

Dose–response modelling for bivariate covariates with and without a spike at zero: theory and application to binary outcomes

E. Lorenz*

*Institute of Public Health, Medical Faculty, University of Heidelberg, Im
Neuenheimer Feld 324, 69120 Heidelberg, Germany*

C. Jenkner

*IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100
Freiburg, Germany*

W. Sauerbrei

*IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100
Freiburg, Germany*

H. Becher

*Institute of Public Health, Medical Faculty, University of Heidelberg, Im
Neuenheimer Feld 324, 69120 Heidelberg, Germany and Institute of
Medical Biometry and Epidemiology, University Medical Center
Hamburg-Eppendorf, Hamburg, Germany*

In epidemiology and clinical research, there is often a proportion of unexposed individuals resulting in zero values of exposure, meaning that some individuals are not exposed and those exposed have some continuous distribution. Examples are smoking or alcohol consumption. We will call these variables with a spike at zero (SAZ). In this paper, we performed a systematic investigation on how to model covariates with a SAZ and derived theoretical odds ratio functions for selected bivariate distributions. We consider the bivariate normal and bivariate log normal distribution with a SAZ. Both confounding and effect modification can be elegantly described by formalizing the covariance matrix given the binary outcome variable Y . To model the effect of these variables, we use a procedure based on fractional polynomials first introduced by Royston and Altman (1994, *Applied Statistics* 43: 429–467) and modified for the SAZ situation (Royston and Sauerbrei, 2008, *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*, Wiley; Becher *et al.*, 2012, *Biometrical Journal* 54: 686–700). We aim to contribute to theory, practical procedures and application in epidemiology and clinical research to derive multivariable models for variables with a SAZ. As an example, we use data from a case–control study on lung cancer.

*eva.lorenz@uni-heidelberg.de

Keywords and Phrases: bivariate dose–response model, spike at zero, regression modelling, fractional polynomials.

1 Introduction

A goal in the analysis of epidemiological or clinical data is often the estimation of a dose–response relationship for risk factors. Some of those risk factors may be semicontinuous variables, that is, a mixture of a point mass at a certain value and continuously distributed positive values. Typical examples in cancer or cardiovascular disease epidemiology are occupational exposures, for example, asbestos exposure or alcohol and tobacco consumption, where a proportion of individuals may be completely unexposed while the exposure follows a continuous distribution. Semicontinuous variables differ from variables that are left censored or truncated in that the zeros are valid data values rather than proxies for negative or missing responses. Conceptually, a true zero is different from a censored zero; a true zero typically represents something not happening, whereas a censored zero only indicates that its occurrence was below a certain threshold. These types of variables have lately received more attention; see, for example, Vink *et al.* (2014) where the authors present imputation methods for semicontinuous data. Dreassi *et al.* (2013) introduce a Bayesian approach using a two-part model where they suggest a gamma model for the skewed continuous distribution of the continuous variable as a more flexible alternative to the typical log transformation. Similar concepts of two-part models have been discussed by Olsen and Schafer (2001). However, in most applications, zero values are observed in the outcome variables, and many approaches to model it have been developed in economic and medical applications (zero-inflated Poisson regression) (Fletcher *et al.*, 2005).

In the following, we consider the case of a point mass at zero only and denote this as spike at zero (SAZ). In earlier research, we have derived the correct model for continuous exposures under some specific assumptions on univariate continuous distributions (Becher, 1992). Recently, we have expanded this by investigating the correct dose–response curve for variables with a SAZ, which are univariate normal, log normal and gamma distributed for the positive values of a continuous covariate X (Becher *et al.*, 2012). However, in practice, we rarely have only one covariate. In a recent paper, Yang and Fu (2013) provide the theoretical vector of regression coefficients for a special case of the bivariate normal distribution within the framework of logistic regression. They investigate the effect on the coefficient of one variable while ignoring the other variable, whereas we consider both variables simultaneously, consider their correlation structure and investigate implications for confounding and interaction.

It has to be emphasized that the model selection procedure in real-data situations is considerably more complex, because (i) usually a larger number of variables need to be considered simultaneously and (ii) the distribution of the covariates may only approximately follow common continuous distributions, for example, normal, log normal or gamma. Therefore, the theoretical considerations in this paper may serve as a guideline in distinguishing which model selection procedures may be useful.

The standard modelling techniques for the investigation of non-linear relationships such as the fractional polynomial (FP) approach (Royston and Sauerbrei, 2008) as well as the spline techniques (De Boer, 2001) do not specifically consider the SAZ situation. In situations where the unexposed are assumed to be different from the exposed, the functional relationship of the continuous variable on the outcome differs from the one between the unexposed individuals and the outcome. To allow for variables with a SAZ, the risk may be modelled as a point estimate for the binary exposure status and as an FP function for the continuous part of the variable. This model might provide more accurate estimates because it fits the distribution more closely. An FP-based function selection procedure with a binary indicator was suggested by extending the standard FP modelling (Royston and Sauerbrei, 2008; Royston *et al.*, 2010) and slightly modified by Becher *et al.* (2012). It is based on the FP procedure, which is a method to investigate whether the assumption of a linear effect is acceptable or whether a non-linear function from the class of FPs improves the data fit severely and is preferable to describe the functional influence of a covariate. This will be explained in detail in section 4. The extension consists of two stages to select a model. In the first stage, the best FP model is chosen, while a binary indicator, V , defined as $V=1$ if $X=0$ and $V=0$ if $X>0$, is included in the model. In the second stage, V and the selected FP function are each tested for removal from the model. For example, in the case-control study on lung cancer (see section 2 for details) with smoking as a covariate with dose X measured as pack-years, we obtain from a regression model with the linear predictor $\alpha + \beta x$ the regression coefficient $\beta = 0.034$. For variables like smoking, a certain proportion of individuals are usually unexposed. Thus, in such an approach, the SAZ was simply ignored. A consequence of this approach is that the odds ratio (OR) depends only on the difference between two measurements, that is, $\text{OR}(X = x + \Delta \text{ vs } x = \Delta) = \text{OR}(X = x \text{ vs } X = 0) = \exp(0.034x)$, which may not be appropriate. Adding a binary indicator V into the model, the linear predictor is $\alpha + \beta_0 v + \beta_1 x$ and yields $\beta_0 = -0.96$ and $\beta_1 = 0.027$, and $\text{OR}(X = x \text{ vs } X = 0) = \exp(0.96 + 0.027x)$, and $\text{OR}(X = x + \Delta \text{ vs } x = \Delta), \Delta > 0 = \exp(0.027x)$. So, for example, in the model with the binary variable V , the estimated OR is $\exp(0.96 + 0.027 \times 20) = 4.48$ for smoking 20 pack-years versus 0 and $\exp(0.027 \times 20) = 1.72$ for smoking 40 pack-years versus 20. In the model without the binary variable, the estimated OR is $\exp(0.034 \times 20) = 1.97$ in both cases.

In practice, modelling of SAZ variables is complicated by the inclusion of more than one covariate with a SAZ; therefore, we expand our investigations to the bivariate case. Results are an important argument for the proposal of different strategies to investigate the influence of two SAZ variables.

The structure of the paper is as follows: In section 2, we introduce the data set, which will be used to illustrate an application. We specifically consider the logistic regression model and investigate the correct ORs for a SAZ situation with selected bivariate distributions in section 3. Here, we consider the bivariate normal and bivariate log normal distribution, both with and without spike, to derive theoretical OR functions. Both confounding and effect modification will be formalized by properties of

the covariance matrix in diseased and non-diseased individuals. We will investigate the effect of correlations between the spike proportions as well as the overall correlation numerically. In section 4, we use direct logistic regression as well as a logistic regression that is based on FPs to obtain a bivariate dose–response curve using data from a case–control study on lung cancer. We describe the modelling approach for multiple SAZ variables based on FPs, which we have used in the data example. Finally, in section 5, we conclude the paper with a discussion.

2 Data

In our example, we use data from a hospital-based case–control study with 1004 lung cancer cases and 1004 population controls matched for region, sex and age. Here, we consider the number of years working in a so-called List A job (jobs with likely exposure to carcinogenic substances) and cumulative dose of smoking (pack-years) as continuous risk factors with a SAZ. For more details on the study, see Jöckel *et al.* (1998). The method used for the quantification of exposure to certain carcinogenic agents is described by Ahrens *et al.* (1993).

A relatively low number of cases (7.1%) and a higher number of controls (23.5%) reported zero exposure for smoking dose (X_1). Zero exposure in ‘List A’ job work duration (X_2) was reported in 54.0% of cases and 66.1% of controls (Table 1).

The positive ($x > 0$) part of the smoking variable is approximately log normally distributed in cases and controls; the positive part of ‘List A’ job work duration also has a right-skewed log normal distribution (Figure 1). The log-transformed positive values for smoking dose have a mean (standard deviation) of 1.83 (1.29) in cases and 1.55 (1.37) in controls. For the ‘List A’ job work duration, the corresponding values are 3.26 (0.87) in cases and 2.65 (1.19) in controls.

