Research Article

Statistics in Medicine

Received 3 March 2009, Accepted 12 January 2010 Published online 26 February 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3864

Modelling continuous exposures with a 'spike' at zero: A new procedure based on fractional polynomials

Patrick Royston[a]*†, Willi Sauerbrei[b] and Heiko Becher[c]

A common task in epidemiology is to estimate the dose–response function for a continuous exposure. Often a proportion of subjects is unexposed. Typical examples are cigarette consumption, alcohol intake, or occupational exposures. The question arises as to how to model such variables statistically. The fractional polynomial method of modelling continuous exposure variables is extended to allow for a proportion unexposed. A binary variable for the unexposed fraction is added to the model. In a two-stage procedure, we assess whether the binary variable and/or the continuous function for the exposed individuals is required for a good fit to the data. Extension to the multivariable situation is described. Three data sets with different characteristics are used as illustrations. The analyses of the three studies using the proposed procedure give differing results. In one example, only the binary variable seems to be required. In the other two examples, the binary variable and fractional polynomial functions of the exposure variable are needed. One function is monotonic and the other has a minimum. In the third example, adjusting for confounders has almost no effect on the function selected. In conclusion, the new procedure offers a worthwhile extension of dose–response modelling with an unexposed fraction. It is simple to carry out with standard software. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: dose–response model; fraction unexposed; fractional polynomials; confounders; regression modelling

1. Introduction

A common task in epidemiology is to estimate the dose–response function for a continuous exposure. Subject matter knowledge, if available, can and should be incorporated in the type of function chosen. Monotonicity is a popular and appropriate assumption in many cases. In particular cases, it may even be possible to consider an alternative metric which may have better biological properties. For example, Thurston *et al*. [1] investigate the influence of smoking on lung cancer. Several continuous smoking metrics such as pack years, smoking duration or smoking intensity are available. Non-smokers are always the same separate group in these metrics. Because of biological considerations, it may be useful to derive a new dose metric by combining the individual metrics in a suitable way. The resulting metric in Thurston *et al*. [1] allowed non-smokers and smokers to be included in the same model without the need to adjust for smoking status. Another approach to the same issue is presented by Leffondré *et al*. [2].

However, in many situations such approaches are infeasible. Furthermore, subject matter knowledge is limited and does not indicate specific functional form(s). In such situations the function must be determined in a data-dependent fashion. For this task, we propose to use fractional polynomials (FPs) [3–5]. However, as is obvious from the smoking example, there is often a proportion of subjects unexposed for a continuous exposure x of interest. Other typical examples are alcohol intake and occupational exposures. This phenomenon may occur in other settings, e.g. number of positive lymph nodes in primary breast cancer, number of tumour blood vessels, etc. The distribution of x therefore has a 'spike' or probability mass at zero, i.e. $\Pr(x=0)>0$.

General FP methodology does not address the question of how to model the factor when there is a spike at zero. Theoretical results [6, 7] in the logistic regression setting indicate that a binary variable, say z, representing exposure/non-exposure should

[a] Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, 222 Euston Road, London NW1 2DA, U.K.
[b] IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100 Freiburg, Germany
[c] Epidemiology and Biostatistics Unit, Medical Faculty, University of Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany
* Correspondence to: Patrick Royston, Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, 222 Euston Road, London NW1 2DA, U.K.
† E-mail: pr@ctu.mrc.ac.uk

Contract/grant sponsor: U.K. Medical Research Council; contract/grant number: U.1228.06.001.00002.01

be included in the model. However, the dose–response function for the exposed individuals still needs to be determined. For univariate dose–response modelling, Royston and Sauerbrei [5] suggested a new approach to the problem by extending FP modelling to include z, and illustrated it using one example in linear regression. The procedure comprises two stages: first, to determine the best FP function when z is included in the model; second, to assess whether z or the FP component can be eliminated without harming the model fit. Here we outline the methodology, extend it to multivariable modelling and further illustrate it with three real examples with different characteristics, including two case-control studies in cancer.

Section 2 introduces the examples. Section 3 describes the basics of FP modelling, and explains 'FSP-spike', the procedure to select a function for a continuous exposure with a spike at zero. We also describe the extension to the multivariable situation. Section 4 gives the results of applying FSP-spike to the three examples, and Section 5 is a discussion.

