CHAPTER 7

Interactions

Summary

1. An interaction is present when the effect of one predictor on the outcome varies according to the value of another.
2. Interactions may be predefined (e.g. in a clinical trial protocol) or emerge as a result of data analysis (hypothesis generation). Caution should be exercised in assessment and interpretation of the latter type. Critical interactions must be confirmed in independent samples (not discussed here).
3. For continuous predictors, categorization is inadvisable because the results depend on the chosen cutpoint(s). Also, power is lost.
4. Extensions of MFP are available for modelling interactions while adjusting for other variables.
5. Two types of interaction with continuous covariates are considered: binary (with obvious extension to categorical) by continuous (MFPI), and continuous-by-continuous (MFPIgen). The treatment effect plot is effective for displaying the former type graphically.
6. Graphical checks of a hypothesized interaction using categorized continuous variables (about four groups) and sensitivity analyses are essential adjuncts to analysis. If these checks confirm the results of the modelling, then the existence of the interaction is more credible.
7. An alternative approach for examining binary-by-continuous interactions, STEPP, is also available. STEPP involves examining interaction effects graphically in overlapping subpopulations of a continuous covariate. Results from STEPP depend on how the subpopulations are defined.

7.1 INTRODUCTION

In this chapter, binary-by-continuous and continuous-by-continuous interactions are considered. Extensions of MFP are described that allow modelling of the two types of interaction, adjusting for other predictors if desired.

Multivariable Model-Building Patrick Royston, Willi Sauerbrei

© 2008 John Wiley & Sons, Ltd

7.2 BACKGROUND

So far, only main effects of predictors on the outcome have been considered. If a factor z_2 explains (at least partially) the relationship between factor z_1 and the outcome y, then confounding is present. Confounding was found in many multivariable models in the earlier chapters. Another important issue is interaction between two or more predictors in a multivariable model. An interaction between z_1 and z_2 is present if z_2 modifies the relationship between z_1 and the outcome. That means that the effect of z_1 is different in subgroups determined by z_2. For example, an interaction between treatment and stage of disease is present if treatment B prolongs survival time in early stage disease compared with treatment A, whereas at a later stage the two treatments result in similar survival times. In an epidemiological study, the effect of an exposure on the probability of developing a disease may be different among smokers and nonsmokers. In the latter context, the equivalent term 'effect modification' is often used. In some disciplines (e.g. behavioural sciences) the preferred term is 'moderation'. In a clinical trial, the concern is typically whether the effect of a prognostic factor is homogeneous across the treatments (no interaction) or whether heterogeneity (interaction) is present.

The general concept and consideration of assessment and testing for interactions are described in many textbooks for different types of outcome data (e.g. Rothman and Greenland, 1998; Woodward, 1999; Cohen et al., 2003; DeMaris, 2004; Vittinghof et al., 2005) . This basic material is not repeated here. We focus particularly on two-way interactions involving at least one continuous covariate. Higher order interactions, which typically play a role in factorial experiments, are ignored.

7.3 GENERAL CONSIDERATIONS

7.3.1 Effect of Type of Predictor

If z_1 and z_2 are both binary, then no modelling issue arises. When z_1 is binary and z_2 is continuous, a common type of analysis is to categorize z_2 into a number of groups according to one or more cutpoints and to analyse the interaction with z_1 in a model with these main effects and multiplicative interaction terms. A trend test of the effect of z_1 over the ordered categories from z_2 may be performed, and if a trend is present is likely to have more power than the more general unordered test (Becher 2005). All of this raises several issues for the analyst, including: dependence of the statistical significance of the interaction on the number and position of the cutpoints; the interpretation of the results when an unstable model with too many cutpoints is fitted; and, in the case of a trend test, possible loss of power and faulty interpretation if a nonlinear relationship is incorrectly assumed to be linear, because of a poor choice of scores. Another approach is not to apply categorization but to assume linearity in z_2 at both levels of z_1. However, the assumption of linearity may, of course, be incorrect.

If z_1 has more than two (ordered or unordered) categories, then interactions are often handled pairwise or by combining factor levels to produce a binary variable again. If both z_1 and z_2 are continuous, then linearity is usually assumed for both, and the product $z_1 \times z_2$ is tested. However, the assumption could result in an erroneous model if the main effects and/or the interaction are in fact nonlinear.

In real data, a mixture of types of predictor is found. This mixture is reflected in our examples.

7.3.2 Power

Since studies are almost invariably powered to detect main effects of interest, power is usually low to detect even moderately large interactions. For a simple special case in the epidemiological context, Greenland (1993) concluded from considerations of asymptotic power functions and from simulation studies that the power of such tests is low in common situations. Results from efficiency theory (Lagakos, 1988; Farewell et al., 2004) show that the power of tests of interaction is much improved if a factor is continuous rather than binary. It is clear, therefore, that a large sample size is in most cases a prerequisite for a sensible analysis of interactions.

7.3.3 Randomized Trials and Observational Studies

Because of their relevance to treating patients, interactions in a randomized controlled trial are probably the most important case. Because of the randomization, the covariates are (at least in a large study) independent of the treatment variable z_1 by design, and model building involving treatment, therefore, is less of an issue. When z_2 is binary, the analysis is straightforward. With more than two categories, difficulties arise, and no standard approach exists with continuous variables.

In this chapter, we aim to use all information from a continuous covariate while allowing possible nonlinearity in z_2 at all levels of z_1, possibly adjusting for other influential covariates. To do this, an MFP interaction (MFPI) algorithm, an extension of MFP, was suggested for investigating treatment–covariate interactions (Royston and Sauerbrei, 2004a). Adjustment for other covariates enables the methodology to be used generally in observational studies for a binary variable of interest. For variables with more than two categories, see comments in Section 7.3.1. Of course, weaknesses of observational studies also manifest themselves in interactions.

The chapter focuses mainly on using MFPI to analyse treatment–covariate interactions in trials. However, we also consider continuous-by-continuous interactions, a topic that arises naturally in observational studies. Components of MFPI are modified for use in the latter case.

7.3.4 Predefined Hypothesis or Hypothesis Generation

Sometimes, there is literature evidence that two factors may interact, and this hypothesis is to be tested in the study. Ideally, such a hypothesis is prespecified, e.g. in a clinical protocol. Another (more common) situation is a data-driven search for possible interactions to improve model fit or to generate hypotheses for further research. In principle, an analysis of interactions may be done in either setting. The results, however, are interpreted differently. With a prespecified hypothesis, all that is required is a single test of interaction. In the hypothesis generation case, model building is necessary, with attendant issues such as which potential interactions to consider, which adjustment model to use, possible correction for multiple testing, use of selection strategies to determine a model including interactions, etc.

In designed experiments, it is usual to test for all possible two-way interactions, even if main effects are not significant (Bishop et al., 1975). In contrast, in model building in observational studies, variables with no significant main effect are often disregarded when considering possible interactions. However, it is conceivable that a variable without a main effect could interact with another variable. Identifying such an interaction may be of interest and can be

done by extending the search for interactions. Of course, the multiplicity issue becomes more severe and interpretation more difficult. The point is discussed and illustrated in examples.

The multiplicity issue is similar to the usual problem of multiple testing. Researchers must interpret P values carefully, and some may prefer to adjust them using procedures such as Bonferroni–Holm (Holm, 1979). However, use of multiplicity adjustment may rob the procedure of most of its (already low) power. It is then likely that even important interactions are missed. The MFPI algorithm was developed to search in a systematic way for possible heterogeneity of treatment effects for a continuous covariate in a randomized trial (Royston and Sauerbrei, 2004a). To generate hypotheses, the MFPI algorithm may also be used for model building in trials and observational studies, interaction terms being added in a forwards stepwise manner. A low nominal P-value, such as 0.01, may be used to control to some extent for overfitting (generating false positive hypotheses).

7.3.5 Interactions Caused by Mismodelling Main Effects

If the main effects of one or both variables of a continuous-by-continuous interaction are incorrectly modelled, e.g. nonlinearity is ignored, then a spurious interaction may be generated. This is a type of residual confounding (Rothman and Greenland, 1998). For example, in the GBSG study, if the main effects of `x3` and `x5` are erroneously assumed to be linear, then there is an interaction between them significant at the 0.01 level. If the nonlinearity in the effect of `x5` is allowed for, then the interaction disappears ($P = 0.2$). Such an effect could also occur if an important adjustment variable was mismodelled. Therefore, attention should be paid to determining an appropriate main-effects model.

7.3.6 The 'Treatment–Effect' Plot

An informative graphical description of the interaction of (binary) z_1 with (continuous) z_2 is a plot of the estimated difference in response between the levels of z_1, as predicted by a model, together with its 95% pointwise confidence band, against z_2. Let $\widehat{f}_0(z_2)$ and $\widehat{f}_1(z_2)$ be the predicted functions at levels 0 and 1 of z_1 respectively. By analogy with a randomized trial, the 'treatment effect' is defined as $\widehat{t}(z_2) = \widehat{f}_1(z_2) - \widehat{f}_0(z_2)$. A significant interaction may be 'qualitative' if $\widehat{t}(z_2)$ crosses zero, because then not only the magnitude but also the *direction* of the treatment effect depends on z_2. In such cases, a treatment may on average be beneficial or harmful depending on the patient's value of z_2. A quantitative interaction means that the treatment effect changes with z_2, but is in the same direction at all relevant values of z_2. See Section 7.5.1 for a detailed example with a plot.

7.3.7 Graphical Checks, Sensitivity and Stability Analyses

To reduce the chance of overfitting and of incorrectly identifying interactions, it is necessary to check the results of complex model-building procedures. Simple graphical checks can be applied to confirm the inferences from the selected model. Sensitivity analyses may involve changing (even removing) the adjustment model or the type of function used to model the interaction. Influential observations may drive an apparent interaction, and this danger should be investigated. In our experience, continuous predictors with a markedly skew distribution are a fertile source of spurious interactions. Application of the robustness transformation to such predictors may help to curb such effects by reducing the influence of extreme observations.

In the case of continuous-by-continuous interactions, further robustification may be achieved by categorizing both variables into, say, three or four groups and graphing the effect estimates against the subgroup means or medians, adjusted for other variables if necessary. One should try to verify the interaction by inspecting the ordering of the results for consistency with that suggested by the model. Disagreement with expectation is an indicator of an erroneous model.

Another possibility is to check the stability of the treatment–effect function by bootstrap resampling. For example, in the kidney cancer example, the interaction between treatment and white cell count was shown to be reproducible when the treatment–effect function was averaged over 1000 bootstrap samples (Sauerbrei and Royston, 2007).

7.3.8 Cautious Interpretation is Essential

Provided that a small number of hypotheses have been prespecified, interpretation of results is clear. However, when model building is used to generate hypotheses or to improve the fit of a model, several types of bias may appear. Currently, the statistical community has no consensus on how to investigate and interpret interactions in clinical trials and observational studies. Furthermore, since power is known to be low, interactions are often not even considered. Many important interactions may be missed.

For studies with a sufficient sample size, we believe that it is worth trying to identify interactions (at least, strong ones). Because of multiplicity, many spurious interactions may be identified by any modelling approach. Therefore, a sceptical attitude to the results of such exercises is essential. 'Reality checks', several of which we describe, must be applied to confirm that interactions are really present in the original data. Interactions should also be checked for consistency with subject-matter knowledge, where available. Finally, confirmation in independent data is always required.

7.4 THE MFPI PROCEDURE

Royston and Sauerbrei (2004a) proposed an extension of the MFP algorithm to investigate interaction between a categorical and a continuous covariate. For simplicity of explanation, z_1 is described as a binary 'treatment' variable, although the methodology applies equally well to any categorical covariate. The context of prognostic factors in a randomized trial is also assumed, though again the restriction is purely linguistic. As above, let z_2 be a continuous covariate and z_1 be a binary treatment variable, coded {0, 1}. In clinical settings, a variable that is prognostic and that interacts with treatment is often called 'predictive' (of response to treatment). A given variable may be both prognostic and predictive, prognostic only, predictive only, or neither prognostic nor predictive.

