Eva Lorenz
Dr. sc. hum.

Dose-response modelling of semicontinuous variables in epidemiology and clinical research

Fach/Einrichtung: Public Health / Institut für Public Health
Doktorvater: Prof. Dr. Heiko Becher

A goal in the analysis of epidemiological or clinical data is often the estimation of a dose-response relationship for semicontinuous risk factors which are composed of positive continuous values and zeros. Typical examples in cancer or cardiovascular disease epidemiology are occupational exposures, e.g. asbestos exposure or alcohol and tobacco consumption where a proportion of individuals may be completely unexposed, and the exposure of those who have been exposed follows a continuous distribution. There are both statistical problems, and problems with regard to interpretation arising from this situation. Recently, the spike at zero (SAZ) situation has been considered using an extended fractional polynomial (FP) approach. In earlier research, the correct model under some specific assumptions on univariate continuous distributions was derived and expanded by investigating the correct dose-response curve for a SAZ situations and univariate normal, log-normal and gamma distribution of the positive part of X.

Theoretical results from this thesis show that even the presumably simple case of two bivariate normally distributed covariates with two variables with SAZ poses some methodological challenges. An important part of modelling SAZ variables is the frequency of zeros and its relation to other covariates. One particular problem is that the four cell distribution (4CD) of two SAZ variables has effects on two levels, the (Spearman) correlation between the positive values of the continuous variables and the OR between binary indicators. Another issue is the correct way to combine variables for investigating an interaction. Depending on the 4CD, it is necessary to include up to three binary indicators into the model, as zero observations in one, the other, and both SAZ variables. An analytical derivation for this situation is presented. For more than two SAZ variables, the situation becomes even more complicated. A practical issue is discussed, namely whether for real data such a complicated model, even theoretically justified, is required or whether a simpler model is sufficient to describe the data.

In order to give practical recommendations for realistic scenarios, two simulation studies were performed. Results from the first simulation study in the Cox PH framework are that with increasing proportion of zeros the standard FP method which ignores the SAZ yielded unsatisfactory dose-response estimates whereas FP-spike gives a better estimate of the true functional relationship by additional modelling of the binary indicator. With a relatively small effect size of the binary indicator, as defined in one of the investigated scenarios, similar results were obtained by both methods. Overall, standard FP overestimated the model complexity more often than FP-spike. Non-linear relationships were detected by both methods since both are based on the same functional class to model the positive continuous functional relationship. Nevertheless, both methods yielded less precise estimates if the true functional relationship was not within the FP class although FP-spike performed slightly better in terms of magnitudes of error and CIs of estimates compared to standard FP.

Four methods for bivariate applications to model two variables with a SAZ, Bi-Sep (two variables with SAZ are each modelled with a separate binary indicator for zero values; V_i=(X_i=0), i=1,2) , Bi-D3 (two variables

with SAZ are modelled with three binary indicators for zero values depending on zero values in the other variable) , Bi-D1 (two variables with SAZ are modelled with one binary indicator for the joint zero category where X_1=0 and X_2=0) and Bi-Sub (two variables with SAZ are modelled with three binary indicators as in Bi-D3 and the positive continuous parts are divided in two submodels for each variable depending on zero values in the other variable) were developed. A second simulation study using the linear regression model was implemented in order to assess their properties, give guidance and to compare them to two recent methods, namely Linear (two variables are modelled untransformed with a separate binary indicator for zero values; V_i=(X_i=0), i=1,2) and MFP (multivariable FP modelling without binary indicators for proportions of zeros). The scenarios and results reported in this thesis assume linear relationships between exposure variables and outcome.

The overall power of all bivariate approaches which were able to model proportions of zeros explicitly was satisfactory and they performed well in modelling two variables with a SAZ in the investigated scenarios. They were all shown to be superior to standard MFP analyses in terms of mean squared difference, choice of linearity/non-linearity and precision of dose-response estimates. The point-wise mean overall estimates were reasonably close to the true effect in all scenarios. Estimates derived by Bi-Sub were rather unstable caused by the high number of degrees of freedom. It was shown that correlation between binary indicators only requires modelling two binary indicators (Bi-Sep). With increasing correlation between continuous variables, Bi-Sep and Bi-D3 were similarly robust and can both be recommended for analyses.

The key components for further comparisons are the relative and absolute sample size in the categories with one variable having positive values and the other zeros. Of course, correlation between covariates affects these rates. Zero values could occur in three different categories. Category A consists of observations with zero in both variables X_1=0 and X_2=0, categories B and C consist of observations with zeros in one variable and positive continuous value in the other variable. Category D consists of observations with positive continuous values in both variables.

Data from two epidemiological studies were used to further illustrate the methods. First, a lung cancer case-control study was used to illustrate the SAZ methods. Lifetime exposure duration to a 'list A job', a job with exposure to carcinogens, and smoking were two SAZ variables in this study. When analysing the effect of lifetime exposure to a 'list A job' on the risk of lung cancer the first stage of FP-spike indicated that a linear function described the influence of the exposure well. The deviances from the second modelling stage showed that neither the linear function nor the binary indicator could be omitted. This particular case highlighted the situation of an exposure variable that has a very strong effect in its dichotomous form, indicating an increased risk irrespective of dose. The correlation structure suggested that either modelling with Bi-D3 or Bi-Sub would be the most suitable bivariate approach since binary indicators for smoking and occupational exposure were strongly correlated.

The second study is a case-control study on premenopausal breast cancer in which duration of breastfeeding and hormone intake were two SAZ variables. The correlation structure suggested that modelling with Bi-Sep would be the most suitable approach since binary indicators for duration of breastfeeding and hormone intake were not correlated. Results from Bi-Sep were consistent with the ones derived by Bi-D3. The binary indicator for duration of hormone intake seemed to have a much larger effect than the one for duration of breastfeeding, as the joint effect of zeros in category A was close to the effect size in category C, indicating that the coefficient is dominated by the effect of duration of hormone intake. The examples showed that the deviance could be slightly decreased with FP-spike and the bivariate SAZ methods compared to the usual FP procedures.

Modelling unexposed individuals as a separate risk group has been shown to be a useful extension to modelling SAZ variables without considering binary indicators. Being theoretically justified already, the comparisons in both simulation studies suggested the use of FP-spike and the bivariate modelling approaches to be preferable to the standard FP or MFP. Furthermore, they mostly performed better in real data applications compared to standard approaches which do not consider proportions of zeros.