2. Data

2.1. Cigarettes and lung cancer

A case-control study of lung cancer in males was performed in Poland [8] with 600 cases (squamous cell, small cell and adenocarcinoma) and 1343 controls frequency matched for age. A summary of smoking exposure and outcome is given in Table I. A large number of controls (21.5 per cent) and only a few cases (2.7 per cent) had zero exposure. The original data file was not available, hence, in this example we use the smoking variable only from the published table. Individual doses were randomly assigned within a dose category. For the upper open-ended category, we simply set 80 cigarettes/day as the upper limit for the simulation. The small number of smokers with missing exposures were randomly assigned to dose categories.

2.2. Alcohol and breast cancer

The second example is based on a case-control study of pre-menopausal breast cancer, which was performed in south-west Germany in 1992–1995 [9]. The aim was to investigate the relation between alcohol consumption and breast cancer. The study had 706 cases and 1381 controls frequency matched by age and study region. Table II gives the average daily alcohol consumption by dose category, with adjusted odds ratios by category indicating a non-monotone dose–response relation. A large number of cases and controls reported zero exposure.

Table I. Distribution of smoking dose (number of cigarettes/day), lung cancer case-control study, males, Poland, and univariate odds ratios for dose categories.

Number of	Controls		Cases			
cigarettes per day	*n*	Per cent	*n*	Per cent	Odds ratio	95 per cent CI
0 (non-smokers)	289	21.5	16	2.7	1.00*	—
1–9	78	5.8	8	1.3	1.85	0.66, 4.79
10–19	247	18.4	73	12.2	5.34	2.98, 10.07
20–29	459	34.2	273	45.5	10.74	6.32, 19.44
30–39	184	13.7	123	20.5	12.07	6.85, 22.42
40+	86	6.4	107	17.8	22.47	12.32, 42.66
	1343	100.0	600	100.0		

*Reference category.

Table II. Distribution of alcohol consumption and univariate odds ratios for breast cancer among participants in a population-based case-control study, Germany, 1992–1995. In the original paper, adjusted odds ratios were presented.

Number of	Controls		Cases			
intake (g/day)	*n*	Per cent	*n*	Per cent	Odds ratio	95 per cent CI
0 (non-drinker)	239	17.3	153	21.7	1.00*	—
1–5	577	41.8	257	36.4	0.70	0.54, 0.90
6–11	295	21.4	124	17.6	0.66	0.49, 0.89
12–18	150	10.9	69	9.8	0.72	0.50, 1.03
19–30	84	6.1	59	8.4	1.10	0.73, 1.65
31+	36	2.6	44	6.2	1.91	1.14, 3.20
	1381	100.0	706	100.0		

*Reference category.

2.3. Gleason score and prostate cancer

Stamey *et al*. [10] studied potential predictors of prostate-specific antigen (PSA) in 97 of 102 patients with surgically treated adenocarcinoma of the prostate. The aim was to see which several factors or combinations of them were associated with a raised PSA level. All observations were made around the time of radical prostatectomy. The percentage of cells in the prostate biopsy sample having Gleason score 4 or 5 was used in the analysis. Of the available variables, we consider only percent Gleason score 4 or 5 because it exhibits the spike at zero feature. The outcome variable is log PSA. See Reference [5] for further details.

3. New FP-based procedure

In this section, we first briefly describe the standard FP function selection procedure (FSP), in which x is assumed to take positive values. We then outline the new FP-spike procedure for cases in which x has a significant proportion of zero exposures. Finally, we consider the multivariable situation.

3.1. FP and functional selection procedure

Suppose that we have an outcome variable y, a positive continuous covariate $x>0$, and a suitable regression model relating them. The starting point is the straight line model, $\beta_0+\beta_1 x$, but other models must be investigated for possible improvements in fit. To simplify the description, we ignore the constant term β_0 in this section, although it is of course a component of most regression models. For example, all linear regression and logistic regression models have a β_0, whereas it is not a part of the linear predictor in the Cox model.

A simple extension of the straight line model is a power transformation model, $\beta_1 x^p$, often used by practitioners in an *ad hoc* way with different choices of p. Royston and Altman [3] formalized the model by calling it a first-degree fractional polynomial or FP1 function. The powers p are chosen from a restricted set, $S=\{-2,-1,-0.5,0,0.5,1,2,3\}$, where x^0 denotes $\log x$. The set S includes no transformation ($p=1$) and the reciprocal, logarithmic, square root and square transformations.