Let x be a vector of potential prognostic factors. With a prespecified hypothesis, z_2 is the only covariate to be investigated for interaction with z_1. For hypothesis generation, z_2 is one of several candidate predictive factors; often, z_2 is part of $\mathbf{x}$. The relationship between the outcome and z_2 is modelled by an FP with the same powers but different regression coefficients at each level of z_1. A standard test of interaction is performed on regression coefficients at the final step. To reduce possible confounding, adjustment for $\mathbf{x}$ may be made. Since z_2 cannot belong

to the adjustment model, a different such model may be chosen for each z_2 that is investigated. The complete procedure, allowing adjustment for $\mathbf{x}$, is as follows:

1. Apply the MFP algorithm to $\mathbf{x}$ (possibly including z_2) with a P-value threshold of α^* for selecting variables and FP transformations. Let $\mathbf{x}^*$ be the resulting covariate vector, called the adjustment model. $\mathbf{x}^*$ may include (transformed) variables in $\mathbf{x}$ selected by the MFP algorithm. If all variables in $\mathbf{x}$ are uninfluential, then $\mathbf{x}^*$ may even be empty, i.e. no adjustment for members of $\mathbf{x}$ appears to be needed. In some cases, parts or even all of $\mathbf{x}^*$ may be formulated from subject-matter knowledge, reducing or avoiding data-driven searching.
2. Find by maximum likelihood the best-fitting FP2 powers $\mathbf{p} = (p_1, p_2)$ for z_2 with $p_1, p_2 \in S$, always adjusting for z_1 and $\mathbf{x}^*$. Denote the FP2 transformations $z_2^{\mathbf{p}} = (z_2^{p_1}, z_2^{p_2})$.
3. For groups $j = 0, 1$ and powers p_i for $i = 1, 2$, define new predictors $z_{ji} = z_2^{p_i}$ if $z_1 = j$, and $z_{ji} = 0$ otherwise.
4. The test of $z_1 \times z_2$ interaction is a likelihood ratio test between the nested models z_1, z_{01}, z_{02}, z_{11}, z_{12}, $\mathbf{x}^*$ and z_1, $z_2^{p_1}$, $z_2^{p_2}$, $\mathbf{x}^*$. The difference in deviance is compared with χ^2 on two d.f.
5. If an interaction is not found, then z_2 is regarded as a potential prognostic factor only. To investigate if an FP2 function is still needed for z_2, the final model is chosen by repeating step 1, but including z_2 as a potential prognostic factor.

The reason to fit FP2 functions to z_2, rather than simpler functions, is to find the best-fitting specification from a flexible class. In terms of bias/variance trade-off, increased variance incurred through the use of FP2 powers for z_2 may be tolerated as the price of low bias. To avoid excessive overfitting, leading to serious artefacts in the fitted functions, estimation of different powers in each treatment group is not entertained. An FP2 function with the same powers in each treatment group is already a flexible specification.

When z_2 is binary or categorical, the approach reduces to the usual procedure of estimating and testing multiplicative interaction term(s), adjusting for $\mathbf{x}^*$.

7.4.1 Model Simplification

Despite consideration of the bias/variance trade-off, overfitting of interaction terms and the consequent instability resulting from use of the MFPI algorithm with FP2 functions of z_2 may be a real concern. Instead of FP2 functions, the MFPI algorithm just outlined may be implemented using FP1 functions or, previously the 'standard' choice, using linear functions. Owing to their mode of construction, these three interaction models (with FP2, FP1 or linear functions of z_2) are not nested. Nested hypothesis testing, therefore, cannot be used to select a model from among them. Instead, a model may be chosen to optimize a penalized likelihood criterion, such as minimum AIC or BIC. However, since the BIC penalty is quite harsh, the AIC criterion may be preferable to reduce the chance of underfitting and consequent bias.

7.4.2 Check of the Results and Sensitivity Analysis

If a $z_1 \times z_2$ (discrete by continuous) interaction is detected by MFPI, then the treatment–effect plot (see Section 7.3.6) indicates where the differences should lie. This may be checked graphically (e.g. in survival data, by graphs of Kaplan–Meier survival curves) and by estimating

the treatment effect in subgroups. This check introduces an element of subjectivity. However, the value of using Kaplan–Meier plots and estimates of treatment effect in a few subgroups has been demonstrated in examples (Royston and Sauerbrei, 2004a; Royston et al., 2004; Sauerbrei et al., 2007d). To show a trend, more than two subgroups must be used. Often, a sensible choice is four subgroups, with cutpoints indicated by the treatment–effect plot and the distribution of the covariate. If an apparent interaction is found following intensive modelling, then it is unlikely that the results in the independent subgroups show a similar pattern. To reduce the chance of artefacts, no subgroup should be too small. This check helps one to identify obvious type I errors. It is also a simple and, it is hoped, convincing way of presenting the main results from MFPI to researchers. When estimating the treatment effect in subgroups, the adjustment model $\mathbf{x}^*$ can also be incorporated. This is less relevant in randomized trials, but important in all types of observational study.

It may be worth additionally performing a sensitivity analysis of interactions to adjustment models of differing complexity. The simplest way to do this is to vary the nominal P-value for selecting a variable. Possible choices are $P = 1$ (full model except z_1 and z_2, perhaps with some additional transformations of continuous covariates), $P = 0.157$ (approximately equivalent to selection by the AIC), $P = 0.05$ (conventional level) and P small, according to the BIC, which depends on the (effective) sample size. At least in a randomized trial, a reliable interaction should survive such modifications to the adjustment model.

7.5 EXAMPLE 1: ADVANCED PROSTATE CANCER

The first example is a well-known trial in patients with advanced prostate cancer, which has been analysed using Cox regression by Byar and Green (1980) and others; the data may be found in Andrews and Herzberg (1985). The main outcome is time to death (overall survival). Of 506 patients randomized to the four treatments under study, only the 475 patients (338 deaths) with complete data on all covariates are considered for this reanalysis. The treatments consisted of a placebo and three dose levels of the synthetic oestrogen drug diethyl stilboestrol. For reasons given by Byar and Green (1980), the placebo and the lowest dose level of diethyl stilboestrol have been combined to give a placebo arm, and the higher doses to give treatment E. This combined treatment variable is called `rx`. Prior to the main analyses, the implausible value of 6 observed in a single patient for the variable `stage` was recoded to 4 (assuming it was a typo), making `stage` binary (3/4), no other stages being represented; and `ekg`, a categorical covariate with seven levels, was recoded as 0 (normal), 1 (abnormal). The prognostic factors are listed in the first column of Table 7.1.

The analysis is focused on the identification of treatment–covariate interactions. Byar and Green (1980) modelled the seven prognostic factors `age`, `hg`, `hx`, `pf`, `sg`, `wt` and `sz` (see Table 7.1), and all seven treatment–covariate interactions. All continuous or multicategory factors were categorized into two or three groups. By contrast, in the present analysis, all continuous factors are kept continuous. Some differences between Byar and Green's (1980) results and those from MFPI would, therefore, be expected.

Since no interactions were predefined, the analysis is an example of hypothesis generation. The MFPI algorithm was applied to all binary and continuous factors using $\alpha^* = 1.0, 0.157$ and 0.05 to select variables, and the default 0.05 level was used to select FP functions. Whether or not the main effects entered the adjustment model, all interactions were tested at the $\alpha = 0.05$ significance level in a model including their main effects. The results are given in Table 7.1.

Table 7.1 Advanced prostate cancer data. Prognostic factors and interactions with treatment. The eight factors in the upper portion of the table are continuous; the remainder are binary.[a]

Prognostic factor	Code	α^* (significance level for adjustment model)					
		1.0		0.157		0.05	
		Main	Int.	Main	Int.	Main	Int.
Continuous							
Age at diagnosis	age[b]	✓	0.018	✓	0.022	✓	0.027
Standardized weight	wt[b]	✓		✓		✓	
Systolic blood pressure	sbp	✓					
Diastolic blood pressure	dbp	✓					
Size of primary tumour	sz[b]	✓		✓		✓	
Serum acid phosphatase	ap	✓	0.044		0.030		(0.064)
Haemoglobin (g/100 ml)	hg[b]	✓					
Gleason stage-grade category	sg[b]	✓		✓			
Binary							
Performance status	pf[b]	✓		✓		✓	
History of cardiovasc. disease	hx[b]	✓		✓		✓	
Presence of bone metastases	bm	✓	0.013	✓	0.008		0.003
Stage 4 vs stage 3	stage	✓				✓	
Abnormal electrocardiogram	ekg	✓		✓		✓	

[a] A tick denotes a variable included in the model. Ticked entries under 'Main' denote prognostic variables selected in the adjustment model by the MFP algorithm using a nominal P-value of α^*. Entries under 'Int.' are P-values for treatment–covariate interactions with $P < 0.05$ according to the MFPI algorithm.
[b] Variable selected by Byar and Green (1980).

The adjustment model selected at $\alpha^* = 0.05$ comprises the binary factors pf, hx, stage and ekg, and the continuous factors age, wt and sz (all except age modelled as linear functions). rx × age and rx × bm interactions are found to be significant at the 0.05 level irrespective of the adjustment model. A significant rx × ap interaction is seen for $\alpha^* = 1.0$, 0.157, but not quite with $\alpha^* = 0.05$.

For $\alpha^* = 0.05$, the effect of age was modelled by FP2 functions with powers (3, 3) in each treatment group (results for other α^* were very similar). Figure 7.1 (left panel) shows that for patients older than 75 years the hazard increases in a similar fashion in both groups, whereas younger patients on treatment E appear to have a lower risk. Figure 7.1 (right panel) is a treatment effect plot by age. The benefit of treatment E seems to be substantial for younger patients but may be lost or even reversed for older ones. The confidence intervals outside the range of (60, 80) years are wide.

7.5.1 The Fitted Model

The linear predictor, $\eta(\mathbf{x}^*, z_1, z_2)$, in the Cox model for the age function ($\widehat{\beta}_1 \text{age}^3 + \widehat{\beta}_2 \text{age}^3 \log \text{age}$) in the advanced prostate data, including an interaction with age selected

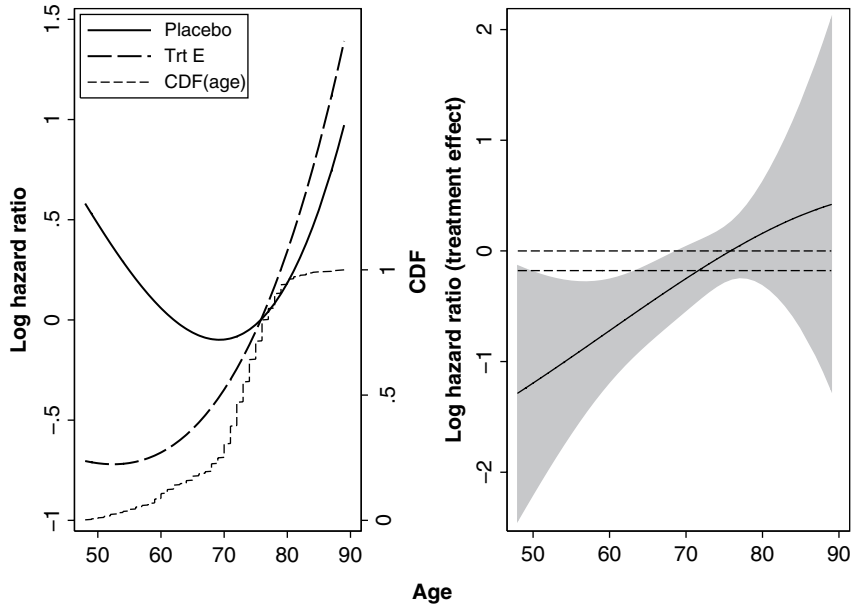

Figure 7.1 Advanced prostate cancer data. rx(treatment) × age interaction. Functions were estimated within a multivariable adjustment model $\mathbf{x}^*$ selected at the $\alpha^* = 0.05$ level and fitted using FP2 functions with powers (3, 3). Left panel: solid line, estimated effect of age in patients on placebo (rx = 0); dashed line, estimated effect of age in patients on treatment E (rx = 1); dotted line and right-hand axis, cumulative distribution function (CDF) of age. Right panel: treatment effect function for age, with 95% pointwise CI. Horizontal dashed lines denote zero and the main effect of rx in a model excluding interaction. (Adapted from Royston and Sauerbrei (2004a) with permission from John Wiley & Sons Ltd.)

at the $\alpha^* = 0.05$ level, is as follows. Let $z_1 = \mathtt{rx}$, $z_2 = \mathtt{age}/10$, $\overline{z}_2 = \text{mean}(\mathtt{age})/10 = 7.156$. Then

$$\begin{aligned}\eta(\mathbf{x}^*, z_1, z_2) = &-0.0114(0.0045)\mathtt{wt} + 0.0185(0.0043)\mathtt{sz} + 0.431(0.170)\mathtt{pf}\\ &+ 0.389(0.136)\mathtt{hx}\\ &+ 0.239(0.115)\mathtt{stage} + 0.303(0.123)\mathtt{ekg} - 0.180(0.112)\mathtt{rx}\\ &- 0.0463(0.0288)z_{01} + 0.0204(0.0128)z_{02} - 0.0226(0.0231)z_{11}\\ &+ 0.0114(0.0100)z_{12}\end{aligned}$$

where

$$\begin{aligned}z_{01} &= (z_2^3 - \overline{z}_2^3)I(\mathtt{rx} = 0), \quad z_{02} = (z_2^3 \log z_2 - \overline{z}_2^3 \log \overline{z}_2)I(\mathtt{rx} = 0)\\ z_{11} &= (z_2^3 - \overline{z}_2^3)I(\mathtt{rx} = 1), \quad z_{12} = (z_2^3 \log z_2 - \overline{z}_2^3 \log \overline{z}_2)I(\mathtt{rx} = 1)\end{aligned}$$

and $I(condition) = 1$ if $condition$ is true, 0 otherwise. Standard errors are given in parentheses. Scaling (dividing by 10) and centering has been applied to age (see section 4.11). The subtraction of the constants $\overline{z}_2^3$ and $\overline{z}_2^3 \log \overline{z}_2$ from the FP2 transformations of z_2 centres the FP functions on the mean of age.

