Chapter 12
Assumptions in Regression Models: Additivity and Linearity

Background In this chapter, we discuss assessment of assumptions in multivariable regression models. Specifically, we consider the additivity assumption, which can be assessed with interaction terms. We also consider the linearity assumption of continuous predictors in a multivariable regression model, where multiple non-linear terms can be included to allow for non-linear relationships between predictors and outcome. Throughout we stress parsimony in strategies to extend a prediction model with interactions and non-linear terms, since better fulfillment of assumptions in a particular sample does not necessarily imply better predictive performance for future subjects. We consider several case studies for illustration of various strategies to deal with additivity and linearity.

12.1 Additivity and Interaction Terms

The generalized linear regression models discussed in this book all have a linear predictor at their core: $\text{lp} = \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_i \times x_i$, for models with i predictors.

The β_1 to β_i are the regression coefficients, referring to the main effects of predictors x_1 to x_i. This formulation implies additivity of effects. For a logistic regression model, we can calculate odds ratios as e^{β}; the odds ratios are multiplied to obtain the odds of the outcome. Hence, effects of predictors are assumed to be multiplicative on the odds scale. For a Cox regression model, e^{β} is the hazard ratio; the assumption is that these hazard ratios can be multiplied on the hazard scale.

The scale is essential for consideration of additivity. If a treatment reduces risk from 20 to 10% in one risk stratum, and from 10 to 5% in another risk stratum, the relative risk is 0.5 in both. The odds ratios are also quite similar (0.44 and 0.47, respectively). Hence, we could say that there is a consistent halving of the risk. But on an absolute scale, the benefit is clearly dependent on the risk (10% vs. 5% reduction).[237]

The most common regression modelling procedure is to start model specification with main effects of predictors only. Some epidemiological text books advice to consider interactions early in the modelling process, with main effects included for

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_12, © Springer Science+Business Media, LLC 2009

all variables that have a relevant interaction term.[234] Interactions between predictors can be considered by multiplicative terms of the form $x1 \times x2$ (two-way or first-order interactions), and $x1 \times x2 \times x3$ (three-way, or second-order interactions); higher-order interactions are uncommon to consider for regression models. The interpretation of a two-way interaction is that the effect of one predictor depends on that of another predictor. The effect is different, depending on the value of another predictor. The effect of a predictor cannot be interpreted alone; we need to know the value of another predictor to interpret its effect.

12.1.1 Potential Interaction Terms to Consider

As for main effects, prior subject knowledge may help to guide us to select interaction terms. For example, interaction terms that were identified in previous studies could be assessed. Clinical insights, e.g. on pathophysiology, are difficult to use, because using main effects in a model is assuming that predictors act in a multiplicative way on the risk scale (e.g. odds ratios and hazard ratios are multiplied). Reasoning why a certain combination of predictors would not act in an additive way on, e.g. the logodds scale, is quite difficult to imagine. Some researchers are motivated to study an interaction term when two predictors are correlated. But correlation does not imply anything on the effects of predictors conditional on each other. Two predictors may not have any correlation, but have interacting effects. Some types of interactions have been suggested that warrant consideration in prediction models (Table 12.1).[174]

12.1.2 Interactions with Treatment

Various interactions with treatment can be considered. The benefit of treatment may depend on the severity of disease, with less relative benefit for those with less severe disease. The reverse may also be true, especially in oncology, where less

Table 12.1 Examples of interactions to consider in clinical prediction models (based on Harrell[174])