In some cases, FP1 functions may not be flexible enough. For example, FP1 functions cannot represent a curve with a maximum or minimum. Extension to the more complex and flexible two-term FP2 functions is straightforward. FP2 functions with powers (p_1,p_2) taken from S are defined as $\beta_1 x^{p_1}+\beta_2 x^{p_2}$. When $p_1=p_2$, a special type of 'repeated powers' FP function is obtained from the mathematical limit of $\beta_1 x^{p_1}+\beta_2 x^{p_2}$ as $p_1 \to p_2$, namely $\beta_1 x^{p_1}+\beta_2 x^{p_1}\ln x$ (see Reference [3]). An FP2 function can be monotonic or can have at most one maximum or minimum. Let m denote the number of FP terms or the degree of FP (e.g. $m=2$ for an FP2 function). Extension to higher-order FPs ($m>2$) is possible [3], but rarely useful in epidemiological applications [4].

To fit an FP1 model each of the 8 values of p is tried. The best-fitting model is the one whose p minimizes the deviance (that is, minimizes minus twice the log likelihood). Similarly, the best fitting FP2 function is the one among the 36 combinations of powers from S (including 8 repeated powers) that minimizes the deviance.

The function selection procedure (FSP) was designed to determine the 'best' function from the FP class [5, Chapter 4]. Before applying the procedure, the user must decide on the significance level (α) and on the degree (m) of the most complex FP model allowed. Typical choices, also used here, are $\alpha=5$ per cent and FP2 ($m=2$).

In the following description of the FSP, a linear function is assumed as the default for x. The procedure runs as follows:

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' at the α level. 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 the procedure.

The test at step 1 is of overall association of the outcome with x. The test at step 2 examines the evidence for non-linearity. The test at step 3 chooses between a simpler or more complex non-linear model.

The FP2 model has 1 d.f. for each of the two power terms and 1 d.f. for each of the two βs, altogether 4 d.f. Asymptotically, the deviance difference between the best FP2 and the null model has 4 d.f. In steps 2 and 3, the d.f. are 3 and 2 because differences are from the linear model (1 d.f.) or the best FP1 model (2 d.f.). It can be shown that the true d.f. of an FP2 model is somewhat lower than 4, but for practical applications the approximation is sufficient [5, Section 4.9.1].

To handle particular situations, e.g. x should be in the model on *a priori* grounds, simple modifications may be made [5].

3.2. FSP-spike procedure

Assume that $x\geqslant 0$ for all individuals. In order for FP functions of x to be defined at $x=0$, required also by the second stage of testing (see below), the origin of x is shifted by adding a small constant, c, before analysis. By default, we take c as the smallest difference between successive positive values of x [3], but other choices are possible [5, p. 79]. Consider a model whose linear predictor, η, is given by

$$\eta=\begin{cases}\beta, & x=0\\ \beta_0+\text{FP2}(x+c;\ p_1,p_2), & x>0\end{cases},$$

where p_1 and p_2 are powers from the standard set S of FP transformations. The linear predictor η is an FP2 function of $x+c$ when $x>0$ and a constant (β) when $x=0$. Thus η is a discontinuous function of x with a possible jump at $x=0$. The expression for η is equivalent to

$$\eta=\beta_0+(\beta-\beta_0)z+(1-z)\mathrm{FP2}^{+}(x+c;\ p_1,p_2)$$

where

$$z=\begin{cases}1, & x=0\\ 0, & x>0\end{cases}, \tag{1}$$

$$\mathrm{FP2}^{+}(x+c;\ p_1,p_2)=\begin{cases}0, & x=0\\ \mathrm{FP2}(x+c;\ p_1,p_2), & x>0\end{cases}. \tag{2}$$

The FSP-spike procedure for selecting a model has two stages. In the first stage, the most complex model comprising z and $\mathrm{FP2}^{+}(x+c;p_1,p_2)$ is compared with the null model on 5 d.f. (4 d.f. from the FP2 model plus one from the binary z term). If the test is significant, the steps of the FSP for selecting an FP function are followed, but with z always included in the model. In the second stage (performed separately), z and the remaining FP or linear component are each tested for removal from the model. If both parts are significant, the final model includes both; if one or both parts are non-significant, the one with the smaller deviance difference is removed. In the latter case, the final model comprises either the binary dummy variable or the selected FP function. If only an FP function is selected, the spike at zero plays no further part. Since the selection of an FP function may be affected by the presence of the binary dummy variable, the resulting model may differ from that from a standard FP analysis.