The estimated prognostic functions of age in treatment (rx) groups 0 and 1 and the treatment–effect function are, respectively,

$$\begin{aligned}\widehat{f}_0(z_2) &= -0.0463(z_2^3 - \overline{z}_2^3) + 0.0204(z_2^3 \log z_2 - \overline{z}_2^3 \log \overline{z}_2)\\ \widehat{f}_1(z_2) &= -0.0226(z_2^3 - \overline{z}_2^3) + 0.0114(z_2^3 \log z_2 - \overline{z}_2^3 \log \overline{z}_2) - 0.180 \qquad (7.1)\\ \widehat{t}(z_2) &= \widehat{f}_1(z_2) - \widehat{f}_0(z_2) = 0.0237(z_2^3 - \overline{z}_2^3) - 0.0090(z_2^3 \log z_2 - \overline{z}_2^3 \log \overline{z}_2) - 0.180\end{aligned}$$

where -0.180 is the main effect regression coefficient of rx in $\eta(\mathbf{x}^*, z_1, z_2)$. These three functions of z_2 (i.e. age) are plotted in Figure 7.1.

7.5.2 Check of the Interactions

Figure 7.2 compares the survival distributions by treatment and age group using age cutpoints of 60, 70 and 75 years. The choice of cutpoints was guided by the treatment effect plot (Figure 7.1). The subgroup age $\leq$ 60 years is relatively small but still has 30 events. Figure 7.2 supports the MFPI analysis of the age interaction. The log hazard ratios for the effects of treatment in these subgroups, unadjusted and using the $\alpha^* = 0.05$ adjustment model, are shown in Table 7.2. They confirm the trend seen in Figures 7.1 and 7.2.

Since bm is binary, plotting Kaplan–Meier curves by rx for the two bm groups provides a check of the rx $\times$ bm interaction, unadjusted for other factors (see Figure 7.3). There is a large treatment effect ($\log \text{HR} = -0.89$) in the subset of patients with bone metastases (bm $= 1$)

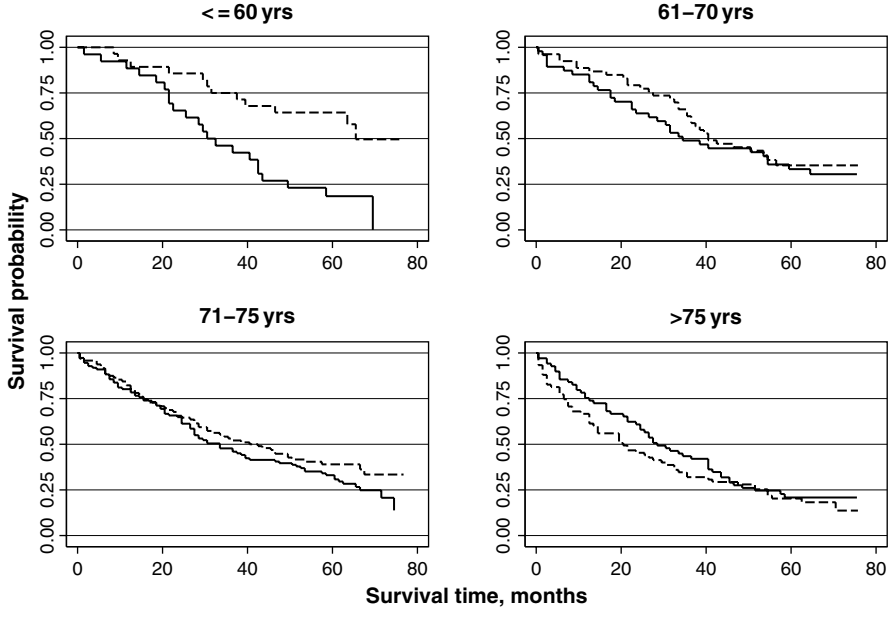

Figure 7.2 Advanced prostate cancer data. Kaplan–Meier survival curves illustrating rx $\times$ age interaction with patients divided into four groups by using age cutpoints of 60, 70 and 75 years. Solid lines: placebo (rx $= 0$); dashed lines: treatment E (rx $= 1$). (Adapted from Royston and Sauerbrei (2004a) with permission from John Wiley & Sons Ltd.)

Table 7.2 Advanced prostate cancer data. Treatment effect (log hazard ratio) in age subgroups. 'Adjusted' means adjusting for factors significant at the 0.05 level, as given in Table 7.1.

Subgroup	Age (years)	Patients (%)	Unadjusted		Adjusted	
			$\widehat{\beta}$	SE	$\widehat{\beta}$	SE
1	≤ 60	10	−1.29	0.39	−1.53	0.49
2	61–70	20	−0.15	0.26	−0.34	0.27
3	71–75	43	−0.25	0.17	−0.16	0.18
4	> 75	27	0.27	0.19	0.16	0.21
All patients		100	−0.18	0.11	−0.16	0.11

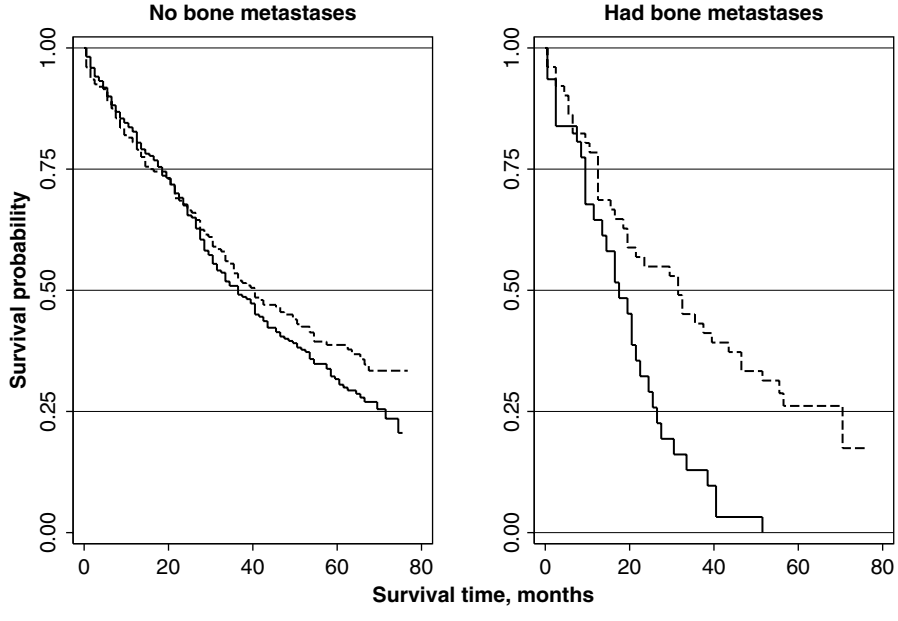

Figure 7.3 Advanced prostate cancer data. Kaplan–Meier curves of overall survival illustrating `rx` × `bm` (bone metastases) interaction. Solid lines: placebo (`rx` = 0); dashed lines: treatment E (`rx` = 1).

and nearly no treatment effect (HR = −0.16) when `bm` = 0. The adjusted estimates shown in Table 7.3 and again using $\alpha^* = 0.05$, are similar to the unadjusted.

7.5.3 Final Model

The interactions identified by MFPI appear to be genuine. To determine a final model, a forwards stepwise (FS) method may be used. Here, FS is preferred to BE because BE would begin with a heavily overfitted, unstable model. It is unlikely that several interactions are

Table 7.3 Advanced prostate cancer data. Treatment effect (log hazard ratio) in subgroups by presence or absence of bone metastases. 'Adjusted' means adjusting for factors significant at the 0.05 level, as given in Table 7.1.

Subgroup of bm	Patients (%)	Unadjusted		Adjusted	
		$\widehat{\beta}$	SE	$\widehat{\beta}$	SE
No bone metastases (bm = 0)	84	−0.16	0.12	−0.06	0.12
Had bone metastases (bm = 1)	16	−0.89	0.26	−0.89	0.28

present. Interactions significant at the 0.05 level are added to the main effects model (see Table 7.1). Because it has the smallest P-value (0.003), the rx × bm interaction is added first. After including rx × bm, the rx × age interaction is significant, and both interactions remain significant in the extended model ($P = 0.036$ and 0.005 respectively). With a more rigorous significance level of 0.01, only rx × bm would remain.

7.5.4 Further Comments and Interpretation

The rx × age interaction was also reported as significant at $P = 0.05$ by Byar and Green (1980). However, Figure 7.1 is more informative than analyses based on cutpoints. Byar and Green (1980) defined three age groups ≤ 74, 75–79 and ≥ 80 years and performed conventional tests of interaction. According to the fitted FP2 functions, there is a substantial beneficial effect of treatment E (rx = 1) for younger patients. The estimated treatment effect is −0.72 (SE 0.24) at age 60 years and decreases to −0.03 (SE 0.13) at age 75 years (see Figure 7.1 and $\widehat{t}(z_2)$ in Section 7.5.1).

Byar and Green (1980) reported no results for the continuous covariate ap (serum acid phosphatase), even though this variable was available in the public-domain dataset. According to the present analysis, there may be a weak rx × ap interaction. A treatment–effect plot (not shown) suggests that treatment E may be effective only in patients with higher ap values.

Results for the variable bm (presence/absence of bone metastases) were also not reported by Byar and Green (1980), despite it appearing to be a fairly strong predictor of response to treatment ($P < 0.01$).

Note that all variables were investigated for possible interaction with rx, irrespective of their prognostic significance. The more common practice is to consider only prognostic variables, in which case the interaction with bm would not have been found at the $\alpha^* = 0.05$ level. The same is true for ap at the $\alpha^* = 0.157$ level.

Note also that the significance of the interaction with ap depends on the adjustment model. The interaction may, therefore, be a weak effect and may be disregarded. A final decision would depend on further checking and on clinical reasoning.

None of these apparently predictive factors would be significant following P-value adjustment by the Bonferroni–Holm method.

An alternative approach to analysing the same dataset was presented by Harrell (2001, p. 516). He applied a global test of all interactions between rx and the other variables, and found $X^2 = 12.2$ on 10 d.f. ($P = 0.27$). He states 'so we ignore the interactions'. For clinical application, it is obviously essential to know which, if any, interactions are present. A global test is unlikely to be helpful, and lacks power.

Table 7.4 Advanced prostate cancer data. `age` × `rx` interaction model selection according to AIC and BIC, based on partial likelihood from Cox models.

Criterion	Model class		
	FP2	FP1	Linear
AIC	3720.66	3718.26	3718.05
BIC	3747.42	3733.55	3729.52

7.5.5 FP Model Simplification

The dataset was initially analysed using FP2 functions for continuous covariates when estimating interactions with treatment. As discussed in Section 7.4.1, the chance of overfitting the data may be reduced by selecting a simpler FP function, if appropriate. For the interaction between `rx` and `age`, Table 7.4 shows the AIC and BIC statistics for FP2, FP1 and linear models, including an adjustment model selected with $\alpha^* = 0.05$. The BIC was calculated using the effective sample size, taken as the number of events (Volinsky and Raftery, 2000). Since both the AIC and the BIC are smallest for the linear model, each criterion would select this model. Also, both criteria suggest that the FP2 model is overfitted. According to the AIC, the FP1 model is a reasonable alternative to a straight line, but the BIC, with its stricter penalty, suggests it is overfitted. The P-value from the likelihood-ratio test for the linear `age` × `rx` interaction is 0.0008. Even though the prognostic effects are different, the treatment–effect function is similar to that in Figure 7.1.