Interaction	Effect
Severity of disease × treatment	Less benefit with less severe disease
Place × treatment	Benefit varies by treament centre
Place × predictors	Predictor effects vary by centre/region
Calender time × treatment	Learning curves for some treatments
Calender time × predictors	Increasing or decreasing impact of predictors over the years
Age × predictors	Older subjects less affected by risk factors; or more affected by certain types of disease
Follow-up time × predictors	Non-proportionality of survival effects, often a decreasing effect over time
Season × predictors	Seasonal effect of predictors

relative benefit occurs for those with more severe disease. For example, surgery in oesophageal cancer can be curative, but only for patients without distant metastases. Note that absolute benefit will anyway depend on the severity of disease, even when the relative benefit is constant. For example, Califf modelled the absolute benefit of tPA treatment for acute myocardial infarction patients in the GUSTO-I trial in relation to predictors. Benefit depended strongly on the risk profile, while it might be assumed that the relative effect of treatment was constant.[63] In addition to severity of disease, a treatment effect may depend on the setting, e.g. the centre where a patient was treated. This is especially the case when specific skills and facilities are required for the treatment. For example, surgical mortality is known to vary widely between centres for some procedures, such as resection of oesophageal cancer. Similarly, some treatments have a learning curve, which can be modelled by including a treatment × calender time interaction term, with calender time reflecting cumulative experience.

In randomized controlled trials, subgroup effects for treatment effects are often performed, e.g. whether treatment works better for older than younger patients. Such subgroup effects should be supported by an interaction test for difference in effect; not with one *p*-value for older and one *p*-value for younger patients.[339] Even when subgroup analyses are pre-specified, results should be cautiously interpreted because of multiple testing of the treatment effect. Multiple testing inflates the risk of false positive conclusions. Subgroup analyses are therefore best interpreted as secondary analyses which motivate further study. This is often not the case in current practice.[18]

*12.1.3 Other Potential Interactions

Predictor effects may differ by place and time, which would limit their generalizability (see Part III). Basic issues to consider are whether predictor definitions were consistent across centres and during time. In some individual patient data analyses, predictor effects were however surprisingly consistent, even when definitions varied over studies (e.g. studies in traumatic brain injury[271,277]). As might be expected, interactions of predictors by place of treatment were small within the GUSTO-I trial, where data were collected in a highly standardized and controlled way.[405]

Various aspects of "time" can interact with predictor effects: calender time (e.g. patients treated during years 1980–2005), age (e.g. 30–90 years), follow-up time (e.g. 0–10 years), and season (months January to December). The effects of predictors may change over the years because of improvements in treatment, or changing definitions. The effects of risk factors for developing cardiovascular disease are known to decrease with aging. Predictors having less effect in the elderly might be explained as that older subjects have proven to survive with the risk factors. For survival analysis, predictors are usually assumed to have proportional effects during follow-up, e.g. in the Cox proportional hazards model, but also in a Weibull model. Such proportionality of effects may not be tenable in the follow-up of

oncological patients, where relative risks of predictors for early events decrease with time, while others may increase. For example, non-proportional effects have been noted in breast cancer survival, with no effect of stage of disease after 10 years of follow-up.[308] The proportionality assumption is equivalent to assuming no interaction effects between predictors and follow-up time.

Furthermore, some predictors may have a different impact during the season, e.g. for infectious and respiratory diseases (Table 12.1). Other interactions may be relevant to consider in specific prediction problems. For example, sex-specific effects of predictors are commonly modelled in cardiovascular disease.

*12.1.4 Example: Time and Survival After Valve Replacement

A follow-up study was done spanning over 25 years for survival of patients after aortic valve replacement.[195] Various changes had taken place in case-mix between the first valve replacement (in 1967) and the latest replacement analysed (in 1994). During the 25+ years period, 1,449 mechanical valves were implanted. Overall early mortality (<30 days) was 5%, and was analysed with logistic regression. Overall survival rates at 5, 10, and 15 years were 80%, 63% and 49%, respectively. Poisson regression analysis was used to disentangle the effects of calender time, age, and follow-up. All three aspects of time appeared to be important. A substantial drop in both early and late mortality was identified around the introduction of cardioplegia (in 1997), but no strong interactions with calender time were found. A changing, non-proportional effect was observed for several prognostic factors during follow-up. For example, increasing effects during follow-up were found for older age ($p<0.05$), urgency (urgent operations and acute endocarditis) ($p<0.05$), and ascending aorta surgery ($p=0.12$). Early year of operation, male gender, and previous cardiac surgery (all $p<0.05$) were more important during early years of follow-up. The effects of concomitant coronary bypass surgery and concomitant mitral valve surgery were more or less constant during follow-up. This study illustrated that a Poisson regression model could be used to disentangle different aspects of time in a survival analysis. This model was more easily to work with compared to the Cox regression model.[195]