3.3. Multivariable case

So far we have considered assessing the effect of an exposure with a spike at zero in univariate models. In reality, however, possible confounders must be considered. The confounder model can easily be determined by using the multivariable FP (MFP) procedure [5, 11]. MFP combines selection of variables (where appropriate) by backward elimination with selection of functions using the FSP for each continuous predictor. For variables with a spike at zero, FSP is replaced with the first stage of FSP-spike. (This is achieved in Stata using the `mfp` command with the `catzero()` option.) The second stage must be done separately for all spike variables after MFP has completed.

With a single exposure x of interest, the first approach is to let MFP determine the model from the candidate variables, including x with the corresponding binary variable z in the model and selecting the potential confounders according to their statistical significance. MFP automatically excludes unimportant variables and where necessary approximates functional relationships for continuous variables.

A second approach is for MFP to select the confounder model independently of x and then apply FSP-spike (univariately) to x, always with adjustment for the selected confounders.

4. Results

The three data sets illustrate different situations. The first two are case-control studies analyzed by logistic regression, whereas the third has a continuous outcome variable requiring linear regression. In the lung cancer example, the percentages of zero values of cigarette consumption differ considerably between cases (2.7 per cent) and controls (21.5 per cent). By contrast, in the breast cancer example the percentages are similar (21.7 per cent among cases, 17.3 per cent among controls). In the prostate cancer example, the percentage of patients with Gleason score not equal to 4 or 5 is large (36.1 per cent).

4.1. Smoking and lung cancer

Table III gives details of the FSP-spike analysis of the lung cancer data. Since cigarette consumption is recorded in units of 1 cigarette per day, the small constant c is taken as 1. For models including the binary variable z, FP transformations with powers -0.5 and $(-2,-1)$ give the best fit within the FP1 and FP2 classes, respectively. The most complex model ($\mathrm{FP2}^{+}+z$) fits substantially better than the null model in which cigarette smoking has no influence on the risk of lung cancer. The deviance difference is 226 ($P<0.001$). The $\mathrm{FP2}^{+}+z$ model also fits much better than the $\mathrm{linear}^{+}+z$ model, but its deviance is similar to that of $\mathrm{FP1}^{+}+z$ (2176.4 vs 2177.0, see Table III). The latter model is therefore selected in the first stage.

In the second stage, we investigate whether either of the components ($\mathrm{FP1}^{+}$ or z) can be removed without harming the fit. The results clearly show that this is not the case, confirming preference for the $\mathrm{FP1}^{+}+z$ model.

The solid line in Figure 1 shows the strong relationship between the odds ratio of lung cancer mortality and cigarette smoking, as estimated from the $\mathrm{FP1}^{+}+z$ model. The reference category is non-smokers (i.e. $x=0$). The figure also shows the curve estimated using a standard FP analysis, ignoring any special role for the zero (non-smoker) category. The FSP selects an FP2 function with powers $(-1,-1)$. The two models have similar deviances (2177.0 for $\mathrm{FP1}^{+}+z$, 2176.8 for FP2) and result in almost identical curves.

Table III. Lung cancer data. Analysis of cigarette consumption (x) with a spike at zero. z is a dummy variable indicating non-smokers. See text for details.

Model	Deviance	Dev. diff.	d.f.	P	Power (s)
First stage					
Null	2402.1	225.7	5	<0.001	—
Linear$^+$+z	2195.4	19.0	3	<0.001	1
FP1$^+$+z	2177.0	0.6	2	0.76	−0.5
FP2$^+$+z	2176.4	—	—	—	−2, −1
Second stage					
FP1$^+$+z	2177.0	—	3	—	−0.5
FP1$^+$ [dropping z]	2384.9	208.0	1	<0.001	−0.5
z [dropping FP1$^+$]	2259.4	82.4	2	<0.001	

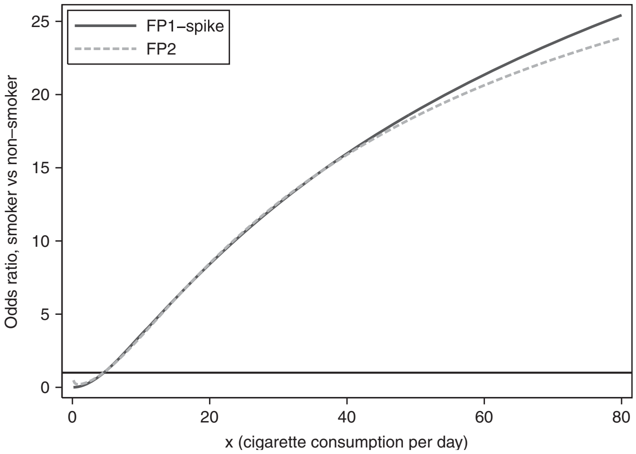

Figure 1. Lung cancer data. Fitted odds ratios for the risk of lung cancer from FP1-spike and standard FP2 models plotted against cigarette consumption. The reference category is non-smokers. The functions are plotted from the minimum positive values of x.

