic studies (REMARK). *Journal of the National Cancer Institute, 97*(16), 1180-1184.

Royston, P., &Altman, D. G. (1994). Regression using fractional polynomials of continuous covariate: parsimonious parametric modelling (with disc), *Applied Statistics, 43,* 429–467

Royston, P, Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine 25*(1), 127-141.

Royston, P., & Sauerbrei, W. (2008). Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley.

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

Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., Schumacher, M., & the German Breast Cancer Study Group (GBSG) (1999). Modelling the effects of standard prognostic factors in node positive breast cancer, *British Journal of Cancer, 79*, 1752-1760.