3 Theoretical odds ratio functions

In this chapter, we derive the dose–response curve under a specific regression model. We use logistic regression because this is the pertinent model for the data example. We assume that the bivariate distribution of two covariates is known and derive the

Table 1. Median values and proportion of spike for smoking and exposure to a ‘List A’ job in the lung cancer case–control study

		Smoking (pack-years) (X_1)				Σ	
		0	>0	0	>0		
‘List A’ job (years) (X_2)		0	0	>0	>0		
	Cases	N (%)	62 (6.2)	480 (47.8)	9 (0.9)	453 (45.1)	1004
		Median (X_1)	0	30.75	0	31.70	
Controls		0	0	4.00	6.20		
	N (%)	190 (18.9)	474 (47.2)	46 (4.6)	294 (29.3)	1004	
	Median (X_1)	0	17.40	0	21.40		
	Median (X_2)	0	0	4.00	4.75		
Σ (%)		252 (12.6)	954 (47.5)	55 (2.7)	747 (37.2)	2008	

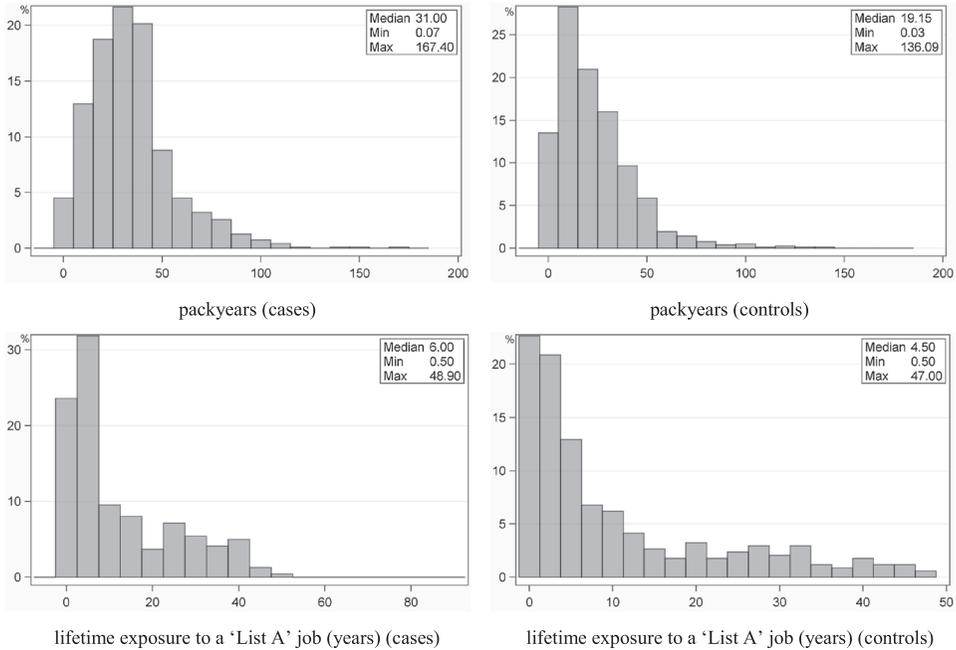


Fig. 1. Distribution of the continuous part of the risk factors smoking and lifetime exposure to a 'List A' job in cases and controls separately, lung cancer case-control study.

theoretical dose-response curve. Although in practice it is rare that the distribution of the covariate is known, the theoretical considerations may help obtain a picture of the underlying process. Here, we consider a bivariate normal and log normal distribution with a spike in none, one or both covariates.

3.1 Notation and definitions

We consider a binary response variable and logistic regression as the regression model. Let Y be a binary response variable (usually the disease of interest), and let X_1 and X_2 be the covariates of interest that have continuous distributions and possibly a SAZ. We denote the corresponding spike probabilities, that is, the probability of taking value zero, with $p_{h,i} = P(X_h = 0 | Y = i)$, where h denotes the two different variables ($h = 1, 2$) and i denotes the disease status ($i = 0, 1$). To widen the assumptions and allow for dependence of the spike proportions, we additionally use the spike probabilities $q_{k,i}$ with $k = 1, 2, 3, 4$, which refer to the four different categories in which the zeros could occur as $q_{1i} = P(X_1 = 0, X_2 = 0 | Y = i)$, $q_{2i} = P(X_1 \neq 0, X_2 = 0 | Y = i)$, $q_{3i} = P(X_1 = 0, X_2 \neq 0 | Y = i)$ and $q_{4i} = P(X_1 \neq 0, X_2 \neq 0 | Y = i) = 1 - q_{1i} - q_{2i} - q_{3i}$ with $0 < q_{ki} < 1$ for $k = 1, 2, 3, 4$ and $\sum_{k=1}^4 q_{ki} = 1$.

The OR function of the general bivariate situation within the logistic regression model can be formalized as follows:

$$\text{OR}_{x_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=x_{20}} = \frac{f_1(x_1^*, x_2^*)f_0(x_{10}, x_{20})}{f_1(x_{10}, x_{20})f_0(x_1^*, x_2^*)} \tag{1}$$

where (x_1^*, x_2^*) and (x_{10}, x_{20}) denote arbitrary realizations of the covariate vector (X_1, X_2) . Depending on the distribution of the covariates of interest, one needs to replace the functional terms f_1 and f_0 with the corresponding density functions.

We will now derive the OR function and the corresponding regression coefficients under the assumption that (X_1, X_2) have a bivariate normal and log normal distribution with and without a SAZ. A combined OR of one normal and one log normally distributed variable is straightforward, and one can easily derive it from the subsequent results.

3.2 Bivariate normal distribution without spike

First, we consider the bivariate normal distribution without a probability mass at zero. Let (X_1, X_2) be the vector of two covariates of interest, and let the second subscript of the realizations denote the case-control status, that is, x_{10} is a realization of covariate X_1 in a control.

For $x_1, x_2 \neq 0$, we have for $Y = 1$

$$X_1 \sim N\left(\begin{pmatrix} \mu_{11} \\ \mu_{11} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \rho_1 \sigma_{11} \sigma_{21} \\ \rho_1 \sigma_{11} \sigma_{21} & \sigma_{21}^2 \end{pmatrix}\right)$$

and for $Y = 0$

$$X_0 \sim N\left(\begin{pmatrix} \mu_{10} \\ \mu_{10} \end{pmatrix}, \begin{pmatrix} \sigma_{10}^2 & \rho_0 \sigma_{10} \sigma_{20} \\ \rho_0 \sigma_{10} \sigma_{20} & \sigma_{20}^2 \end{pmatrix}\right)$$

with density functions

$$f_i(X_1, X_2) = \left(2\pi\sigma_{1i}\sigma_{2i}\sqrt{1-\rho_i^2}\right)^{-1} \exp\left(-\frac{1}{2(1-\rho_i^2)}\left(\frac{X_1-\mu_{1i}}{\sigma_{1i}}\right)^2 - 2\rho_i\left(\frac{X_1-\mu_{1i}}{\sigma_{1i}}\right)\left(\frac{X_2-\mu_{2i}}{\sigma_{2i}}\right) + \left(\frac{X_2-\mu_{2i}}{\sigma_{2i}}\right)^2\right) \tag{2}$$

with means μ_{11}, μ_{21} in the diseased and μ_{10}, μ_{20} in the non-diseased, σ_{hi} as standard deviation and correlation coefficient

$$\rho_i(X_1, X_2|Y = i) = \frac{\text{cov}(X_1, X_2|Y = i)}{\sigma_{1i}\sigma_{2i}} \tag{3}$$

The following can be derived for general variances, but for a simpler presentation, we assume equal variances $\sigma_{hi}^2 = 1$. The density function is then

$$f_i(X_1, X_2) = (2\pi\sqrt{1-\rho_i^2})^{-1} \exp\left(-\frac{1}{2(1-\rho_i^2)}\left((X_1 - \mu_{1i})^2 - 2\rho_i(X_1 - \mu_{1i})(X_2 - \mu_{2i}) + (X_2 - \mu_{2i})^2\right)\right) \tag{4}$$

In the numerical analysis (section 3.6), we also consider more general cases. Algebraic calculations show that the OR given in (1) can be expressed as

$$\exp\left[\beta_1(x_1^* - x_{10}) + \beta_2(x_2^* - x_{20}) + \beta_3(x_1^{*2} - x_{10}^2 + x_2^{*2} - x_{20}^2) + \beta_4(x_1^*x_2^* - x_{10}x_{20})\right] \tag{5}$$

with

$$\begin{aligned} \beta_1 &= \frac{\mu_{11}}{(1-\rho_1^2)} - \frac{\mu_{10}}{(1-\rho_0^2)} + \frac{\mu_{20}\rho_0}{(1-\rho_0^2)} - \frac{\mu_{21}\rho_1}{(1-\rho_1^2)} \\ \beta_2 &= \frac{\mu_{21}}{(1-\rho_1^2)} - \frac{\mu_{20}}{(1-\rho_0^2)} + \frac{\mu_{10}\rho_0}{(1-\rho_0^2)} - \frac{\mu_{11}\rho_1}{(1-\rho_1^2)} \\ \beta_3 &= \frac{1}{2(1-\rho_0^2)} - \frac{1}{2(1-\rho_1^2)} \\ \beta_4 &= \frac{\rho_1}{(1-\rho_1^2)} - \frac{\rho_0}{(1-\rho_0^2)} \end{aligned} \tag{6}$$

Note that we summarized the coefficient for X_1 and X_2 squared to one joint coefficient β_3 , which is only possible when $\sigma_{hi}^2 = 1$. Then the coefficient only depends on the correlation and therefore is independent of the mean values.