The fitted functions (standard errors in parentheses) are as follows:

$$\log OR = \begin{cases} \text{FP1-spike:} & -2.89(0.26)z + (1-z)[1.49(0.25) - 10.31(1.25)(x+1)^{-0.5}] \\ \text{FP2:} & 0.87(0.19) - 3.76(0.31)(x+1)^{-1} - 9.96(1.36)(x+1)^{-1}\log(x+1) \end{cases}$$

Neither of the models is preferred on statistical grounds.

4.2. Alcohol and breast cancer

Table IV gives details of the FSP-spike analysis of the breast cancer study. Since values of alcohol consumption were rounded to the nearest integer, the small constant c is taken as 1. A comparison between the null model and the selected FP2$^+$+z model shows a highly significant ($P<0.001$) association between alcohol consumption and case-control status. The most complex model (FP2$^+$+z) also fits significantly better ($P<0.05$) than both the linear$^+$+z and FP1$^+$+z models. If a more stringent significance level, say $P<0.01$, had been applied, the linear$^+$+z model would have been chosen.

Simplifying the first stage (FP2$^+$+z) model in the second stage is not possible here, since dropping either the z or the FP2$^+$ terms result in a significant ($P<0.001$) worsening of the fit.

Figure 2 shows the fitted odds ratio of breast cancer, comparing drinkers with non-drinkers. Fitted curves above 50 g/day have been truncated; the curves tend to diverge, but are based on less than 1 per cent of the distribution of alcohol consumption. According to the best-fitting FP2$^+$+z model, a small intake of alcohol (up to 20 g/day) is associated with a slightly reduced risk of breast cancer. With a standard FP analysis an FP2 model with different powers of (0, 0.5) is selected. According to the deviance criterion, its fit is only slightly worse (2636.2 vs 2635.4) than the FP2$^+$+z model. The two fitted curves are very similar, the odds ratio from the standard FP2 model being somewhat larger at high alcohol consumption levels. The fitted functions for x=alcohol

Table IV. Breast cancer data. Analysis of alcohol consumption (x) with a spike at zero. z is a dummy variable indicating non-drinkers. See text for details.

Model	Deviance	Dev. diff.	d.f.	P	Power (s)
First stage					
Null	2670.9	35.5	5	<0.001	—
Linear^{+} $+z$	2644.1	8.7	3	0.033	1
FP1^{+} $+z$	2642.5	7.1	2	0.028	2
FP2^{+} $+z$	2635.4	—	—	—	−0.5, 0.5
Second stage					
FP2^{+} $+z$	2635.4	—	5	—	−0.5, 0.5
FP2^{+} [dropping z]	2661.3	24.9	1	<0.001	−0.5, 0.5
z [dropping FP2^{+}]	2665.2	29.8	4	<0.001	

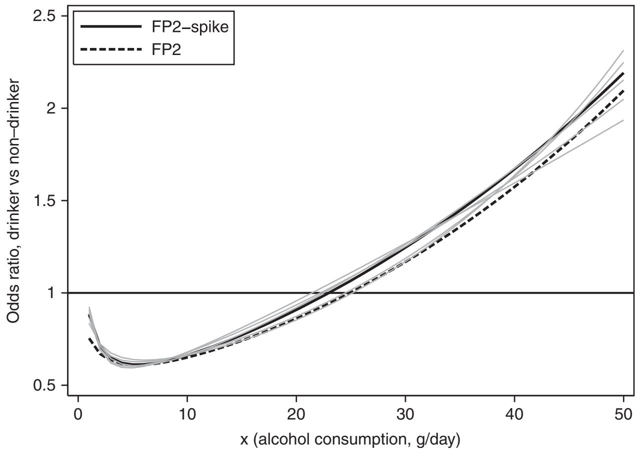

Figure 2. Breast cancer data. Thick dark lines show the fitted odds ratios for the risk of breast cancer from FP2-spike and standard FP2 models plotted against daily alcohol consumption. Thin pale lines show the 5 next-best-fitting FP2-spike models. The reference category is non-drinkers. The functions are plotted from the minimum positive value of x to $x=50$.