7.6 EXAMPLE 2: GBSG BREAST CANCER STUDY

7.6.1 Oestrogen Receptor Positivity as a Predictive Factor

It is well established that oestrogen receptor (ER) status `er` is a predictive factor for response to hormonal therapy with tamoxifen (`tam`). The risk of disease recurrence is reduced to a much greater extent by `tam` in ER-positive patients (Early Breast Cancer Trialists' Collaborative Group, 1998). For present purposes, therefore, the `tam` × `er` interaction is an evidence-based hypothesis. However, the choice of cutpoint on `er` which defines ER-positivity is controversial.

Here, the data are used as an example of applying the MFPI algorithm to investigate a predefined interaction. In Section 7.7, it is used to demonstrate difficulties with categorization when investigating interactions.

7.6.2 A Predefined Hypothesis: Tamoxifen–Oestrogen Receptor Interaction

The adjustment model is considered first. Because of the high correlation between `pgr` and `er`, the vector $\mathbf{x}$ was taken as all the available prognostic factors except for `pgr`. The adjustment model $\mathbf{x}^*$ selected by MFP comprised an FP2 function for `age` with powers $(-2, -1)$, $\exp(-0.12 \times$ `nodes`$)$ and `gradd1` (grade 1 versus 2 or 3), at both α^*-levels of 0.157 and 0.05.

For the interaction, MFPI selected an FP2 function for `er` with powers $(-2, -1)$ or an FP1 function with power -0.5. The AIC values for linear, FP1 and FP2 models were 3459.6, 3448.1 and 3449.1 respectively. The FP1 model may, therefore, be preferred.

The FP1 function of `er` is shown for the two `tam` groups in Figure 7.4 (left panel). For clarity, the range of `er` values in the plot has been restricted to [0, 100]. The treatment–effect plot (Figure 7.4, right panel) illustrates the difference clearly. The test for interaction has a P-value of 0.042.

For large `er` values (say, > 20 fmol l^{-1}) the estimated treatment effect is nearly constant (log hazard ratio about -0.5, hazard ratio 0.6), but it changes sharply for small values. To check this result, the `tam` effect was considered in five subgroups with cutpoints partly determined by the treatment–effect function, partly by the CDF of `er`, and partly for clinical reasons. Biologically, zero is a special group, and 10 is the cutpoint most often used in the clinical literature. Table 7.5 shows the estimated treatment effect in subgroups of `er`, unadjusted and adjusted for other factors. The unadjusted estimates summarize the comparison between Kaplan–Meier curves presented in Figure 7.5. For these plots, groups 4 and 5 were amalgamated. Although some of the groups are small, analysis in subgroups confirms the treatment–effect function seen in Figure 7.4. In the small subgroup with `er` $= 0$ ($n = 82$ with 45 events), the patients treated with `tam` have a higher hazard, whereas in all other subgroups the hazard is reduced. These checks clearly indicate that the form of dependence of the estimated treatment effect on `er` is not the result of an artefact generated by the MFPI algorithm.

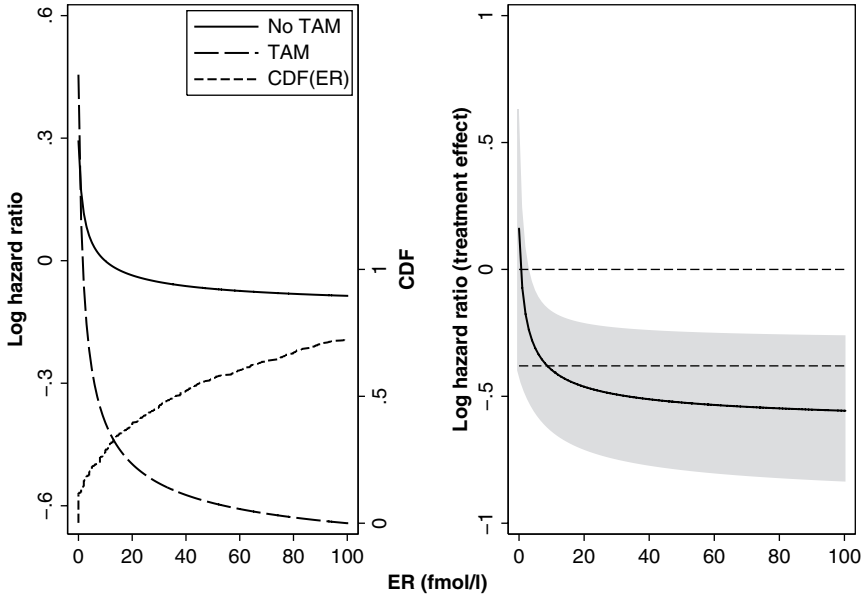

Figure 7.4 GBSG breast cancer data. Analysis of `tam` × `er` interaction keeping `er` continuous, fitted by FP1 functions with power -0.5. Functions were estimated within multivariable models; for details of the adjustment model, see the text. Left panel: solid line, estimated effect of `er` in patients not treated with `tam`; dashed line, estimated effect of `er` in patients treated with `tam`; dotted line and right-hand axis, CDF of `er`. Right panel: effect of `tam` by ER status, with 95% pointwise CI. Horizontal dashed lines denote zero and the main effect of `tam` in the absence of an interaction.

Table 7.5 GBSG breast cancer data. Effect of `tam` (log hazard ratio) in `er` subgroups. 'Adjusted' means adjusted for factors significant at $\alpha^* = 0.05$.

Subgroup	ER (fmol l^{-1})	Patients (%)	Unadjusted		Adjusted	
			$\widehat{\beta}$	SE	$\widehat{\beta}$	SE
1	0	12	0.67	0.31	0.47	0.32
2	1–10	17	−0.61	0.30	−0.41	0.31
3	11–36	21	−0.58	0.30	−0.68	0.32
4	37–115	25	−0.34	0.25	−0.40	0.26
5	>115	25	−0.31	0.26	−0.42	0.27

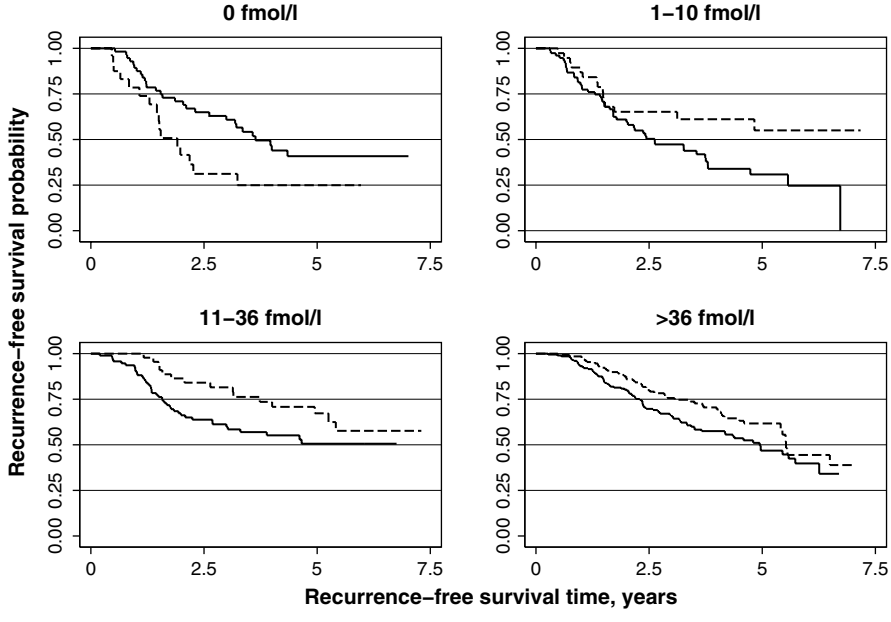

Figure 7.5 GBSG breast cancer data. Kaplan–Meier survival curves illustrating `tam` × `er` interaction with patients divided into four groups by using `er` cutpoints of 0, 10 and 36 fmol l^{-1}. Solid lines: `tam`; dashed lines: no `tam`. (Adapted from Royston and Sauerbrei (2004a) with permission from John Wiley & Sons Ltd.)

7.7 CATEGORIZATION

7.7.1 Interaction with Categorized Variables

In many fields, continuous variables are often converted into categorical variables by grouping values into two or more categories. As discussed in Section 3.4, categorization of continuous data is unnecessary for statistical analysis, and is not a natural way of analysing continuous variables. Usually, it is done to make the analysis and interpretation of results simpler. It appears to be the standard approach for handling interactions, at least in clinical research (Assmann et al., 2000).

In this section, the tam × er interaction in the GBSG data is used to illustrate the difficulties caused by categorizing a continuous variable. Often, only one cutpoint is used and it is not specified in advance. It is tempting to search for a suitable cutpoint to answer the two questions 'Does the study show that the effect of the treatment t depend on the continuous variable z?', and 'Which cutpoint on z best separates nonresponders from responders to t?' The first question might be answered by investigating P-values for the $t \times z$ interaction with different cutpoints on z, and the second by exploring possibly different treatment effects associated with each cutpoint.

7.7.2 Example: GBSG Study

In practice, patients with er > 20 fmol l^{-1} are always regarded as having high ER. Therefore, clinically relevant cutpoints lie below 20 fmol l^{-1}. In the GBSG study, 60% of patients had er > 20 fmol l^{-1}. For each integer cutpoint c on er in the range $[0, 20]$, binary dummy variables er_c were defined as 0 if er $\leq c$ and 1 if er $> c$. Cox regression models were fitted, adjusted for $\mathbf{x}^*$ as in Section 7.6.2. For each c the P-value for the interaction tam × er_c was calculated. Additionally, adjusted Cox models were fitted separately in each subgroup defined by er_c (i.e. low er and high er). The regression coefficients and associated P-values for the effect of tam were computed in each subgroup.

As shown in Figure 7.6 (left panel), the tam × er_c interaction is significant at $P < 0.05$ only for two cutpoints: $c = 0$ and $c = 8$. In the high ER group, the regression coefficient

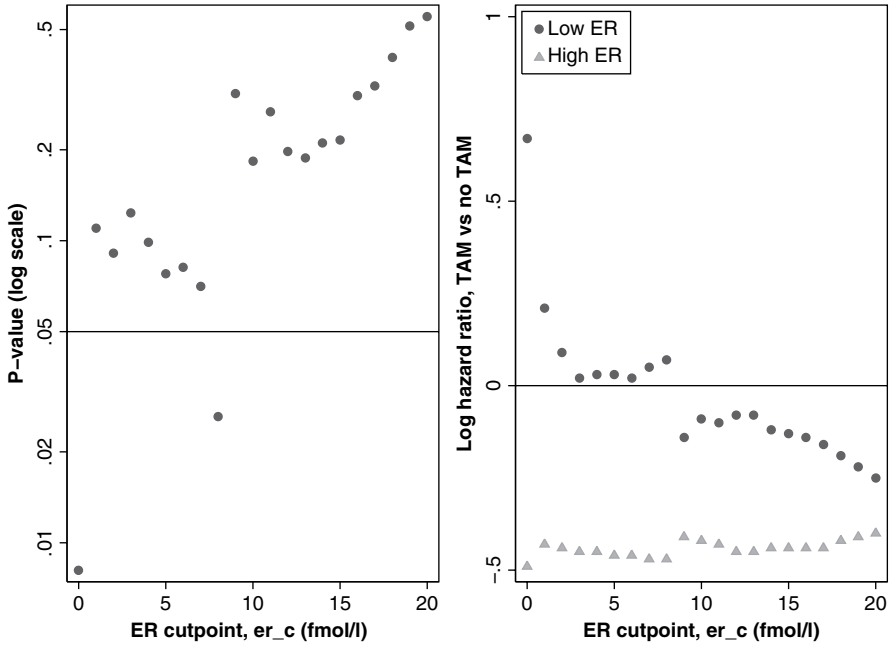

Figure 7.6 GBSG breast cancer data. Cutpoint analysis of interactions in 21 subgroups, adjusting for other prognostic factors. Cutpoints er_c in the range $[0, 20]$ are used to define low and high ER subgroups. Left panel: P-values for tests of tam × er interaction. Right panel: regression coefficients (log hazard ratios) from Cox models for effects of tam in subgroups with ER low (circles) and high (triangles). (Adapted from Royston and Sauerbrei (2004a) with permission from John Wiley & Sons Ltd.)

for tam is significantly different from zero for all c, usually with $P < 0.01$ (Figure 7.6, right panel). In the low ER group, the regression coefficient is not significantly different from zero for any $c > 0$. Exceptionally, at $c = 0$ the regression coefficient is significantly positive ($P < 0.05$), suggesting at face value that tam might actually harm patients in this subgroup (see also Figure 7.5).