From this, we can derive some properties given in the previously defined assumptions:

1. The correct model to express this OR requires X_1 and X_2 untransformed, X_1 and X_2 squared and the multiplicative term X_1X_2 .
2. For $\rho_0 = \rho_1 = 0$, we obtain the same results for the regression coefficients as in the univariate case (Becher *et al.*, 2012).
3. If $\rho_0 = \rho_1 = \rho$, we have $\beta_3 = \beta_4 = 0$, and the correct model simplifies to

$$\text{logit } P(Y = 1|X_1 = x_1, X_2 = x_2) = \alpha + \beta_1x_1 + \beta_2x_2$$

with

$$\begin{aligned} \beta_1 &= \frac{1}{(1-\rho^2)} (\mu_{11} - \mu_{10} + \rho(\mu_{20} - \mu_{21})) \\ \beta_2 &= \frac{1}{(1-\rho^2)} (\mu_{21} - \mu_{20} + \rho(\mu_{10} - \mu_{11})) \end{aligned} \tag{7}$$

Thus, it defines the condition for confounding, but no interaction. There is a correlation between the two variables, and this correlation does not depend on the disease status.

4. If $\rho_0 \neq \rho_1$, the correct model requires X_1 and X_2 squared and the multiplicative term X_1X_2 . The latter is known as an interaction term.
5. $\rho_0 = 0$ implies that the variables are not correlated in the controls, and thus, there is no confounding. Therefore, $\rho_0 = 0$ and $\rho_1 \neq 0$ would be the natural conditions for interaction without confounding. It is easily seen that the regression coefficients β_1 and β_2 however change (Equation (6)) unless $\mu_{11} = \mu_{21} = 0$. For this case, we obtain $\beta_1 = -\mu_{10}$, $\beta_2 = -\mu_{20}$, $\beta_3 = (1/2) - [1/2(1 - \rho_1^2)]$ and $\beta_4 = \rho_1 / (1 - \rho_1^2)$. The corresponding OR function in the case of $\rho_0 = 0$ is

$$\begin{aligned}
 & \text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=x_{20}} = \\
 = & \exp \left[-\mu_{10}(x_1^* - x_{10}) - \mu_{20}(x_2^* - x_{20}) + \left(\frac{1}{2} - \frac{1}{2(1 - \rho_1^2)} \right) (x_1^{*2} - x_{10}^2 + x_2^{*2} - x_{20}^2) \right. \\
 & \left. + \left(\frac{\rho_1}{(1 - \rho_1^2)} \right) (x_1^* x_2^* - x_{10} x_{20}) \right] \tag{8}
 \end{aligned}$$

In Figure 2, we illustrate the OR function for a few parameter combinations. We use $\rho_0 = 0, \rho_1 = 0, 0.2, 0.4, \sigma_{hi}^2 = 1, \mu_0 = \begin{pmatrix} -0.4 \\ -0.2 \end{pmatrix}, \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $x_{10} = 0.5$. We keep the second variable constant with $x_2^* = 0.5$. The OR function $\text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_{10}=0.5, X_2=x_2^*}$ then becomes

$$\left[-\mu_{10}(x_1^* - 0.5) + \left(\frac{1}{2} - \frac{1}{2(1 - \rho_1^2)} \right) (1 - 0.5^2) + \left(\frac{\rho_1}{(1 - \rho_1^2)} \right) (x_1^* - 0.5) \right]$$

In Figure 2, we show that the OR function increases more strongly if the correlation between X_1 and X_2 becomes larger.

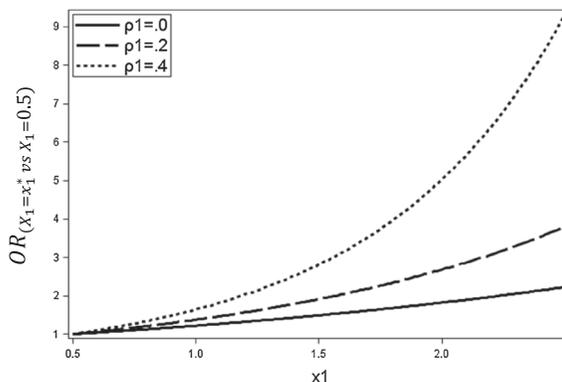


Fig. 2. Dose-response curve for X_1 for two variables with normal distribution and different correlation coefficients.

Table 2 gives some numerical examples of the regression coefficients for different values of ρ_0 and ρ_1 with $\sigma_{hi}^2 = 1$ and means (A) $\mu_0 = \begin{pmatrix} -1.0 \\ -0.5 \end{pmatrix}$, $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and (B) $\mu_0 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$, $\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, as well as the resulting ORs for both cases (A) $OR_{X_1=0, X_2=0 \text{ vs } X_1=-1, X_2=-0.5}$ and (B) $OR_{X_1=1, X_2=1 \text{ vs } X_1=-0, X_2=-0.5}$. The mean values in A and B are chosen such that the mean differences in cases and controls remain the same, that is, $0 - (-1) = 1 - 0$ for X_1 and $0 - (-0.5) = 0.5 - 0$ for X_2 . This is to show the situation for which the ORs are independent of the reference value.

Case (i) in Table 2 is the simple case without correlation, that is, no confounding between the two variables. The regression coefficients are calculated as the difference of means in the diseased and non-diseased. Cases (ii), (iii), (iv), (v) and (vi) show numerically that positive confounding without interaction yields reduced regression coefficients. The absolute and relative reduction is smaller for the stronger risk factor X_1 . Cases (vii) and (viii) show the relevance of the interaction term. While in the first six cases, the ORs are identical for A and B, that is, here, ORs do not depend on the baseline value, we have a different situation here. If a positive interaction is present, the ORs become larger with increasing reference value (3.34 vs 3.82 and 3.09 vs 3.45, respectively).

3.3 Bivariate normal distribution with spike in one covariate

We now extend the bivariate normal distribution, again with $\sigma_{hi} = 1$, by a SAZ in covariate X_2 and define the following density function

$$f_{s,i}(x_1, x_2) = \begin{cases} p_{2i}f_i(x_1) & x_1 \neq 0, x_2 = 0 \\ (1 - p_{2i})f_i(x_1, x_2) & x_1 \neq 0, x_2 \neq 0 \end{cases} \quad (9)$$

where $p_{2i} = P(X_1 \neq 0, X_2 = 0 | Y = i)$. Here, $f_i(x_1, x_2)$ is the bivariate density function as before, and $f_i(x_1)$ is the density function of the marginal distribution, which is a normal distribution with mean μ_{1i} .

The calculation of the theoretical OR function becomes more complicated because different cases have to be considered as given in (a)–(d) in the

Table 2. Theoretical regression coefficients for two normally distributed variables and ORs for selected parameter combinations

	ρ_0	ρ_1	β_1	β_2	β_3	β_4	OR (A)	OR (B)
(i)	0.0	0.0	1.00	0.50	0.000	0.00	3.49	3.49
(ii)	0.1	0.1	0.96	0.40	0.000	0.00	3.20	3.20
(iii)	0.2	0.2	0.94	0.31	0.000	0.00	2.99	2.99
(iv)	0.3	0.3	0.93	0.22	0.000	0.00	2.84	2.84
(v)	0.4	0.4	0.95	0.12	0.000	0.00	2.75	2.75
(vi)	0.5	0.5	1.00	0.00	0.000	0.00	2.72	2.72
(vii)	0.0	0.1	1.00	0.50	-0.005	0.10	3.34	3.82
(viii)	0.1	0.2	0.96	0.40	-0.016	0.11	3.09	3.45

Note: OR, odds ratio.

following list. In the first case, the baseline value is a positive value for both variables X_1 and X_2 . The second case describes the OR for $X_2 = x_2^*$ versus zero exposure to variable X_2 . The third case describes the OR for $X_1 = x_1^*$ versus $X_1 = x_{10}$ given zero value for X_2 . We have to consider OR functions as follows:

- (a) $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=x_{20}}$
- (b) $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_1^*, X_2=0}$
- (c) $OR_{X_1=x_1^*, X_2=0 \text{ vs } X_1=x_{10}, X_2=0}$
- (d) $OR_{X_1=x_1^*, X_2=x_{20} \text{ vs } X_1=x_{10}, X_2=x_{20}}$

which are

$$(a) \quad OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=x_{20}} = \frac{(1 - p_{21})(1 - p_{20})f_1(x_1^*, x_2^*)f_0(x_{10}, x_{20})}{(1 - p_{21})(1 - p_{20})f_1(x_{10}^*, x_{20}^*)f_0(x_1^*, x_2^*)}$$