consumption (standard errors in parentheses) are as follows:

$$\log OR=\begin{cases}\text{FP2-spike:} & -0.45(0.10)z+(1-z)[-3.09(0.52)+2.71(0.70)(x+1)^{-0.5}+0.43(0.08)(x+1)^{0.5}]\\ \text{FP2:} & -1.02(0.10)-0.77(0.14)\log(x+1)+0.62(0.11)(x+1)^{0.5}\end{cases}$$

A general phenomenon is that several FP2 models have a deviance that is similar to the best FP2 model. The thin pale lines in Figure 2 show the five best-fitting FP2-spike models. Although they all have different power terms, the fitted functions are very similar. Note that the power terms themselves are not interpretable.

The potential confounders available for a multivariable analysis are age, parity, total length of breastfeeding, education, menopausal status and family history of breast cancer. All of these variables were adjusted for in the original analysis [9]. Using a significance level of 0.2, only breastfeeding (linear function) ($P=0.005$) and family history ($P<0.001$) were significant, irrespective of whether alcohol consumption (x) was included (first approach) or not (second approach). For both approaches, the same function, namely FP2 with powers (−0.5, 0.5), was selected as in the univariate case just described. Because of the adjustment by breastfeeding and family history, the regression coefficients and therefore the function for the FP2 model changed only slightly.

4.3. Gleason score and prostate cancer

We apply the FSP-spike procedure to percent Gleason score 4 or 5 (x) in the prostate cancer data as a predictor of log PSA (y). Figure 3 shows the original data for the 97 patients in the study. Because of the large number of zeros, a suitable analysis of log PSA is not immediately apparent. One possible approach is linear regression, but the model does not fit well. For positive values of x, y and x are almost uncorrelated, suggesting a simple model with a binary variable z indicating $x=0$. The mean value of y is smaller for the zero group (1.74 vs 2.90). The FSP selects a log function. The latter shows a steep increase in fitted y for

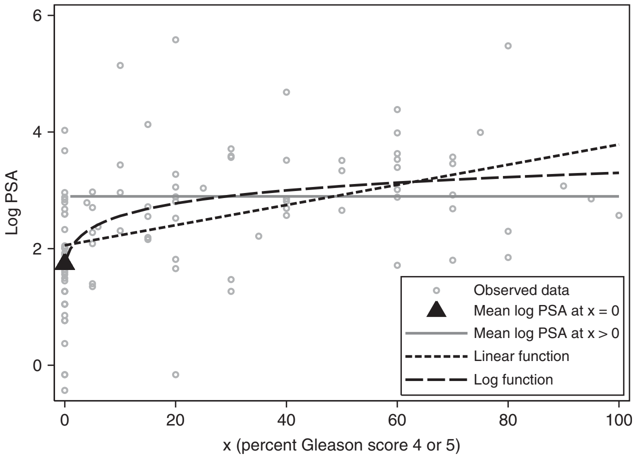

Figure 3. Prostate cancer data. Observed values of log PSA (y) and fitted linear and log functions of per cent Gleason score 4 or 5 (x). Large triangle and horizontal line show the mean value of y for $x=0$ and $x>0$, respectively.

Table V. Prostate cancer data. Analysis of percentage of cells with Gleason score 4 or 5 (x), with a spike at zero. z is a dummy variable indicating $x=0$. See text for details.

Model	Deviance	Dev. diff.	d.f.	P	Power (s)
First stage					
Null	302.1	29.8	5	<0.001	—
Linear$^{+}+z$	273.7	1.4	3	0.73	1
FP1$^{+}+z$	272.7	0.4	2	0.84	−0.5
FP2$^{+}+z$	272.3	—	—	—	1, 3
Second stage					
Linear$^{+}+z$	273.7	—	2	—	
Linear^{+} [dropping z]	282.7	9.0	1	0.003	
z [dropping Linear^{+}]	276.2	2.5	1	0.12	

values of x near zero and a very small trend for larger x values (Figure 3). Variance explained (R^2) for these three models is 17.8 per cent (linear), 23.4 per cent (binary) and 25.9 per cent (log).