The predominant impression from Figure 7.6 is of the instability of P-values from cutpoint analyses, especially in the important range [0, 10]. Consequently, it is hard to assess how strongly the study supports an influence of er on the effect of tam. The regression coefficients in the ER subgroups show no effect of tam in the low ER group (except possibly at $c = 0$, as just noted), whereas in the high ER group there is a reduction in the log hazard ratio of about 0.5. Although the interaction is of questionable statistical significance, the treatment effect in the high ER group does appear to be real. For higher values of c, a substantial proportion of patients with positive ER values enters the low ER group, causing a small trend towards a tam effect in this subgroup. No allowance has been made in this analysis for the multiple testing implicit in the use of 21 cutpoints, nor for the fact that for extreme cutpoints the subgroups may be small and the estimated regression coefficients imprecise.

7.8 STEPP

A more exploratory approach for the investigation of interactions, called the 'subpopulation treatment effect pattern plot' or STEPP, is based on dividing the observations into overlapping subgroups defined with respect to z_2 and estimating the effect of treatment z_1 separately within each subpopulation (Bonetti and Gelber, 2000). To create the subpopulations, sliding window (SW) and tail-oriented (TO) variants were proposed. STEPP was extended and further illustrated by Bonetti and Gelber (2004). The SW version of STEPP is a type of moving average and is also related to varying-coefficient models (Hastie and Tibshirani, 1993).

To increase the number of patients that contribute to each point estimate, subpopulations are allowed to overlap. This increases the precision of the individual estimates. The subpopulations may be defined in two different ways (SW and TO), as indicated in Figure 7.7 The horizontal axis in Figure 7.7 indexes the various subpopulations for which treatment effects are estimated, and shows the range of covariate values used to define the cohort of patients included in each subpopulation. The TO version has the overall population as the centre group. With increasing distance from the centre, more and more patients with high covariate values (to the left side) or low covariate values (to the right side) are deleted. Subpopulations in the SW version have an overlapping part and a part that differs from neighbouring subpopulations. The number of subpopulations and the percentage of overlapping patients are important parameters of STEPP. To define the size of the subpopulations, the SW version has two parameters: n_1 and n_2. A subpopulation must have at least n_2 patients, of which at least $(n_2 - n_1)$ patients are required to be different between neighbouring subpopulations. The degree of discreteness of a continuous variable determines the size of a subpopulation. The TO version has a parameter g giving $(g - 1)$ subpopulations, where patients with larger values are eliminated and $(g - 1)$ subpopulations excluding patients with smaller values. For further details of how the subpopulations are created, see Bonetti and Gelber (2000).

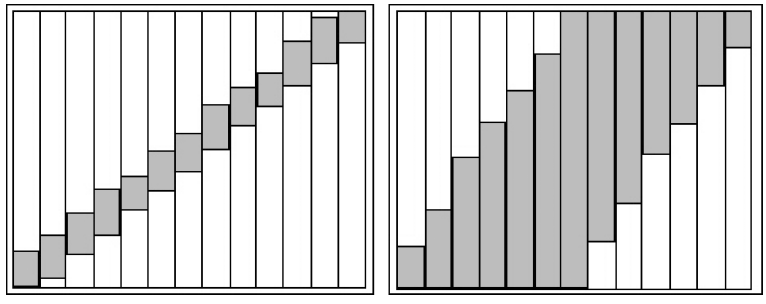

Figure 7.7 Schematic depiction of the two sets of subgroups used in STEPP. SW (left) and TO (right). The horizontal axis indexes the various subpopulations for which treatment effects are estimated, and shows the range of covariate values (vertical axis) used to define the cohort of patients included in each subpopulation. The TO version has the overall population as the centre group. (Adapted from Sauerbrei et al. (2007d) with permission from Elsevier Limited.)

The estimated treatment effects in the subpopulations defined by z_2 should be similar to the treatment effect in the overall population if z_2 has no influence on the treatment effect, i.e. no interaction exists between z_1 and z_2. Plots showing the estimated treatment effect with corresponding CIs in the subpopulations and tests based on the deviation of treatment effects in the subpopulations from the corresponding estimate in the overall population may be used to investigate an interaction between z_1 and z_2. For z_1, each subpopulation is represented by its mean. For more details see Bonetti and Gelber (2004).

7.9 EXAMPLE 3: COMPARISON OF STEPP WITH MFPI

7.9.1 Interaction in the Kidney Cancer Data

Royston et al. (2004) used the kidney cancer data from the MRC RE01 trial (see Appendix A.2.9) with the MFPI procedure and found a significant interaction between treatment `trt` and white cell count `wcc`. Altogether, they considered 10 potential predictive factors, of which six were continuous. Interest centres on the `trt` × `wcc` interaction.

Figure 7.8 displays the results of several STEPP analyses of the interaction between treatment `trt` and white cell count `wcc`. It is clear that use of small subpopulations ($n_1 = 25$, $n_2 = 40$; upper left-hand plot) with the SW method results in considerable variation caused by overfitting the data. Increasing the sample size in each subpopulation reduces the variation and leads to treatment estimates which show a similar dependence on `wcc`. For example, for $n_1 = 50, n_2 = 80$, there is only one additional 'blip' for `wcc` around 7. The lower panel clearly indicates that results from the TO version are less noisy and, hence, easier to interpret. The graph with $g = 4$ (lower right-hand plot) clearly indicates an interaction, and can be regarded as a rough approximation to the treatment effect function from MFPI (see Figure 7.9).

7.9.2 Stability Investigation

(In)stability, loosely defined as a vulnerability of modelling results to small changes in the data, is a critical issue when working with flexible models (Breiman, 1996b). See Chapter 8

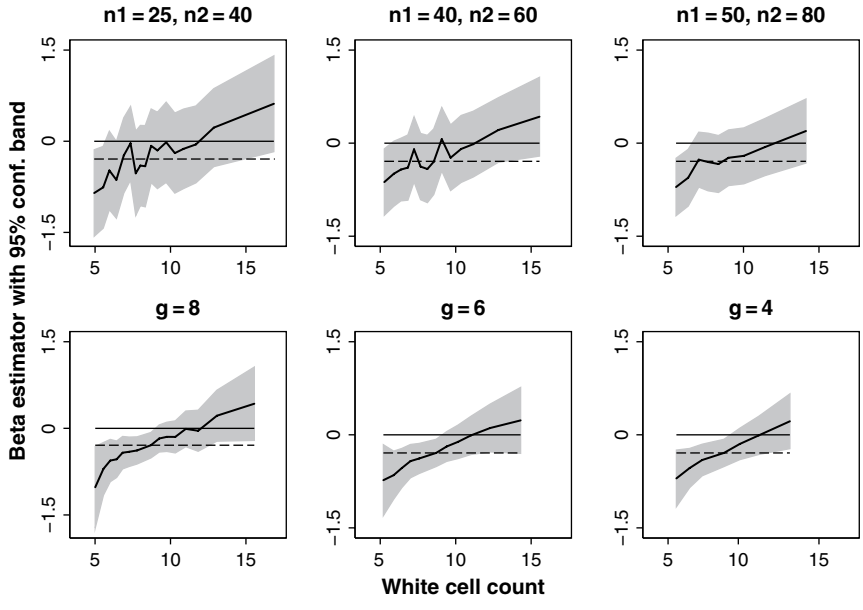

Figure 7.8 Kidney cancer data. STEPP plots for the interaction between `trt` and `wcc`, constructed with several choices of parameter values. Upper panel: SW; lower panel: TO. The plotted points represent the estimated treatment effects in each subgroup, with 95% CIs (faint dashed lines). The solid horizontal lines at zero represent no treatment effect and the dashed lines the estimated overall treatment effect. (Adapted from Sauerbrei et al. (2007d) with permission from Elsevier Limited.)

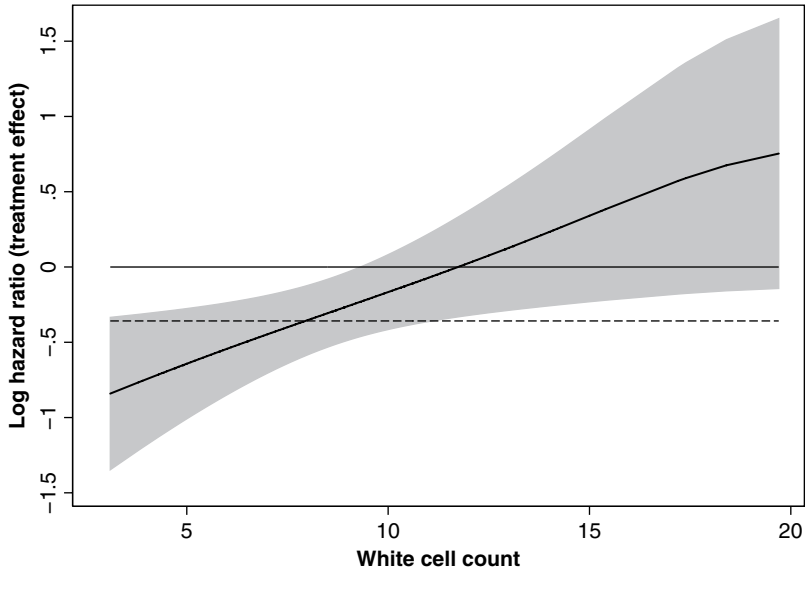

Figure 7.9 Kidney cancer data. Treatment effect plot for `wcc` (white cell count) from an MFPI analysis. Estimated treatment effect with pointwise 95% CI. The dashed line denotes the overall effect of treatment. (Adapted from Sauerbrei et al. (2007d) with permission from Elsevier Limited.)

for more details and examples. When estimating a treatment effect function for a continuous covariate, a small number of influential points may drive the function, thus indicating an interaction which is mainly a result of overfitting the data.

To explore the stability of MFPI, the treatment effect function was estimated in bootstrap replications in the kidney cancer example. Included as potential confounders were all variables other than the one under investigation (wcc). MFP was used to select the confounder model with a nominal significance level of $\alpha = 0.05$. wcc is the only variable identified by MFPI as a predictive factor. The stability of the function chosen was investigated, adjusting for the other variables. As in Royston and Sauerbrei (2003), a mean function and empirical bootstrap CIs were determined. Bootstrap resampling was also used to investigate the stability of STEPP, but without adjusting for covariates. For each STEPP group, the bootstrap mean and empirical 95% CI for the treatment effect were computed, using SW and TO variants.

The stability of the treatment effect function and of the corresponding STEPP functions were assessed in 1000 bootstrap samples. For the MFPI analyses, the multivariable adjustment model was selected, and then the FP2 treatment effect function for wcc (with preliminary robustness transformation; see Section 5.5). Selecting a new adjustment model in each bootstrap replication increases instability.

As an illustration, 20 randomly selected curves from MFPI analyses are shown in the left-hand panel of Figure 7.10. Most of the individual curves are similar to the curve from the original analysis presented in Figure 7.9. The mean of the 1000 bootstrap replications gives a nearly identical curve, with small differences appearing for more extreme values (wcc < 5

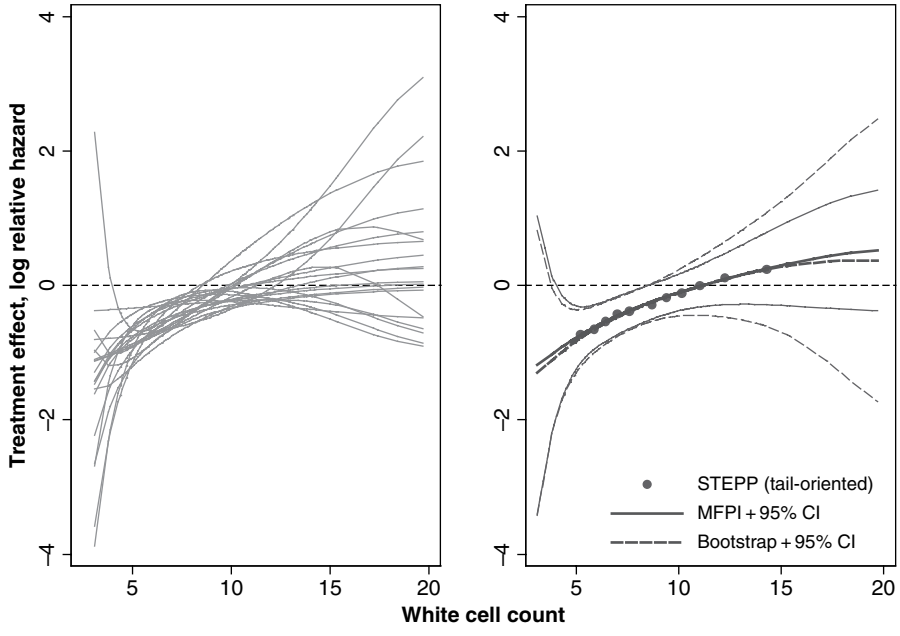

Figure 7.10 Kidney cancer data. Bootstrap analysis of stability of interaction between trt and wcc. Left panel: treatment effects plots in 20 bootstrap replications, using MFPI with data-driven adjustment model. Right panel: mean and 95% CI of treatment effect function from 1000 bootstrap replications, also showing the treatment effect function on the original data with 95% CI, and TO variant of STEPP. (Adapted from Sauerbrei et al. (2007d) with permission from Elsevier Limited.)