This is the OR function for the bivariate case without a spike as given in (5) and (6). For

$$(b) \quad OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \frac{(1 - p_{21})(p_{20})f_1(x_1^*, x_2^*)f_0(x_{10})}{(p_{21})(1 - p_{20})f_1(x_{10})f_0(x_1^*, x_2^*)}$$

we obtain $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \exp(\beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1^2 + x_2^2) + \beta_4 (x_1 x_2))$ with

$$\begin{aligned} \beta_{02} &= \ln \left(\frac{(1 - p_{21})(p_{20})}{(p_{21})(1 - p_{20})} \sqrt{\frac{1 - \rho_0^2}{1 - \rho_1^2}} \right) \\ &\quad + \frac{-\mu_{11}^2 + 2p_1\mu_{11}\mu_{21} - \mu_{21}^2 + \mu_{21}^2(1 - \rho_1^2)}{2(1 - \rho_1^2)} \\ &\quad + \frac{-\mu_{10}^2 + 2p_0\mu_{10}\mu_{20} + \mu_{20}^2 - \mu_{20}^2(1 - \rho_0^2)}{2(1 - \rho_0^2)} \\ \beta_1 &= \frac{\mu_{11}}{(1 - \rho_1^2)} - \frac{\mu_{10}}{(1 - \rho_0^2)} + \frac{\mu_{20}\rho_0}{(1 - \rho_0^2)} - \frac{\mu_{21}\rho_1}{(1 - \rho_1^2)} + \mu_{10} - \mu_{11} \\ \beta_2 &= \frac{\mu_{21}}{(1 - \rho_1^2)} - \frac{\mu_{20}}{(1 - \rho_0^2)} + \frac{\mu_{10}\rho_0}{(1 - \rho_0^2)} - \frac{\mu_{11}\rho_1}{(1 - \rho_1^2)} \\ \beta_3 &= \frac{1}{2(1 - \rho_0^2)} - \frac{1}{2(1 - \rho_1^2)} \\ \beta_4 &= \frac{\rho_1}{(1 - \rho_1^2)} - \frac{\rho_0}{(1 - \rho_0^2)} \end{aligned} \tag{10}$$

For the third case,

$$(c) \text{ OR}_{X_1=x_1^*, X_2=0 \text{ vs } X_1=x_{10}, X_2=0} = \frac{(p_{21})(p_{20})f_1(x_1^*)f_0(x_{10})}{(p_{21})(p_{20})f_1(x_{10})f_0(x_1^*)}$$

we obtain $\text{OR}_{X_1=x_1^*, X_2=0 \text{ vs } X_1=x_{10}, X_2=0} = \exp(\beta_1 x_1 + \beta_2 x_1^2)$ with

$$\beta_1 = \frac{\mu_{11}}{(1 - \rho_1^2)} - \frac{\mu_{10}}{1 - \rho_0^2}$$

$$\beta_2 = \frac{1}{2(1 - \rho_0^2)} - \frac{1}{2(1 - \rho_1^2)}$$

and the coefficients for X_2 and x_2^2 both cancel out.

For the fourth case,

$$(d) \text{ OR}_{X_1=x_1^*, X_2=x_{20} \text{ vs } X_1=x_{10}, X_2=x_{20}} = \frac{(1 - p_{21})(1 - p_{20})f_1(x_1^*, x_{20})f_0(x_{10}, x_{20})}{(1 - p_{21})(1 - p_{20})f_1(x_{10}, x_{20})f_0(x_1^*, x_{20})}$$

we obtain $\text{OR}_{X_1=x_1^*, X_2=x_{20} \text{ vs } X_1=x_{10}, X_2=x_{20}} = \exp(\beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1 x_2)$ with coefficients β_1 and β_2 equal to the results from the third case earlier and an additional interaction term with the following coefficient:

$$\beta_3 = \frac{\rho_1}{(1 - \rho_1^2)} - \frac{\rho_0}{(1 - \rho_0^2)}$$

The only difference between the third and fourth cases is the additional interaction term, which cancels out if $\rho_0 = \rho_1 = \rho$.

3.4 Bivariate normal distribution with spike in both covariates

We now extend the bivariate normal distribution, again with $\sigma_{hi} = 1$, by a SAZ in both covariates and define the following density function:

$$f_{s,i}(x_1, x_2) = \begin{cases} q_{1i} & x_1 = 0, x_2 = 0 \\ q_{2i}f_i(x_1) & x_1 \neq 0, x_2 = 0 \\ q_{3i}f_i(x_2) & \text{if } x_1 = 0, x_2 \neq 0 \\ q_{4i}f_i(x_1, x_2) & x_1 \neq 0, x_2 \neq 0 \end{cases} \tag{11}$$

where $f_i(x_1, x_2)$ is defined as before and we have now spike probabilities q_{ki} with $q_{1i} = P(X_1 = 0, X_2 = 0 | Y = i)$, $q_{2i} = P(X_1 \neq 0, X_2 = 0 | Y = i)$, $q_{3i} = P(X_1 = 0, X_2 \neq 0 | Y = i)$ and $q_{4i} = P(X_1 \neq 0, X_2 \neq 0 | Y = i) = 1 - q_{1i} - q_{2i} - q_{3i}$ with $0 < q_{ki} < 1$ for $k = 1, 2, 3, 4$ and $\sum_{k=1}^4 q_{ki} = 1$. The joint distribution of the SAZ in both variables can be expressed with these four probabilities, of which the fourth is one minus the other three, and we call this four-cell distribution (4CD) in the following.

Here $f_i(x_1)$ and $f_i(x_2)$ are the marginal distributions, which are normal distributions with means μ_{1i} and μ_{2i} , respectively. If $X_2=0$, we assume that the conditional distribution of $X_{1i}|X_{2i}=0$ is a normal distribution with mean μ_{1i} . The equivalent holds for the distribution of $X_{2i}|X_{1i}=0$. If the spikes are independent, we can express the bivariate spike probabilities with the parameters p_{1i}, p_{2i} as $q_{1i}=p_{1i}p_{2i}, q_{2i}=p_{2i}(1-p_{1i}), q_{3i}=p_{1i}(1-p_{2i})$ and $q_{4i}=(1-p_{1i})(1-p_{2i})$ with the following density function:

$$f_{s,i}(x_1, x_2) = \begin{cases} p_{1i}p_{2i} & x_1 = 0, x_2 = 0 \\ p_{2i}(1-p_{1i})f_i(x_1) & x_1 \neq 0, x_2 = 0 \\ p_{1i}(1-p_{2i})f_i(x_2) & \text{if } x_1 = 0, x_2 \neq 0 \\ (1-p_{1i})(1-p_{2i})f_i(x_1, x_2) & x_1 \neq 0, x_2 \neq 0 \end{cases} \quad (12)$$

From (11) and (12), it can be seen that it is not straightforward to describe the covariance matrix of (X_1, X_2) . There are two different levels of correlation, between the binary indicators and between the continuous components. This has to be considered in deriving the correct OR functions and will be shown in the succeeding text.

Figure 3 gives the empirical density function for a bivariate normally distributed variable when simulating 100 000 random values, which are distributed according to (Equation 12) with $p_1=0.2, p_2=0.3, \mu_1=4, \mu_2=4$ and $\rho=0.3$ between X_1 and X_2 .

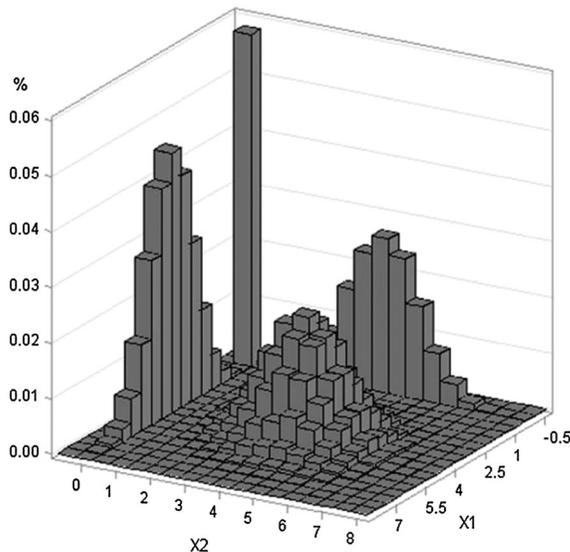


Fig. 3. Empirical density function of a bivariate normally distributed set $x = (x_1, x_2)$. The joint bivariate distribution is displayed in the centre of the graph, the marginal distributions of X_1 and X_2 are displayed on the x and y axes, respectively, and the zero proportion is displayed in the origin.

The calculation of the theoretical OR function becomes more complicated because different cases have to be considered.

In the first case, the baseline value is the non-exposure to both variables X_1 and X_2 and in the second case only to one of both. This has implications for the regression coefficients, as shown in the succeeding text.