With the FSP-spike procedure, a model with z and a linear function of $x+c$ for positive x is selected in the first stage (Table V) Because Gleason score increases in steps of 1, the small constant c is taken as 1. The test of FP2$^{+}+z$ versus null is highly significant ($P<0.001$). The next test, FP2$^{+}+z$ versus Linear$^{+}+z$, has $P=0.73$. The procedure terminates and the selected model from the first stage is therefore Linear$^{+}+z$. At the second stage, dropping the selected linear term does not significantly worsen the fit ($P=0.12$), whereas dropping z is highly significant ($P=0.003$). Therefore, the FSP-spike procedure selects a model with a binary factor for $x=0$. The coefficient for z is -1.16 (SE 0.21).

In Figure 4, the three selected functions from the first stage of FSP-spike are plotted, together with the mean of y for $x=0$ and $x>0$. There is no evidence that any of these more complex functions fit the data substantially better than the binary model. Their R^2 values lie between 25.4 and 26.5 per cent.

5. Discussion

We describe a procedure to deal with the spike at zero problem which commonly arises in epidemiology and other fields. As suggested previously [6, 7] and used without theoretical justification [8], a binary indicator z of positivity of an exposure x is added to the model. The data from Reference [8] are used in the first example here. In our approach, the positive component is modelled using FP functions, selected systematically using the FSP. Adding a small constant before performing the analysis allows us to use all the data in each step of the procedure. The FSP involves fitting an FP2 model to assess whether x is associated with the outcome, and if so, attempts to simplify the model. FSP-spike is an extension of the FSP to include the binary variable, z. As described in Section 3, the final model is determined in two stages. In the multivariable situation, extension is straightforward by using the MFP approach [5, 11]. Since FSP-spike is an extension of the FSP, analysis of various types of outcome, such as time-to-event data, requires nothing new. In principle, if even greater flexibility were desired, the FP component could be replaced with other functions, such as splines.

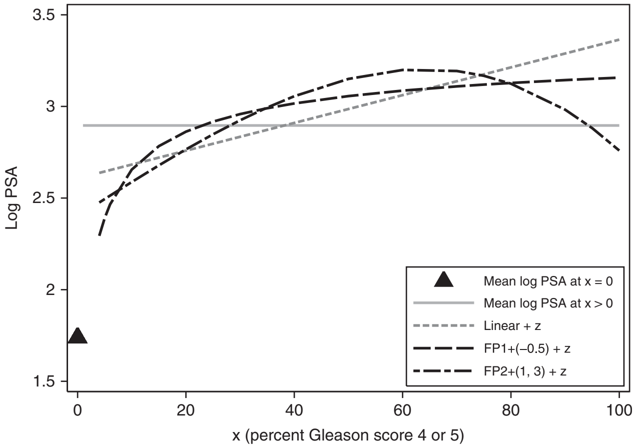

Figure 4. Prostate cancer data. Fitted linear and FP-spike functions for per cent Gleason score 4 or 5 (x) with a dummy variable (z) for zero values. $\mathrm{FP1}^{+}$ and $\mathrm{FP2}^{+}$ are FP functions estimated from positive values of x (see text). Large triangle and horizontal line show the mean value of y for $x=0$ and $x>0$, respectively. Note that the vertical scale is enlarged compared with Figure 3.

To illustrate potential applications of FP-spike, we used the procedure to reanalyse three studies with different characteristics and give differing results. The first two examples are case-control studies. The cases in the lung cancer study have a much lower proportion of zeroes. The third study involves a continuous outcome variable and is analysed using linear regression. In the lung cancer study, the first stage clearly indicates that cigarette consumption has a non-linear effect; a monotonic FP1 function describes the influence of the exposure well. The binary indicator emphasizes the role of the non-smokers. The deviances from the second modelling stage show that neither the FP1 component nor the binary indicator can be omitted. The standard FP approach selects an FP2 function. Deviances and functional forms are very similar between the two models, but the interpretation of the FP-spike model seems much more natural.

In the breast cancer study, a non-monotonic FP2 function was selected. The deviance is only slightly lower than that from a model including a linear effect (plus binary indicator) in the first stage, but the functional forms and corresponding interpretations are very different. Here, a standard FP approach selects a very similar FP2 model. The FP method detects a slight decrease in risk for women with a low alcohol consumption followed by a non-linear increase with more alcohol, but it is unimportant whether the version with or without the spike component is used.