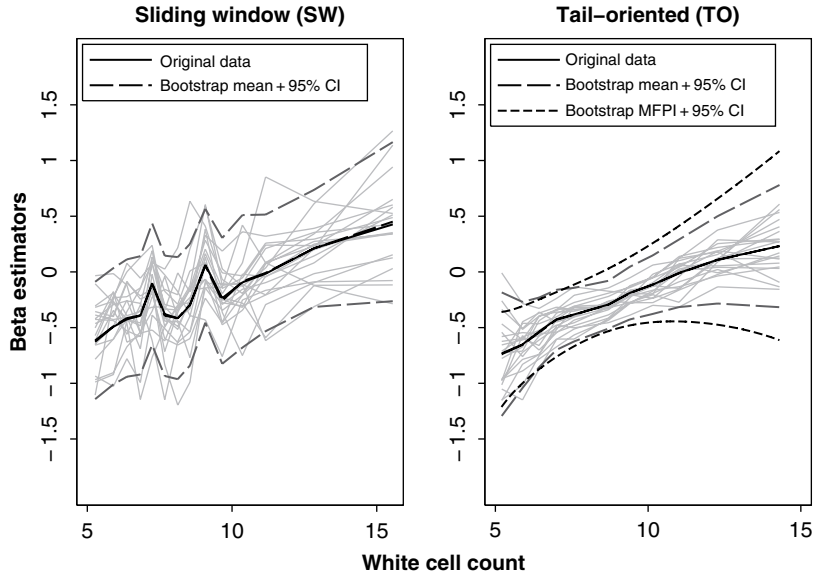

Figure 7.11 Kidney cancer data. Results from 1000 bootstrap replications of STEPP analysis of `wcc`. Left panel: SW ($m = 40$, $n = 60$); right panel, TO ($g = 6$). Thick lines represent the original data, bootstrap mean and 95% CI. Note that the bootstrap mean and original estimates are indistinguishable. Thin lines are results from 20 bootstrap replications selected at random. (Adapted from Sauerbrei et al. (2007d) with permission from Elsevier Limited.)

or > 15). The estimated effects from 11 subpopulations using the TO version of STEPP with $g = 6$ agree very closely with the functions from MFPI. For the bulk of the distribution of `wcc` values, the 95% pointwise CI derived from the 1000 bootstrap replications is a little wider than the interval from the original analysis. For larger values (say, > 12), the data become sparse and the bootstrap intervals become much wider, reflecting greater uncertainty in the FP2 functions selected by MFPI.

Figure 7.11 shows a random sample of 20 curves from 1000 bootstrap replications using STEPP with $n_1 = 40$, $n_2 = 60$ (SW) or $g = 6$ (TO). Major instability is apparent for the SW version. Functions for the TO version are more variable than the functions from MFPI, but the trend representing the result from the original analysis stands out clearly. For TO, the bootstrap interval is narrower than the corresponding interval from MFPI, where selection and estimation of the treatment effect function was repeated in each bootstrap replication. The wider intervals are partly due to the additional variation introduced in the MFPI analyses by selecting the adjustment model. Note that with STEPP the range of the covariate is restricted in the tails by the grouping process. No information about the treatment effect function is available beyond the range of the mean covariate values in the extreme groups.

7.10 COMMENT ON TYPE I ERROR OF MFPI

Little is known about the type I and type II errors of MFPI. In one example (Sauerbrei et al., 2007d), the values of the continuous variable haemoglobin in the kidney cancer dataset were

repeatedly permuted at random. Independence of the effects of treatment and the continuous variable was thereby simulated. Using 1000 permutations, the distribution of P-values from a test of interaction was close to uniform on $(0, 1)$ (data not shown). In 54 of the 1000 permutations the P-value was < 0.05, showing that the type I error of the MFPI procedure was close to its nominal level.

7.11 CONTINUOUS-BY-CONTINUOUS INTERACTIONS

As has already been discussed and illustrated, discrete-by-continuous interactions involving randomized treatments are particularly important in clinical trials. The topic of continuous-by-continuous interactions is also of interest, perhaps having more relevance to observational studies than to trials. A popular approach is to assume linearity for both variables and test the multiplicative term for significance. The model may fit poorly if one or both of the main effects is nonlinear (Cohen et al., 2003).

Despite its importance in many areas, modelling of interactions seems often to be ignored in practical analyses (Ganzach, 1998). Uncertainty about how to proceed in the context of multivariable modelling may be a reason. Even in the simplest case, the decision on whether to include an interaction requires a comparison between additive models $\beta_1 z_1 + \beta_2 z_2$ and interaction models $\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2$. Three questions immediately arise:

1. Is the sample size sufficient to allow detection of a 'nonnegligible' interaction?
2. Is the assumption of linear effects of z_1 and z_2 justifiable?
3. Is the increased complexity of the model resulting from including the interaction worthwhile?

Certainly, the presence of an interaction complicates the interpretation and presentation of the model.

Because of lack of power, a small sample size may result in erroneous rejection of the interaction (Greenland, 1983). Incorrectly assuming linear effects for z_1 and z_2 may lead to a wrong decision to include the interaction. The severity of the problem and the probability of selecting the interaction model instead of a model with nonlinear terms for z_1 and/or z_2 depend on several factors. They include the correlation between z_1 and z_2, the magnitude of β_1, β_2 and β_3, the sample size, and possible measurement errors in z_1 and z_2. MacCallum and Mar (1995) reported a large simulation study looking at factors influencing the chance of selecting the correct model. Whereas the interpretation of estimates from an additive model is straightforward, e.g. increasing z_1 from a to b leads to an increase in $E(y)$ of $\beta_1(b - a)$, it is much harder to interpret the increase in an interaction model where it depends on z_2, e.g. $(\beta_1 + \beta_3 z_2)(b - a)$. The meaning of β_1 changes when the product term is included (Greenland, 1989).

In real analyses, more than two predictors must be considered, increasing the difficulties. If one variable, z_1 say, is of particular interest, e.g. an exposure variable in an epidemiological study, then the analysis strategy is more straightforward. The main interest resides in the estimation of a satisfactory functional form for z_1, adjusted for possible confounders. The potential addition of interactions between z_1 and confounders can be done in a second step. If a product term is to be included, then so also should be the relevant main effects (the hierarchy principle: see Bishop et al. (1975)). Including several interactions produces a complex and

perhaps uninterpretable model. Subject-matter knowledge should be taken into account when considering such model extensions, and could result in adding nonsignificant interaction terms (Pearce and Greenland, 2005, p. 388).

For general model building without any particular variable of interest, all two-way interactions may be considered. Since the main-effects model may include nonlinear functions, products of these functions should be included as candidates. To handle this, we propose a procedure which is a natural extension of MFPI. More model building and data-dependent decisions are involved, leading to possible overfitting. Several 'reality checks' are needed to assess whether or not significant interactions are artefacts of mismodelling. The aim should be to select models which are 'as simple as possible' but which nevertheless show the important dependencies in the data. Most of the 'significant' interactions are not expected to reveal any crucial aspect of the data and, hence, are not required in a final model. However, a detailed search for interactions may improve model fit and may display interesting features of the data. Whether these features arise by chance or are reproducible must be assessed in external validation studies.

7.11.1 Mismodelling May Induce Interaction

Mismodelling a curved regression relationship, say as linear, may induce a spurious interaction with another variable (Lubinski and Humphreys, 1990; MacCallum and Mar, 1995). We illustrate this effect by considering predictors of 10-year all-cause mortality `all10` in the Whitehall I dataset in a logistic regression analysis. In an MFP model for `all10`, both `age` and `weight` are significant predictors of outcome. We consider only these two predictors and their possible interaction. MFP selects a model with linear `age` and an FP2 transformation of `weight` with powers $(-1, 3)$. The interaction between `age` and `weight` is computed by including the multiplicative terms $\texttt{age} \times \texttt{weight}^{-1}$ and $\texttt{age} \times \texttt{weight}^{3}$ in the logistic model. The likelihood ratio test (two d.f.) for this interaction has $\chi^2 = 5.27$ ($P = 0.07$), i.e. is not significant at the 0.05 level. However, if `weight` is erroneously modelled as linear, then $\texttt{age} \times \texttt{weight}$ is highly significant ($\chi^2 = 8.74$, $P = 0.003$), suggesting that mismodelling of the main effect of `weight` has induced a spurious interaction.

We first show that there is no strong interaction between `age` and `weight` when `weight` is modelled as a nonlinear function. To check the model and illustrate the effect, we divide `age` into four equal (quartile) groups and compute a running line smooth of the binary outcome `all10` on `weight` in each age group. Because the definitive analysis involves a logistic regression model, the smoothed probabilities are transformed to the logit scale and the results plotted against `weight`. With no interaction between `age` and `weight`, we would expect the curves to be roughly parallel. The results (see Figure 7.12) show that the logits of the smoothed probabilities are indeed approximately parallel across age groups, suggesting no (strong) interaction. Also shown in Figure 7.12 are the estimated slopes from logistic regressions on `weight` in each age group, erroneously assuming linearity. The lines are clearly not parallel. The sign of the slope even changes across age groups, explaining why a significant interaction is found when the effect of `weight` is assumed linear.

Other types of mismodelling, such as omission of correlated influential variable(s), may also introduce spurious interactions.

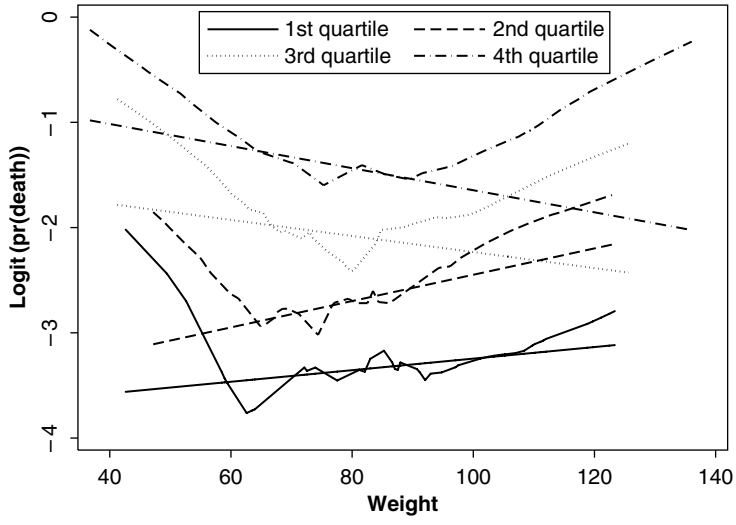

Figure 7.12 Whitehall I data. Graphical exploration of possible interaction between `age` and `weight`. The four pairs of lines show running line smooths (jagged lines) and linear fits (straight lines) in each of the four quartile groups by `age`. The changing slopes of the linear fits erroneously indicate an interaction, which disappears when a nonlinear function of `weight` is fitted.

7.11.2 MFPIgen: An FP Procedure to Investigate Interactions

With MFPI, a model for discrete z_1 and continuous z_2 is determined by finding the best FP transformation of z_2 and forming multiplicative interaction terms with the dummy variable(s) for z_1. A joint test of interaction involving the FP terms for z_2 and all dummy variables for z_1 is performed.