In the first case, allowing the spikes to be dependent as defined in (11), we obtain

$$\begin{aligned}
 \text{(a) } \text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=0} &= \frac{f_{s,1}(x_1^*, x_2^*)f_{s,0}(0, 0)}{f_{s,1}(0, 0)f_{s,0}(x_1^*, x_2^*)} \\
 &= \frac{q_{10}q_{41}f_1(x_1^*, x_2^*)}{q_{40}q_{11}f_0(x_1^*, x_2^*)}
 \end{aligned}
 \tag{13}$$

and in the second case, we obtain

$$\begin{aligned}
 \text{(b) } \text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} &= \frac{f_{s,1}(x_1^*, x_2^*)f_{s,0}(x_{10}, 0)}{f_{s,1}(x_{10}, 0)f_{s,0}(x_1^*, x_2^*)} \\
 &= \frac{q_{20}q_{41}f_1(x_1^*, x_2^*)f_0(x_{10})}{q_{40}q_{21}f_1(x_{10})f_0(x_1^*, x_2^*)}
 \end{aligned}
 \tag{14}$$

3.4.1 Theoretical odds ratio when baseline value is zero exposure to both variables X_1 and X_2

The first case describes the OR for $X_1 = x_1^*$ and $X_2 = x_2^*$ versus zero value for X_1 and X_2 .

Algebraic calculations give the OR function for case (a) as follows:

$$\text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_{10}=0, X_{20}=0} = \exp(\beta_{01} + \beta_1x_1 + \beta_2x_2 + \beta_3(x_1^2 + x_2^2) + \beta_4x_1x_2)
 \tag{15}$$

with

$$\beta_{01} = \ln\left(\frac{q_{10}q_{41}}{q_{40}q_{11}}\sqrt{\frac{1-\rho_0^2}{1-\rho_1^2}}\right) + \frac{-\mu_{11}^2 + 2\rho_1\mu_{11}\mu_{21} - \mu_{21}^2}{2(1-\rho_1^2)} + \frac{-\mu_{10}^2 + 2\rho_0\mu_{10}\mu_{20} - \mu_{20}^2}{2(1-\rho_0^2)}
 \tag{16}$$

and $\beta_1, \beta_2, \beta_3$ and β_4 as before in the case without a spike (6).

If $\rho_0 = \rho_1 = \rho$, the correct model simplifies to $\text{logit } P(Y = 1 | X_1 = x_1, X_2 = x_2) = \beta_{01} + \beta_1x_1 + \beta_2x_2$ with

$$\begin{aligned} \beta_{01} &= \ln\left(\frac{q_{10}q_{41}}{q_{40}q_{11}}\right) + \frac{-\mu_{11}^2 + 2\rho\mu_{11}\mu_{21} - \mu_{21}^2 + \mu_{10}^2 - 2\rho\mu_{10}\mu_{20} + \mu_{20}^2}{2(1-\rho^2)} \\ \beta_1 &= \frac{1}{(1-\rho^2)} (\mu_{11} - \mu_{10} + \mu_{20}\rho = -\mu_{21}\rho) \\ \beta_2 &= \frac{1}{(1-\rho^2)} (\mu_{21} - \mu_{20} + \mu_{10}\rho = -\mu_{11}\rho) \end{aligned} \tag{17}$$

3.4.2 Theoretical odds ratio when baseline value is the non-exposure to only one variable

The second case describes the OR for $X_1 = x_1^*$ and $X_2 = x_2^*$ versus $X_1 = x_{10}$ given zero value for X_2 . From algebraic calculations on case (b), we obtain $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \exp(\beta_{02} + \beta_1x_1 + \beta_2x_2 + \beta_3(x_1^2 + x_2^2) + \beta_4(x_1x_2))$ with coefficients $\beta_1, \beta_2, \beta_3$ and β_4 as in (10), and for β_{02} , we obtain

$$\begin{aligned} \beta_{02} &= \ln\left(\frac{q_{20}q_{41}}{q_{40}q_{21}}\sqrt{\frac{1-\rho_0^2}{1-\rho_1^2}}\right) \\ &+ \frac{-\mu_{11}^2 + 2\rho_1\mu_{11}\mu_{21} - \mu_{21}^2 - \mu_{21}^2 + \mu_{21}^2(1-\rho_1^2)}{2(1-\rho_1^2)} \\ &+ \frac{-\mu_{10}^2 - 2\rho_0\mu_{10}\mu_{20} + \mu_{20}^2 - \mu_{20}^2 - \mu_{20}^2(1-\rho_0^2)}{2(1-\rho_0^2)} \end{aligned} \tag{18}$$

If $\rho_0 = \rho_1 = \rho$, the correct OR function simplifies further to $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \exp(\beta_{02} + \beta_1x_1 + \beta_2x_2)$ with

$$\begin{aligned} \beta_{02} &= \ln\left(\frac{q_{20}q_{41}}{q_{40}q_{21}}\right) \frac{-\mu_{11}^2 + 2\rho\mu_{11}\mu_{21} + \mu_{21}^2\rho^2 + \mu_{10}^2 - 2\rho\mu_{10}\mu_{20} - \mu_{20}^2\rho^2}{2(1-\rho_1^2)} \\ \beta_1 &= \frac{1}{(1-\rho^2)} (\mu_{11} - \mu_{10} + \mu_{20}\rho - \mu_{21}\rho) \\ \beta_2 &= \frac{1}{(1-\rho^2)} (\mu_{21} - \mu_{20} + \mu_{10}\rho - \mu_{11}\rho) \end{aligned} \tag{19}$$

For the symmetric case (c), we obtain $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=x_{20}} = \exp(\beta_{03} + \beta_1x_1 + \beta_2x_2 + \beta_3(x_1^2 + x_2^2) + \beta_4x_1x_2)$ with

$$\begin{aligned} \beta_{03} &= \ln\left(\frac{q_{30}q_{41}}{q_{40}q_{31}}\sqrt{\frac{1-\rho_0^2}{1-\rho_1^2}}\right) \\ &+ \frac{-\mu_{11}^2 + 2\rho_1\mu_{11}\mu_{21} - \mu_{21}^2 + \mu_{21}^2 + \mu_{11}^2(1-\rho_1^2)}{2(1-\rho_1^2)} \\ &+ \frac{-\mu_{10}^2 - 2\rho_0\mu_{10}\mu_{20} + \mu_{20}^2 - \mu_{20}^2 - \mu_{10}^2(1-\rho_0^2)}{2(1-\rho_0^2)} \end{aligned}$$

It is of interest to consider the coefficients β_{0j} , with $j = 1, 2, 3$, further. They depend on both the spike probabilities and the means and correlation coefficients. In fitting a model, different spike variables need to be included for X_1 and X_2 . These spike variables are defined as follows:

$$\begin{aligned} \text{logit}(Y|X_1X_2) &= \alpha + \beta_{01}Z_1 + \beta_{02}Z_2 + \beta_{03}Z_3 + \beta_1X_1 + \beta_2X_2 \\ \text{with } Z_1 &= 1 \text{ if } X_1 \neq 0, X_2 \neq 0, 0 \text{ otherwise} \\ \text{and } Z_2 &= 1 \text{ if } X_1 = 0, X_2 \neq 0, 0 \text{ otherwise} \\ \text{and } Z_3 &= 1 \text{ if } X_1 \neq 0, X_2 = 0, 0 \text{ otherwise} \end{aligned} \tag{20}$$

It may be sufficient in practice to use the model with a different parameterization of the spike, which is also theoretically correct in the special case for $\rho_0 = \rho_1 = 0$

$$\begin{aligned} \text{logit}(Y|X_1X_2) &= \alpha + \gamma_{01}V_1 + \gamma_{02}V_2 + \beta_1X_1 + \beta_2X_2 \\ \text{with } V_1 &= 1 \text{ if } X_2 \neq 0, 0 \text{ otherwise} \\ \text{and } V_2 &= 1 \text{ if } X_1 \neq 0, 0 \text{ otherwise} \end{aligned} \tag{21}$$

All three indicators Z_1, Z_2 and Z_3 describe distinct proportions of the population, which might differ in their correlation structure and therefore need to be included. The relation between Z_i and V_i is given as $V_1 = Z_1 + Z_2$ and $V_2 = Z_1 + Z_3$.

A further property is of interest here. If there is a risk associated with the exposure, however independent of its level, we have from Equations (13) and (14) $f_0(x_1, x_2) = f_1(x_1, x_2)$ and $f_0(x_i) = f_1(x_i)$, respectively. The resulting OR functions reduce to

$$\text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=0} = \frac{q_{10}q_{41}}{q_{40}q_{11}}$$

and

$$\text{OR}_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \frac{q_{20}q_{41}}{q_{40}q_{21}}$$

that is, the OR depends on the spike probabilities only. These results correspond to the univariate situation given by Becher *et al.* (2012).

If the spike proportions are not independent, as in the general definition of the bivariate normal distribution with a spike, we have confounding of the two binary variables, which indicate the zero/non-zero status, and if this differs in the diseased and non-diseased, then interaction is present as well. In the real world, one can expect confounding and interaction of the binary indicators V_1 and V_2 and on the level of the positive continuous part of the variables. Therefore, proper modelling in the multivariable situation with spike variables is rather challenging.

As a result, from this section, we can state that in the case of two normally distributed variables with a spike in both, the correct model for the most general case (spikes are not independent, and the continuous parts have unequal variances and are

correlated) requires three binary variables Z_1 , Z_2 and Z_3 , the continuous variables X_1 and X_2 untransformed, X_1 and X_2 squared and the multiplicative term X_1X_2 .