The prostate cancer study illustrates another situation. Plotting the data does not clearly indicate whether a simple binary variable is suitable, which (if any) curve shape represents the functional relationship between x and y, or whether a binary indicator plus a function improves the fit. A standard linear regression model gives a bad fit, standard FP methodology selects a log function, whereas the second stage of the FP-spike approach eliminates the FP component selects a simple model comprising only a binary indicator, z. Obtaining the binary model within a formalized framework may reassure the analyst that no better model has been overlooked.

As in all model-building strategies, the chosen significance level critically affects the final model. In the breast cancer study, for example, had we chosen 1 per cent as the significance level then FSP-spike would have selected linear^{+} $+z$ as the final model, instead of the non-monotonic $\mathrm{FP2}^{+}+z$ function. If a monotonic function is required, the procedure may be modified by allowing an FP1 function to be the most complex, rather than FP2. This has the additional advantage of increasing the power to select a non-linear function. Another possible case is rare exposures, where the limited sample size confers low power. In this situation one may consider restricting the model choice to FP1 functions or even only to log or linear functions.

The decision to use a model including z as just described, or to work within the standard FP class, is best made on subject-matter grounds rather than by considering the fit of functions with or without z. Judgment of the relative merits of the two approaches is needed. For example, in some cases a discontinuous function may make no scientific sense. Since in an $\mathrm{FP}^{+}+z$ model the FP function is not 'anchored' by the data at $x=0$, the function may be unstable and poorly estimated for small positive x. When the proportion of zero values is small, it may be preferable to ignore the special role of the zero subset and apply the standard FSP to x. In the first two examples, a standard FP analysis would have resulted in curve fits very similar to those from the FSP-spike analysis. In the third example, the standard FP approach would have selected a log function of Gleason score. In contrast, FSP-spike selected only the binary indicator z.

All approaches in which functions are determined in a data-dependent manner are less satisfactory if several highly correlated exposures are considered. In such situations it could be preferable to derive a new metric with better biological interpretation for this 'complex' of variables. The principle is illustrated for several smoking variables [1, 2].

FSP-spike selects a model in a data-dependent fashion without any theoretical justification for the chosen functional form. For case-control studies, some theoretical results are available in some simple cases, such as an exposure with a lognormal distribution in cases and controls. In these situations, the proportion of zeroes and the distribution of the exposure will affect the most suitable function. Extension to the gamma distribution is possible. However, in real data the exposure distribution often takes no recognizable or standard form, and simulation studies are required to investigate when and how often a 'suitable' function, with or without the binary term, is selected with our algorithm.

Acknowledgements

The original data set for the breast cancer study was provided by Jenny Chang-Claude.

Sources of financial support: Royston was supported by U.K. Medical Research Council grant number U.1228.06.001.00002.01. Royston and Sauerbrei received some support from the RiP-program, Mathematisches Forschungsinstitut, Oberwolfach, Germany.

References

1. Thurston SW, Liu G, Miller DP, Christiani DC. Modeling lung cancer risk in case-control studies using a new dose metric of smoking. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:2296–2302.
2. Leffondré K, Abrahamowicz M, Yongling X, Siemiatycki J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Statistics in Medicine* 2006; **25**:4132–4146.
3. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). *Applied Statistics* 1994; **43**(3):429–467.
4. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**:964–974.
5. Royston P, Sauerbrei W. *Multivariable Model-building*: *A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester, 2008.
6. Robertson C, Boyle P, Hsieh C-C, Macfarlane GJ, Maisonneuve P. Some statistical considerations in the analysis of case-control studies when the exposure variables are continuous measurements. *Epidemiology* 1994; **5**:164–170.
7. Pregibon D. Data analytic methods for matched case control studies. *Biometrics* 1984; **40**:639–651.
8. Jedrychowski W, Becher H, Wahrendorf J, Basa-Cierpialek Z, Gomola K. Effect of tobacco smoking on various histological types of lung cancer. *Journal of Cancer Research and Clinical Oncology* 1992; **118**:276–282.
9. Kropp S, Becher H, Nieters A, Chang-Claude J. Low and moderate alcohol consumption and breast cancer risk by age 50 among women in Germany. *American Journal of Epidemiology* 2001; **154**:624–634.
10. Stamey TA, Kabalin JN, McNeal JE, Johnstone IM, Freiha F, Redwine EA, Yang N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *Journal of Urology* 1989; **141**:1076–1083.
11. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model-building. *Statistics in Medicine* 2007; **26**:5512–5528.