For general z_1 and z_2 (both possibly continuous) and confounders $\mathbf{x}$, a new procedure called MFPIgen, in the same spirit as MFPI, is as follows:

1. Apply MFP to $\mathbf{x}, z_1, z_2$ with significance level α^* for selecting members of $\mathbf{x}$ and FP functions of continuous variables. Force z_1 and z_2 into the model and apply the FSP to them. In the notation of Chapter 6 (see Section 6.1), MFP(α^*, α^*) is applied to $\mathbf{x}$, while simultaneously MFP(α_1, α_2) with $\alpha_1 = 1$ and chosen α_2 is applied to z_1, z_2. This step requires a single run of MFP.
2. Multiplicative interaction terms are calculated between the FP transformations selected for z_1 and z_2, or between untransformed z_1 and z_2 if no FP transformation is needed. For example, if both variables need FP2 transformation, then four interaction terms are created.
3. The model selected on $\mathbf{x}, z_1, z_2$ is refitted with the interaction terms included. The latter are tested in the usual way using a likelihood ratio test. If k interaction terms are added to the model, then the interaction χ^2 test has k d.f. For example, if FP2 functions were selected for both z_1 and z_2, then $k = 2 \times 2 = 4$.
4. All pairs of predictors are considered for possible interaction, irrespective of the statistical significance of the main effects in the MFP model. If z_1 and/or z_2 is binary or forced to be linear, then the procedure simplifies to the usual situation. If z_1 and/or z_2 are categorical,

then joint tests on all dummy variables are performed. An option is to treat the dummy variables as separate predictors.
5. All interactions should be checked for artefacts (see Section 7.4.2) and ignored if they fail the check.
6. If more than one interaction is detected, then a forward stepwise procedure can be used to extend the main-effects model.

There is one difference between this algorithm, MFPIgen, and MFPI. In MFPI, the confounder model $\mathbf{x}^*$ is selected independently of z_1 and z_2, whereas a joint model is selected in MFPIgen. The reason for the difference is that MFPI is principally intended for use with data from a randomized trial in which the effect of the treatment covariate z_1 is by design independent of other covariate effects. Therefore, adjustment by $\mathbf{x}^*$ is less important (see, for example, Table 7.1). In observational studies, however, it may be necessary fully to adjust the effects of z_1 and z_2 for confounders before investigating their interaction.

Since MFPIgen addresses dozens of potential interactions, multiple testing is a major issue. Results must be checked in detail and interpreted cautiously as hypotheses only. See Section 7.3 for further comments.

7.11.3 Examples of MFPIgen

Simplest Case

We consider the simplest case of a continuous-by-continuous interaction, i.e. that of a continuous outcome and two continuous covariates. Cohen et al. (2003) describe a study in 250 individuals of the intention to quit smoking y as a function of a measure of the fear of health ill-effects of smoking x and of self-efficacy for quitting smoking z. The correlation between x and z is 0.3. On applying MFPIgen, linear functions are selected for x and z and the $\mathtt{x} \times \mathtt{z}$ interaction is significant ($P = 0.008$).

Figure 7.13 shows the relationship between y and z in the four quartile groups of x. The relationship between y and z appears linear in each group. The regression slopes on z are nearly the same in the first and second quartile groups but subsequently increase with x. The largest slope is seen in the highest quartile group.

The MFPIgen model selected for this dataset is $E(\mathtt{y}) = \beta_0 + \beta_1 \mathtt{x} + \beta_2 \mathtt{z} + \beta_3 \mathtt{xz}$, whereas Cohen et al. (2003) included additional quadratic terms $\mathtt{x}^2$ and $\mathtt{x}^2\mathtt{z}$. Although $\mathtt{x}^2$ and $\mathtt{x}^2\mathtt{z}$ are jointly significant at the 0.05 level, they complicate the interpretation of the model and add little to R^2, which increases from 0.52 to 0.54.

More Complex Cases (1): Prostate Cancer Dataset

Table 7.6 shows the results of applying MFPIgen to all 21 pairs of variables in the prostate cancer dataset. In principle, each of the 21 analyses could involve a different main-effects model, to which is added one interaction term. Consider, for example, the interaction between age and pgg45. MFP is applied to all variables, with age and pgg45 forced into the model and the other variables selected at the 0.05 significance level. FP functions for all variables are selected in the usual way at the 0.05 level. Finally, the interaction between the selection functions for age and pgg45 is tested.

Linear terms were selected for all variables considered in interactions except for cavol, for which a log transformation was always selected. Six and two of the 21 interactions are significant at the 0.05 and 0.01 levels respectively, the strongest being $\mathtt{pgg45} \times \mathtt{cp}$ ($P = 0.003$).

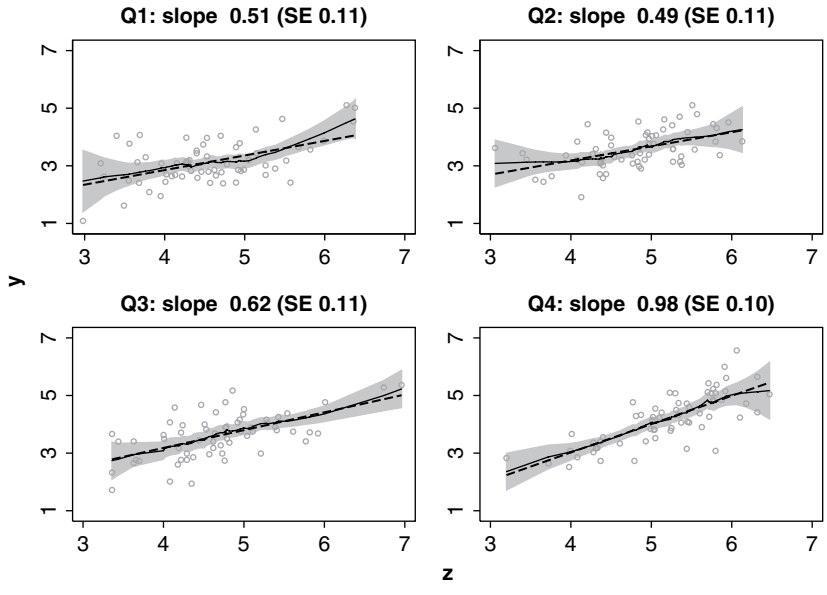

Figure 7.13 Quit-smoking study. Investigation of x by z interaction in four quartile groups of x (Q1–Q4). Solid lines, running line smooth with pointwise 95% CI; dashed line, fit from linear regression.

Table 7.6 Prostate data. Analysis of two-way interactions. Entries are *P*-values from tests of interaction using the MFPIgen procedure.

Variable	age	pgg45	cavol[a]	bph	cp	weight	svi
age	–						
pgg45	0.3	–					
cavol[a]	0.2	0.2	–				
bph	0.8	0.2	0.1	–			
cp	0.04	0.003	0.2	0.006	–		
weight	0.9	0.06	0.03	0.1	0.09	–	
svi	0.3	0.06	0.9	0.02	0.6	0.03	–

[a] cavol log transformed, all other variables untransformed.

Checking the Model The possible effect of influential observations on the six interactions age×cp, pgg45×cp, bph×cp, log cavol×weight, svi×bph and svi×weight was assessed. DFBETA (Belsley et al., 1980) is a measure that quantifies the effect of removing an observation on the regression coefficient of a predictor. Leverage is a measure of the influence of an observation on the predicted values from the model. Points that appear to be outliers in a scatter plot of absolute values of DFBETA against leverage may indicate undue influence of the relevant observations. Figure 7.14 shows such plots for the multiplicative interaction terms in the six models under consideration. Observation 47 appears influential for interactions (a), (b), (c) and (e). Observation 89 appears influential for (d) and (f). Observations 47 and 89 were omitted and all six interactions were retested. The resulting *P*-values were 0.2, 0.08, 0.1, 0.1, 0.3 and 0.2 respectively. From this it may be concluded that there are no important interactions in the dataset.

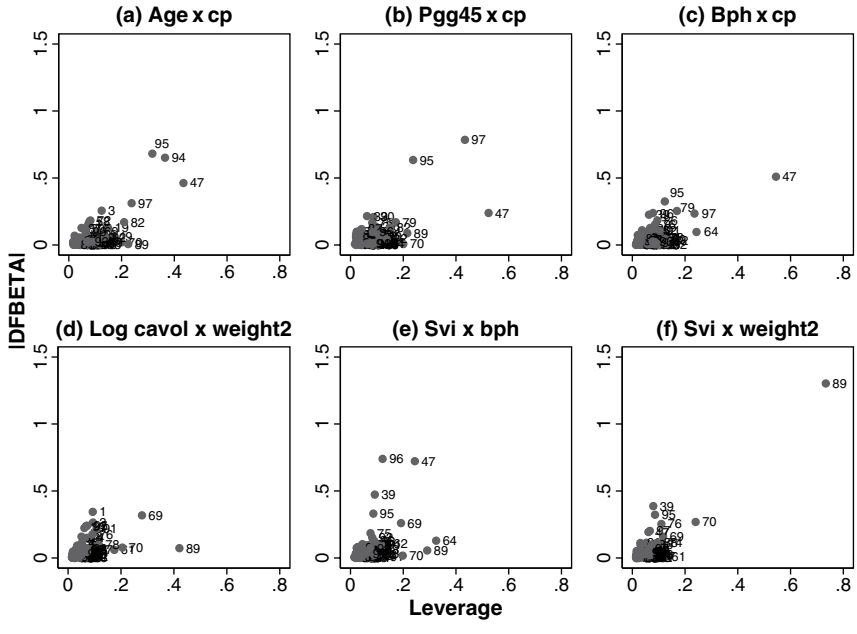

Figure 7.14 Prostate data. Influence plots for six interactions significant at the 0.05 level according to MFPIgen. Observation numbers are used to label points.

More Complex Cases (2): Whitehall I

An analysis of the CHD outcome in the Whitehall I study was described in Section 6.7.3. Here, we focus on analysing 10-year all-cause mortality `all10`, a binary outcome, using logistic regression.

An MFP(0.05) analysis of the covariates `cigs`, `sysbp`, `age`, `height`, `weight`, `chol` and `grade` (four-level categorical variable, three dummy variables tested jointly) gives FP powers 0.5 for `cigs`, $(-2, -2)$ for `sysbp`, $(-2, 3)$ for `weight` and linear terms for the other continuous variables. Main effects for all variables are required; therefore, the adjustment model used in assessing interactions in turn is always the same. Pairwise interactions among these variables are now considered (see Table 7.7). The `grade` × `weight` interaction is

Table 7.7 Whitehall I data. Analysis of two-way interactions. Entries are P-values from tests of interaction using the MFPIgen procedure.

Variable	`cigs`[a]	`sysbp`[a]	`age`	`height`	`weight`[a]	`chol`
`cigs`[a]	–					
`sysbp`[a]	0.7	–				
`age`	0.9	0.2	–			
`height`	0.1	0.5	1.0	–		
`weight`[a]	0.9	0.5	0.1	0.4	–	
`chol`	0.2	0.07	0.001	0.8	0.2	–
`grade`	0.2	0.2	0.2	0.2	0.04	0.4

[a] With FP transformation (see text), all other variables untransformed.

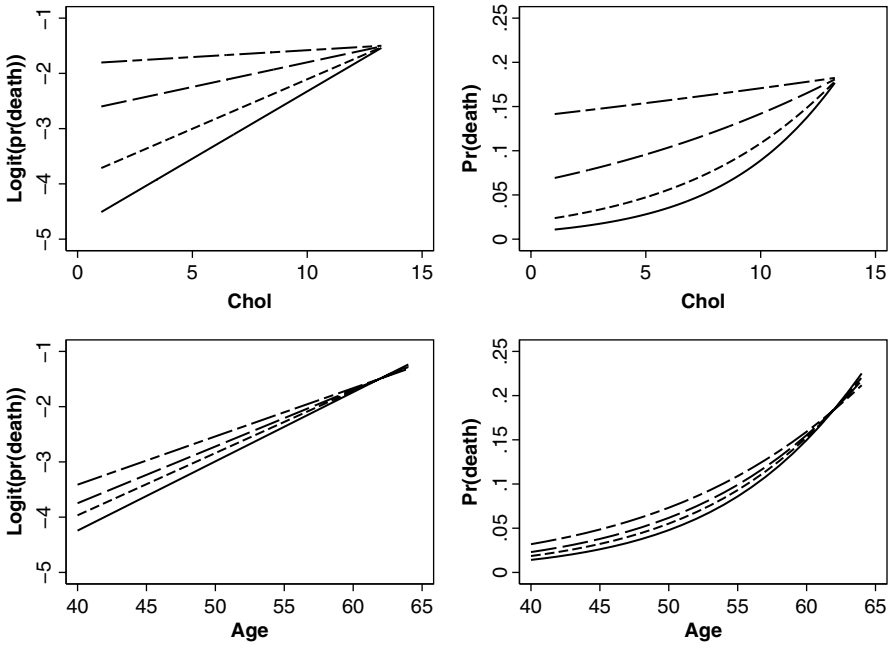

Figure 7.15 Whitehall I data. Graphical presentation `age` × `chol` interaction. Left-hand panels show the logistic scale, right-hand panels the probability scale. Upper panels show results for the 10th, 35th, 65th and 90th centiles of the distribution of `age`, lower panels for the same centiles of `chol`. Adjusted for other covariates.

significant at the 0.04 level but is not considered further. One interaction stands out: `age` × `chol`. The relationship between mortality predicted by the fitted interaction model and `chol` at the 10th, 35th, 65th and 90th centiles of the distribution of `age` is presented in Figure 7.15. The `age` by `chol` relationship is adjusted for other covariates centred on their means (see Section 4.11). The results show that the risk gradient on `chol` is much steeper in younger men than in older ones. The lines converge at a `chol` of 12 mmol l^{-1}. A very high cholesterol level (12 mmol l^{-1}) is associated with a probability of 10-year mortality of about 18%, irrespective of age. The lower right panel shows that the risk gradient on cholesterol disappears at about age 62 years.