3.5 Log normal distribution

Becher *et al.* (2012) showed for the univariate case that the log transformation is required if the covariate is log normally distributed. Algebraic calculations show that this also holds if two covariates have a bivariate log normal distribution. In the case of equal variances, the correct model requires the terms in $\ln(x_1^*)$, $\ln(x_2^*)$, the quadratic log-transformed part of both variables $\ln(x_1^*)^2$, $\ln(x_2^*)^2$ and a product of the log-transformed positive part $\ln(x_1^*)$, $\ln(x_2^*)$ to consider the interaction, and three binary indicators for the distinct spike proportions.

If X_1 is normally distributed and X_2 is log normally distributed, the correct model is obtained with similar algebraic calculations, using formula (1). Note that $(X_1, \ln X_2)$ then has a bivariate normal distribution. It requires the terms in $\ln(x_1^*)$, $\ln(x_2^*)$, the quadratic part of both variables $\ln(x_1^*)^2$, $\ln(x_2^*)^2$ and a product of both $x_1^* \ln(x_2^*)$ to consider the interaction, and three binary indicators for the distinct spike proportions.

3.6 Numerical analysis of the variation of regression coefficients with varying correlations

We have shown under the bivariate normal distribution for the continuous parts that the correct model requires a binary exposure indicator for each of the three distinct spike proportions, and the corresponding formula for the spike regression coefficients β_{0j} is given in the previous section in Equations (16–19). However, in this section, we investigate numerically the difference between the correct modelling of the spike with binary indicator variables Z_1 , Z_2 and Z_3 and the simpler indicator variables V_1 and V_2 . Modelling with three binary indicators enables one to consider possible correlation between covariates. The latter indicators model the data under the strong assumption of independent covariates. They are however more straightforwardly modelled and may therefore be preferable.

For numerical illustration, we consider the parameters from Table 2, that is, $\sigma_{hi}^2 = 1$ and $\mu_0 = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$, $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, with additional spike probabilities as follows: $p_{01} = 0.2$, $p_{02} = 0.2$, $p_{11} = 0.1$ and $p_{12} = 0.1$. Table 3 shows the regression coefficients of three special cases with independent spike proportions for illustration.

It is seen that the spike coefficients β_{02} and β_{03} add up to the overall theoretical spike β_{01} for $\rho_0 = \rho_1 = 0$. If $\rho_0 = \rho_1 > 0$, the coefficients β_{02} and β_{03} do not add up to the overall theoretical spike β_{01} , although the difference is not very large. For model fitting with FPs, it is therefore important to investigate numerically the difference. The OR functions (here for the second case) are

Table 3. Theoretical regression coefficients of two normally distributed variables for selected parameter combinations with three binary indicators

ρ_0	ρ_1	β_{01}	β_{02}	β_{03}	β_1	β_2
0.0	0.0	2.25	1.31	0.94	1.00	0.50
0.1	0.1	2.19	1.26	0.88	0.96	0.40
0.2	0.2	2.13	1.19	0.81	0.94	0.31

$$OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=0} = \exp(2.19 + 1.26x_1^* + 0.88x_2^*)$$

$$OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_1^*, X_2=0} = \exp(0.88 + 0.4x_2^*)$$

$$OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=x_2^*} = \exp(1.26 + 0.96x_1^*)$$

We generated a very large data set with 500 000 cases and 500 000 controls to minimize sampling variation and compare the coefficients of the spike indicators when using the two different definitions from (20) and (21) in section 3.4.2. Spike probabilities in cases and controls were chosen as aforementioned.

We first assume the spikes to be independent with $q_{11} = 0.01$, $q_{21} = 0.09$, $q_{31} = 0.09$, $q_{41} = 0.81$ and $q_{10} = 0.04$, $q_{20} = 0.16$, $q_{30} = 0.16$, $q_{40} = 0.64$. The correct model is logit $(Y | X_1, X_2) = \beta_{01}Z_1 + \beta_{02}Z_2 + \beta_{03}Z_3 + \beta_1X_1 + \beta_2X_2$, which yields regression coefficients as given in Table 4 for different values of ρ_0 , ρ_1 and σ^2 . In this case, the coefficients β_{02} and β_{03} are almost the same as γ_{01} and γ_{02} , and each pair sums up to β_{01} (columns β_{01} , $(\beta_{02} + \beta_{03})$ and $(\gamma_{01} + \gamma_{02})$). Hence, it would already be sufficient to use the parameterization as given in (20). Note that β_1 and β_2 are identical for each parameterization (Tables 4 and 5).

Second, we assume that the spikes are not independent (Table 6). There is now an increased amount of observations in the intersection denoting zeros in both variables and a reduced number of observations with value zero in only one variable. This gives an OR of 2.2 in cases and 2.6 in controls and leads to following regression coefficients shown in Table 5. Here, the coefficients β_{02} and β_{03} are not the same as γ_{01} and γ_{02} , and neither of these pairs sums up to β_{01} (columns β_{01} , $(\beta_{02} + \beta_{03})$ and $(\gamma_{01} + \gamma_{02})$).

As a result, we can state that as soon as the spike proportions of both variables X_1 and X_2 are correlated, those coefficients for the binary indicators that consider both

Table 4. Regression coefficients of two normally distributed variables for selected parameter combinations (Table 3) to compare spike coefficients of independent spike proportions defined as given in (20) and (21)

ρ_0	ρ_1	σ^2	β_{01}	β_{02}	β_{03}	γ_{01}	γ_{02}	$(\beta_{02} + \beta_{03})$	$(\gamma_{01} + \gamma_{02})$	β_1	β_2
0.0	0.0	1	2.23	1.31	0.94	1.31	0.93	2.25	2.24	1.00	0.50
0.1	0.1	1	2.19	1.29	0.91	1.29	0.91	2.20	2.20	0.96	0.42
0.2	0.2	1	2.16	1.28	0.90	1.27	0.89	2.18	2.16	0.94	0.34
0.5	0.5	1	2.10	1.28	0.83	1.28	0.83	2.11	2.11	0.96	0.10
0.0	0.0	2	1.91	1.06	0.87	1.06	0.87	1.93	1.93	0.50	0.25
0.1	0.1	2	1.90	1.05	0.86	1.05	0.86	1.91	1.91	0.48	0.21
0.2	0.2	2	1.88	1.04	0.85	1.04	0.85	1.89	1.89	0.47	0.17
0.0	0.0	3	1.81	0.98	0.85	0.97	0.85	1.83	1.82	0.33	0.17
0.1	0.1	3	1.79	0.97	0.84	0.97	0.84	1.81	1.81	0.32	0.14
0.2	0.2	3	1.79	0.96	0.84	0.96	0.83	1.80	1.79	0.31	0.11

Table 5. Regression coefficients of two normally distributed variables for selected parameter combinations (Table 3) to compare spike coefficients of dependent spike proportions defined as given in (20) and (21)

ρ_0	ρ_1	σ^2	β_{01}	β_{02}	β_{03}	γ_{01}	γ_{02}	$(\beta_{02} + \beta_{03})$	$(\gamma_{01} + \gamma_{02})$	β_1	β_2
0.0	0.0	1	2.24	1.23	0.86	1.26	0.90	2.09	2.16	1.00	0.50
0.1	0.1	1	2.20	1.21	0.84	1.24	0.88	2.05	2.12	0.96	0.42
0.2	0.2	1	2.17	1.20	0.82	1.23	0.86	2.02	2.09	0.94	0.34
0.5	0.5	1	2.12	1.21	0.76	1.24	0.80	1.97	2.04	0.96	0.09
0.0	0.0	2	1.92	0.98	0.80	1.01	0.83	1.78	1.84	0.50	0.25
0.1	0.1	2	1.90	0.97	0.79	1.00	0.82	1.76	1.82	0.48	0.21
0.2	0.2	2	1.89	0.96	0.78	1.00	0.81	1.74	1.81	0.47	0.17
0.0	0.0	3	1.82	0.90	0.78	0.93	0.81	1.68	1.74	0.33	0.17
0.1	0.1	3	1.81	0.89	0.77	0.92	0.80	1.66	1.72	0.32	0.14
0.2	0.2	3	1.80	0.89	0.76	0.92	0.80	1.65	1.72	0.31	0.11

Table 6. A 2×2 table of the spike probabilities q_{ki}

		Cases			Controls			
		X_1		0	X_1		0	>0
		0	>0		0	>0		
X_2	0	0.02	0.085	0.105	0.08	0.14	0.22	
	>0	0.085	0.81	0.895	0.14	0.64	0.78	
		0.105	0.895	1	0.22	0.78	1	

Note: Numerical example for dependent spike proportions.

continuous variables simultaneously differ from the ones defined as given in (Equation 21). Thus, in order to take a possible correlation between the spike proportions into account, the correct model requires all three binary indicators. If the difference of $\beta_{02} + \beta_{03}$ and β_{01} is relatively small, an FP procedure with two binary indicators can be used. We demonstrate this with data from the aforementioned study.

3.7 Note on other distributions

The earlier calculations can be carried out with arbitrary density functions, although depending on the complexity of the density function, this must be performed numerically. For the univariate case, this has been carried out for several other distributions, and one can show that different transformations of the covariates yield the correct model. However, for practical applications, it is preferable to select the model with a procedure based on a goodness-of-fit statistics rather than to identify a distribution that comes closest to the observed data. One option is kernel density estimation, which attempts to estimate the density directly from the data without assuming a particular form for the underlying distribution and to obtain the OR function with a kernel density estimator. However, the resulting functions are rather wiggly and non-linear depending on the chosen bandwidth, which makes it difficult to interpret it at specific positions.

Here, the FP procedure has been shown to be useful (Royston and Sauerbrei, 2008), also in the presence of a spike (Royston *et al.*, 2010; Becher *et al.*, 2012). In

the following, we illustrate model-fitting procedures with a real-data example. In a first attempt, we consider the distribution of the covariates and fit a model according to the model specification, which theoretically emerges from the assumed distributions. In a second attempt, we use the FP procedure to obtain a bivariate dose-response model.

4 Data example and model-fitting aspects

We present results on the analysis of the data from the case-control study on lung cancer. The variables of interest, smoking and number of years exposed to a ‘List A’ job (jobs with likely exposure to carcinogenic substances), are slightly correlated. Among those exposed to both variables, the Spearman correlation coefficient is 0.09, and between both variables including zeros, it is 0.20. While medians of both variables are different in cases and controls (Figure 1), the variances (after log transformation) seem rather similar. The correlation between both variables appears to be slightly higher in controls than in cases, indicating a small negative interaction.

We fit a model under the assumption that the distribution of the exposure variables is a log normal distribution with equal variances in cases and controls, possibly different correlations between the positive parts of the variable in cases and controls and an unknown relationship between the spike proportions. Here, we have in controls $(q_{10} + q_{30}) = 0.05$, $(q_{10} + q_{20}) = 0.61$ and $q_{10} = 0.035$, which is close to the product of the marginal probabilities 0.031, and positive parts were only slightly correlated. Based on the results in section 3, the correct model would include two or three binary indicators, the two log-transformed exposure variables and possibly the interaction term. These models were fitted for both variables separately in univariate analyses and jointly in multivariable analyses. The results are given in Table 7 and will be discussed together with the results of the following model-fitting procedure in section 4.2.

4.1 Modelling continuous covariates using fractional polynomials

The common procedures for model fitting are however different, because covariates rarely follow exactly a given distribution. When modelling continuous variables, it is often preferable to model the relationship allowing non-linear functions using FPs. The functional form of an FP1 is defined as $\beta_1 x^{p_1}$ with p_1 taken from the set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where $p = 0$ denotes $\ln(x)$. The functional form of the more complex and flexible FP2 function is defined as $\beta_1 x^{p_1} + \beta_2 x^{p_2}$. If $p_1 = p_2$, the functional form with so-called repeated powers is defined as $\beta_1 x^p + \beta_2 x^p \log(x)$ (Royston and Sauerbrei, 2008). An FP2 allows more flexible non-monotone dose-response functions.

The FP procedure to model one continuous variable with a SAZ was described in a modified version by Becher *et al.* (2012). We will refer to this modified version as ‘FP-spike’ throughout our manuscript. In a SAZ situation, we have an additional coefficient β_0 , which refers to the binary indicator V . The indicator

Table 7. Lung cancer case-control study: results of modelling with prespecified models and with the FP-spike procedure

Number	Variables included in final model (power value)	$\gamma_{01}(p)$	$\gamma_{02}(p)$	$\beta_{00}(p)$	$\beta_{02}(p)$	$\beta_{03}(p)$	$\beta_1(p)$	$\beta_2(p)$	$\beta_4(p)$	Dev. Diff. to null model	df
0	Null model	—	—	—	—	—	—	—	—	1391.83*	—
Univariate analyses											
1	$V_1, \log X_1$	0.17 (0.52)	—	—	—	—	0.63 (<0.001)	—	—	257.53	2
2	$X_1(1, 1)^\dagger$	—	—	—	—	—	0.18, -0.03 [‡] (<0.001)	—	—	268.57	3
3	$V_2, \log X_2$	—	-0.27 (0.06)	—	—	—	—	0.17 (0.01)	—	42.11	2
4	$V_2, X_2(1)^\ddagger$	—	-0.42 (0.001)	—	—	—	—	0.01 (0.04)	—	38.43	2
Bivariate analyses											
5a	$V_1, V_2, \log X_1, \log X_2$	0.22 (0.40)	-0.35 (0.03)	—	—	—	0.63 (<0.001)	0.10 (0.12)	—	283.90	4
5b	$V_1, V_2, \log X_1, \log X_2, (\log X_1^\dagger \log X_2)$	0.19 (0.57)	-0.35 (0.04)	—	—	—	0.63 (<0.001)	0.10 (0.13)	-0.02 (0.91)	283.92	5
6a	$Z_1, Z_2, Z_3, \log X_1, \log X_2$	—	—	-0.10 (0.73)	-0.004 (0.99)	-0.37 (0.02)	0.63 (<0.001)	0.10 (0.13)	—	284.31	5
6b	$Z_1, Z_2, Z_3, \log X_1, \log X_2, (\log X_1^\dagger \log X_2)$	—	—	1.09 (0.18)	-0.52 (0.35)	-0.80 (0.01)	0.56 (<0.001)	0.14 (0.05)	-0.01 (0.11)	286.87	6
7 [§]	$X_1(1, 1)^\dagger$	—	-0.54 (<0.001)	—	—	—	0.17, -0.03 [‡] (<0.001)	—	—	292.49	2
8 [§]	$V_2, X_2(1)^\ddagger, X_1(1, 1)^\dagger$	—	-0.46 (<0.001)	—	—	—	0.17, -0.03 [‡] (<0.001)	0.01 (0.32)	—	293.48	4

Notes: df, degree of freedom; X_1 , smoking (untransformed); X_2 , number of years exposed to a 'List A' job (untransformed); Z_1, Z_2 and Z_3 , binary indicators as defined in 3.4.2 (20); V_1, V_2 , binary indicators as defined in 3.4.2 (21); $X_1^\dagger X_2$, multiplicative interaction term; $\log X_1$, log-transformed smoking; $\log X_2$, log-transformed number of years exposed to a 'List A' job.

*Deviance of the null model.

[†] Selected fractional polynomial (FP) power values.

[‡] Fractional polynomial-spike procedure in an iterative Expectation-Maximization (EM) approach separately applied to both variables.

[§] X_1 enters the model as an FP2 with both parameters listed, that is, the OR function is $OR(X_1 = x^* \text{ vs } X_1 = 0) = \exp(0.17x_1^* - 0.03x_1^{*2} \log(x_1^*))$.

distinguishes individuals with $X=0$ from those where $X>0$. The unexposed individuals are defined as a distinct subpopulation, for which it is necessary to model the outcome explicitly. The rest of the distribution is modelled as a positive continuous variable ($X>0$) using FPs. Including the binary variable V , the model is

$$\varphi(x, \beta) = \beta_0 v + (\beta_1 f(x) + \alpha)(1 - v)$$

where φ represents the outcome depending on the type of regression model.

Theoretical justifications for the univariate situation and results of a simulation study are shown by Becher *et al.* (2012).

Building on these results, we extended the theoretical investigations for specific bivariate situations described in section 3. In real data, there are different situations concerning the correlation structures of spike variables and their influence on the outcome. To handle such situations, we propose different options to deal with the SAZ. In the easiest case, one can assume independence of the spike variables and use the univariate approach separately for each variable with some slight modifications. Further ideas are to use combinations of dummy variables, which vary the influence of observations that are positive in one variable and zero in the other one.

4.2 Comparison of direct logistic regression and fractional polynomial modelling results

We now use the FP-spike procedure to search the most appropriate model and compare the results with the direct approach.

Table 7 shows the results of the logistic regression analysis for different models. Models 1, 3, 5 and 6 result from distributional assumptions of the covariates. The graphical analysis in Figure 1 suggests that both variables might be log normally distributed. This implies that the positive values must be log-transformed before including them in the model. Both variables have a similar variance in cases and controls, indicating that a squared term is not needed, and are positively correlated, indicating confounding. However, the correlation appears a little smaller in cases than in controls, indicating little negative interaction. The spikes also seem to be independent, indicating that modelling with two spike variables V_1 and V_2 might be sufficient rather than using three spike variables Z_1 , Z_2 and Z_3 .

In comparison with the null model, both X_1 and X_2 , and V_1 and V_2 significantly improve the model fit when included separately (models 1 and 2, Table 7). This would be the correct model for both variables with neither confounding nor interaction. In model 5a, both variables and the two spike indicators V_1 and V_2 are included. This would be the correct model with confounding, without interaction. The resulting OR function estimates are $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=0, X_2=0} = \exp(0.22 - 0.35 + 0.62 \log(x_1^*) + 0.10 \log(x_2^*))$ and $OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=0} = \exp(-0.35 + 0.62 [\log(x_1^*) - \log(x_{10})] + 0.10 \log(x_2^*))$. The regression coefficient

estimates β_1 and β_2 become slightly smaller in model 5 compared with models 1 and 3, respectively, indicating moderate positive confounding. In model 6a, the three spike indicators Z_1 , Z_2 and Z_3 are included, which leads to the same conclusion concerning confounding as obtained from model 5a. The model fit is slightly better. Models 5b and 6b additionally include the common interaction term, that is, by multiplying both variables. These interaction terms are clearly non-significant.