The validity of this putative interaction was checked in four equal-sized age groups as follows. The slopes from logistic regression on `chol` were computed, adjusted for other factors in the MFP model (see above). The results are shown in Figure 7.16. The linearity assumption in each age group was checked by cubic regression splines (four d.f.) and confirmed to be well supported by the data. However, the slopes on `chol` are not monotonically ordered across the age groups, suggesting a possible lack of fit of the linear-by-linear interaction model.

In the MFP model, an FP1 term for `age` with power -1 is close to significant at the 0.05 level. To try to improve the fit, a model including $\texttt{age}^{-1}$, `chol` and their product was fitted. The deviance was reduced by 7.6 compared with that for the model with `age`, `chol` and their product. A common approach is to consider quadratic terms. A similar deviance reduction was obtained by adding quadratic interaction terms ($\texttt{age}^2$, $\texttt{age}^2$ × `chol`) to the linear by linear interaction model. The quadratic terms were highly significant ($\chi^2 = 9.67$, two d.f.,

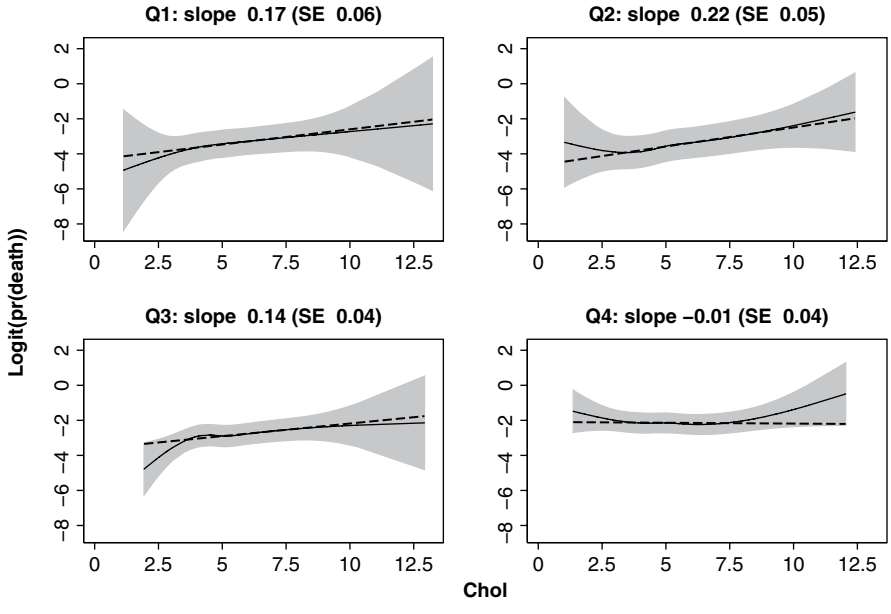

Figure 7.16 Whitehall I data. Interaction between `age` and `chol`. The relationship between `all10` and `chol` is explored in four equal-sized age groups. Solid lines: fit and 95% pointwise CIs from cubic spline regression with four d.f. Dashed lines: fitted line on `chol` from linear logistic regression model.

$P = 0.008$). Since the FP1 interaction model is more parsimonious and its fit is similar to the quadratic model, it is preferable.

Figure 7.17 repeats Figure 7.15 but includes the interaction between $\texttt{age}^{-1}$ and `chol`. The slight curvature on `age` can be seen, but the overall message from the revised model is not very different. It requires a large sample size to be able to distinguish between such subtly different models.

The example is instructive in that MFPIgen found a highly significant interaction between `age` and `chol`, which subsequent checking showed to be susceptible to some improvement. In a complete analysis of the dataset, the researcher might go on to consider other extensions of the `age` function and its interaction. The lesson to learn is that simple approaches to interaction modelling may not always be sufficient. Care is needed in checking the fit of the model and improving it if necessary.

7.11.4 Graphical Presentation of Continuous-by-Continuous Interactions

Since a two-way interaction involves three variables, a three-dimensional surface plot might be considered the preferred way to present it. However, such a plot may be hard to interpret, since its appearance depends critically on the orientation of the axes. Also, not all statistical software can produce three-dimensional graphics.

In our view, a better option is a 'sliced' plot, as in Figures 7.15 and 7.17. The fitted function at selected centiles of the distribution of the first variable is plotted against the second variable. The plot is repeated, reversing the roles of the two variables. In the Whitehall I example,

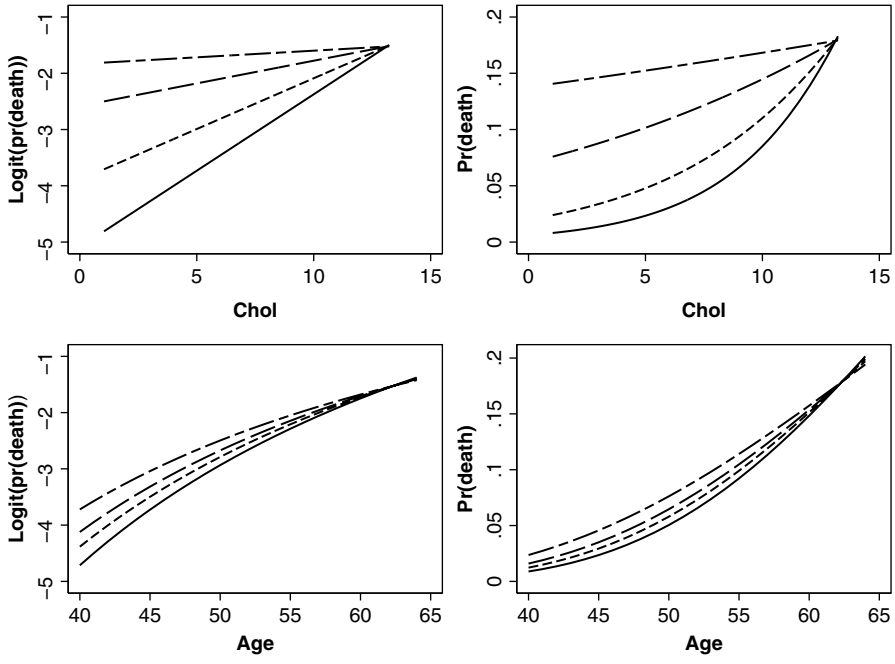

Figure 7.17 Whitehall I data. Graphical presentation of $\texttt{age}^{-1} \times \texttt{chol}$ interaction. Details as for Figure 7.15.

we chose to plot the fit at the 10th, 35th, 65th and 90th centiles. Probably four centiles is enough in most cases. Since distributions of observations are usually more 'bunched' around the centre than at the extremes, we selected two values near the extremes and two surrounding the median.

7.11.5 Summary

We have suggested MFPIgen, an extension to MFPI, for modelling continuous-by-continuous interactions in a multivariable context. We limit our analyses to first-order interactions because higher order interactions are unstable and hard to interpret. In the simplest case, all variables have a linear main effect. An interaction between two variables, z_1 and z_2, can be assessed by testing the estimated regression coefficient of the product $z_1 z_2$ for significance. Such product terms may introduce collinearity, resulting in difficulties in computation and interpretation. Centring continuous variables (e.g. on their means) improves matters. A more detailed discussion is given in standard textbooks, such as Vittinghof et al. (2005), DeMaris (2004), or Cohen et al. (2003).

As elsewhere in our book, we examine the data systematically for convincing evidence of nonlinearity of the main effects of continuous variables. By considering just two continuous variables in the Whitehall I study, we demonstrated that mismodelling a main effect by erroneously assuming linearity can introduce a spurious interaction. We propose graphical checks

of whether an interaction is genuine or may have arisen from mismodelling. Such checks are essential in a multivariable context where several potential interactions are considered.

For example, the interaction between `age` and `weight` in the Whitehall I example was caused by mismodelling the main-effect function for `weight` as linear when it should have been an FP2. The significance of several interaction terms in the prostate cancer data depended critically on two influential observations among the 97 patients. The analysis of 10-year all-cause mortality in Whitehall I revealed a linear-by-linear interaction for `age` × `chol` undetected in earlier analyses (including our own). Because of the large sample size (1670 events), the interaction is highly significant. Checks suggest that it is real, but also hint at a nonlinear effect of `age`. Even with its substantial power, the Whitehall study does not provide assurance as to whether the effect of `age` is linear or nonlinear.

In general, we hypothesize that many interactions remain undetected because nonlinearity is not considered or because the sample size is too small. Simulation studies are necessary to gain more insight into the characteristics (e.g. the power) of MFPI and MFPIgen.

7.12 MULTI-CATEGORY VARIABLES

The natural extension for an unordered multi-category variable z_1 with $k > 2$ levels is to estimate a function for each level, using the FP2 power terms from the main-effects model for z_2, adjusting for $\mathbf{x}^*$ (see Section 7.4). As before, the test of interaction compares deviances between the main-effects model and the extended model with parameters estimated for each category. The d.f. for the test are $2(k-1)$. Comparisons between functions at different levels depend on the reference category. Treatment-effect plots are constructed as for the binary case (see Section 7.3.6), one for each of the $k-1$ comparisons.

For an ordinal covariate, several options are available. The first option (which has serious drawbacks) is to ignore the ordering and proceed as above. The second option is to define scores (coding schemes) z_1^* for z_1 (see Section 3.3), analyse z_1^* as though it were continuous, and then use MFPIgen. If MFPIgen finds that an FP transformation of z_1^* improves the fit, then in effect the scores are redefined. Finally, the theory of trend tests and their associated contrasts may be used to define suitable tests based on the functions estimated for each level of z_1, as described above for the unordered case.

7.13 DISCUSSION

In this chapter, we have considered methods to identify interactions with continuous variables. Regarding treatment effects in randomized trials, current practice still seems to be either to assume linearity or to categorize the variable into a number of groups according to one or more cutpoints. The linearity assumption may be incorrect, and categorization reduces power and raises questions of the number and position of the cutpoints. MFPI can in principle be used for any binary variable (not just treatment).

In observational studies, modelling of confounder variables is important and a careful interpretation of the 'treatment effect' function is required. The confounder issue is not specific to our modelling approach, but relates to the fact that, in randomized trials, treatment is by design independent of all other covariates, whereas treatment and other such variables are almost always confounded by other factors in observational studies.

We emphasize the importance of several model checks to ensure that a postulated interaction is supported by the data and not a result of mismodelling or driven by a few influential observations. We also draw attention to the important distinction between a prespecified interaction and one that is identified in a data-dependent fashion. The strength of evidence for an interaction differs considerably between these two cases.

STEPP has been used several times in the literature with data from randomized trials to demonstrate interaction between treatment and a continuous covariate. The approach compares treatment effects in overlapping sub-intervals of the covariate. It may be considered as midway between the still-popular dichotomization with the comparison of treatment effects in two subgroups and our method, which estimates a treatment effect function. In one example using the TO version of STEPP with a smaller number of subgroups, the results from MFPI and STEPP agreed remarkably closely.

The principles of MFPI have been extended to investigate continuous-by-continuous interactions. We have shown in examples the importance of considering possibly nonlinear main effect functions. If nonlinearity is ignored by using the common approach of testing linear-by-linear product terms, then spurious interactions can be introduced into a model. Models derived with MFPIgen may not only fit the data better, but if an interaction term is not required, they may even give results which are easier to interpret. However, the sample size may not be large enough to discriminate well between several models with similar fits. Subject-matter knowledge plays an important role here.

For detailed discussion of various aspects of interactions between different types of variables, see for example DeMaris (2004, chapters 4 and 5) and Cohen et al. (2003, chapters 7 and 9). A fuller discussion of interaction among continuous variables, including higher order interactions, is given by Aiken and West (1991). A more flexible approach using semiparametric modelling for investigating interactions with continuous predictors is described by Ruppert et al. (2003, chapter 12). None of these books considers the FP framework.