When applying the FP-spike procedure for the original variables smoking and lifetime exposure to a 'List A' job separately, we obtain the result from models 2 and 4. Smoking is entered as an FP of second degree with powers (1, 1) without the binary indicator. The corresponding OR function is $\text{OR}(X_1 = x_1 \text{ vs } X_1 = 0) = \exp(0.18X_1 - 0.03X_1 \log(X_1))$. Lifetime exposure to a 'List A' job enters the model linearly with the binary indicator. The corresponding OR function for any dose $x > 0$ is $\text{OR}(X_1 = x \text{ vs } X_1 = 0) = \exp(-0.42 + 0.01X_2)$.

The best-fitting models are obtained in models 2, 5a, 5b, 6a and 7. Although the deviances are almost the same, models 2 and 7 would be preferably chosen because of the lower degree of freedom. Furthermore, it needs to be considered that the degrees of freedom between the FP models and the models chosen by regular logistic regression are not directly comparable as we already log-transformed the continuous variables before applying logistic regression. Therefore, the difference between the degrees of freedom of both methods is rather conservative. In models 2 and 7, we did not pretransform the covariates as obtained from a given distributional assumption. The transformed data have only an approximate normal distribution with a right tail of the data, which usually is too long. In the iterative approach (model 7), we kept the result from the FP fitting procedure with one variable (model 2 or 4) while fitting the best FP model for the other variables. Here, the improvement of fit was small, and the occupational exposure variable was not selected in the procedure. Starting with either variable, the result is identical in this example. The advantage of FP modelling is that the initial value is modelled directly without the restrictions of a distributional assumption. For illustration purposes, we also fitted model 8 wherein the selected transformations from model 4 were added into the final model 7. The deviance difference of this model was 293.48. The corresponding regression coefficients are -0.46 (V_2) and 0.01 (X_2). When testing both separately for removal, neither test is significant. When testing both for removal simultaneously, occupational exposure has a significant independent effect.

In Figure 4a, we display the dose-response functions for smoking as given in Table 7. Model 5a is a model from a multivariable fit where smoking was adjusted for lifetime exposure to a 'List A' job. In model 2, the spike was dropped from the final model because of deviance criteria. The diamond represents the coefficient for the binary indicator, that is, the non-smokers, in model 5a. The bubbles represent the categorized OR per group of 3 pack-years to the reference of zero. The size of the bubbles varies depending on the number of individuals per category and approximates the original data. The results of the second risk factor, lifetime

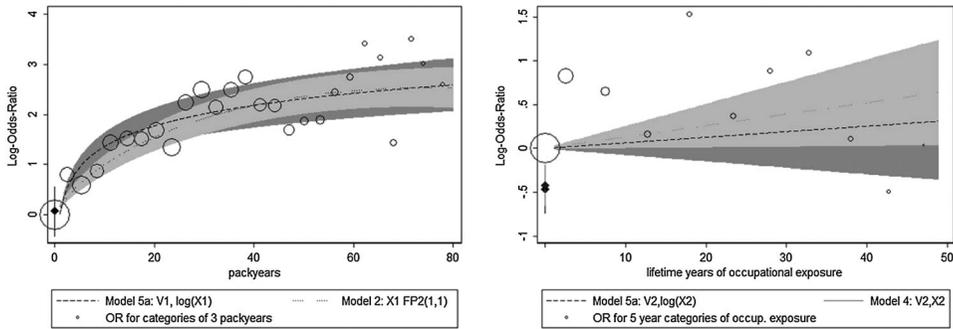


Fig. 4. Lung cancer case-control study: visual comparison of dose-response functions as given in Table 7 for (a) smoking (models 5a and 2) and (b) lifetime years of occupational exposure (models 5a and 4).

exposure to a ‘List A’ job, are displayed in Figure 4b. Model 5a is the same model from a multivariable fit as used in Figure 4a where we plot lifetime exposure to a ‘List A’ job adjusted for smoking. In models 4 and 5a, the binary indicator and X_2 were kept in the final model. In model 5a, neither the binary indicator nor the log-transformed X_2 was significant. In the univariate model 4, both variables were significant. The upper diamond represents the coefficient for the binary indicator, that is, the non-smokers, in model 5a. The lower diamond represents the coefficient for the binary indicator, that is, the non-smokers, in model 4. The bubbles represent the categorized OR per group of 5 years of lifetime occupational exposure to the reference of zero. The size of the bubbles varies depending on the number of individuals per category.

5 Discussion

This paper provided a contribution towards a joint dose-response modelling of covariates with a SAZ. We have presented some theoretical results on the OR functions for a binary response variable within the logistic regression model for two normally distributed variables with and without spikes at zero. Furthermore, we have compared the performance of two modelling approaches in a practical data example from a lung cancer case-control study.

The results have shown that even the presumably simple case of two bivariate normally distributed cases 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 joint distribution of the zero cells (4CD) of two SAZ variables has effects on two levels, the 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 may be necessary to include up to three binary indicators into the model, as zero observations in one, the other and both SAZ

variables. For more than two SAZ variables, the situation becomes even more complicated. A practical issue, however, is whether for real data such a complicated model, even theoretically justified, is required or whether a simpler model is sufficient to describe the data.

The data example shows that it is difficult to derive a suitable dose–response curve based on assumptions of the distribution, which were made in advance. Observed data usually do not exactly follow common continuous distributions. Our approach allows modelling in the class of FP functions and uses a function selection procedure to determine a function, which fits the data best within the flexible class of FP functions. In our data example, we have a distribution that is close but certainly not exactly a log normal distribution. The model that is based on this distributional assumption did not give a fit as good as the one derived without prespecified distributional assumptions. The FP-spike procedure, on the other hand, selected a model that gave a better fit, as seen, for example, from the deviance in model 1 in comparison with model 2.

Summing up, modelling unexposed individuals as a separate risk group has been shown to be useful and preferable to modelling semicontinuous variables without considering binary indicators. The developed methods are consistent with the theoretical solution when method-specific assumptions hold. Further work is underway to investigate the performance of the FP-spike procedure in the bivariate situation in different regression models with simulation.

Acknowledgements

E. Lorenz was supported by the German Research Foundation (DFG), BE 2056/10-1. C. Jenkner was supported by the German Research Foundation (DFG), SA 580/7-1. We thank Dr Hermann Pohlabein, Prof. Wolfgang Ahrens and Prof. Karl-Heinz Jöckel for providing the data of the lung cancer study and a reviewer for detailed comments that helped to clarify several issues and to improve the presentation of the paper.

Conflict of interest

The authors have declared no conflict of interest.

References

- AHRENS, W., JÖCKEL, K.-H., BROCHARD, P., BOLM-AUDORFF, U., GROSSGARTEN, K., IWATSUBO, Y., ORLOWSKI, E., POHLABELN, H. and F. BERRINO (1993), Retrospective assessment of asbestos exposure-I. Case–control analysis in a study of lung cancer: efficiency of job-specific questionnaires and job exposure matrices, *International Journal of Epidemiology* **22**, 83–95.
- BECHER, H. (1992), The concept of residual confounding in regression models and some applications, *Statistics in Medicine* **11**, 1747–1758.

- BECHER, H., LORENZ, E., ROYSTON, P. and W. SAUERBREI (2012). Analysing covariates with spike at zero: a modified FP procedure and conceptual issues, *Biometrical Journal* **54**, 686–700. DOI: 10.1002/bimj.201100263
- DE BOER, C. (2001). *A practical guide to splines*, revised edn, Springer, New York.
- DREASSI, E., PETRUCCI, A. and E. ROCCO (2013). Small area estimation for semicontinuous skewed spatial data: an application to the grape wine production in Tuscany, *Biometrical Journal* **56**(2014) 1, 141–156.
- FLETCHER D., MACKENZIE D., E. VILLOUTA (2005). Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics* **12**, 45–54.
- JÖCKEL, K.-H., POHLABELN, H., AHRENS, W. and M. KRAUSS (1998). Environmental tobacco smoke and lung cancer, *Epidemiology* **9**, 672–675.
- OLSEN M. K., J. L. SCHAFFER (2001). A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association* **96**, 730–745.
- ROYSTON, P. and W. SAUERBREI (2008). *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester.
- ROYSTON, P., SAUERBREI, W. and H. BECHER (2010). Modelling continuous exposures with a ‘spike’ at zero: a new procedure based on fractional polynomials, *Statistics in Medicine* **29**, 1219–27. DOI: 10.1002/sim.3864
- VINK, G., LAURENCE, E. F., PANNEKOEK, J. and S. VAN BUUREN (2014). Predictive mean matching imputation of semicontinuous variables, *Statistica Neerlandica* **68**, 61–90.
- YANG Y.-W., C. Y. FU (2013). Two advanced methods for adjusting the main coefficient in logistic regression. *Computational Statistics* **28**, 199–218.

Received: 20 May 2014. Revised: 19 January 2015.