12.2 Selection, Estimation and Performance with Interaction Terms

In clinical prediction models with a typical number of predictors, say 5–10, the number of potential interactions is substantial. If interactions are considered, it has been suggested to first perform an overall interaction test.[174] We can also obtain partial overall p-values, e.g. for all interactions with age. If this p-value is low, we may consider proceeding with studying specific interactions for inclusion in the

model. This approach limits the multiple testing problem, at the price of lower power for including specific interactions. An alternative is to perform interaction tests for individual combinations of predictors, but use a rather stringent p-value, such as 0.01 for inclusion. We illustrate the problems with selection of interaction terms with a small subsample from the GUSTO-I study.

12.2.1 Example: Age Interactions in GUSTO-I

We study interaction with age in the relatively large subsample from GUSTO-I (sample5, n=785, 52 deaths). We first fit all interactions, and then perform and overall test based on the Wald statistics. The overall test has a p-value of 0.14; but the interaction AGE×HRT is statistically significant (p=0.03, not adjusted for multiple testing). Some might be tempted to include this interaction in the model. It appears that tachycardia (HRT) has a stronger effect at higher age (a positive interaction). Equivalently, we can state that age has more effect in strength in those with tachycardia (Fig. 12.1).

12.2.2 Estimation of Interaction Terms

A first distinction that epidemiologists like to make is between "qualitative" and "quantitative" interactions. A qualitative interaction means that a predictor has an opposite effect in one group vs. another group of patients. Quantitative interaction means that the effect of a predictor is in the same direction, but different in strength

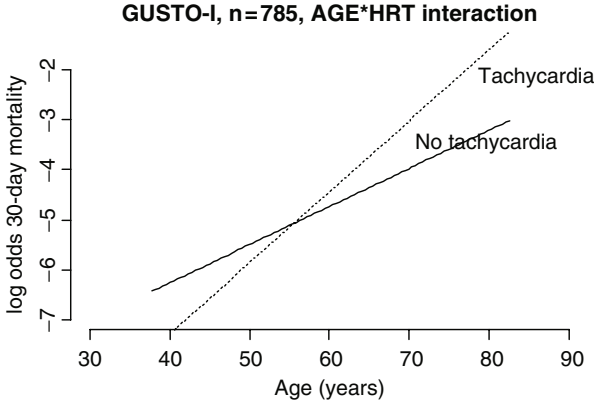

Fig. 12.1 Age by tachycardia interactions in a subsample of GUSTO-I (n=785, 52 deaths), revealing a positive interaction

in one group than another group of patients (see e.g. Fig. 12.1). This distinction is especially important when we aim to interpret the effects of predictors; we will more be tempted to include a qualitative interaction than a quantitative interaction. For predictive performance, the distinction between qualitative and quantitative interaction is less relevant.

Another issue is that we can have somewhat counterintuitive effects of interactions. For example, Fig. 12.1 suggests that the presence of tachycardia is protective for 30-day mortality at ages younger than 55. If we consider this implausible, we can code the interaction such that no effect of tachycardia is present below age 55 (Fig. 12.2). Admittedly, the age cut-point of 55 years is data-driven. But the general idea is that we incorporate subject-specific knowledge to prevent incorporation of random noise in the model.

More generally, we should use a smart coding for interaction terms once we decide to include an interaction term in a model. This is especially useful when we want to readily obtain standard errors and confidence intervals for predictors in interaction with other predictors.[122] The approach is to test for interactions in models with standard multiplicative terms of the form $x1 \times x2$. But we can estimate effects with a smarter coding of the form $x1 + (1 - x1) \times x2 + x1 \times x2$ instead of $x1 + x2 + x1 \times x2$. More details are on the book's web page.

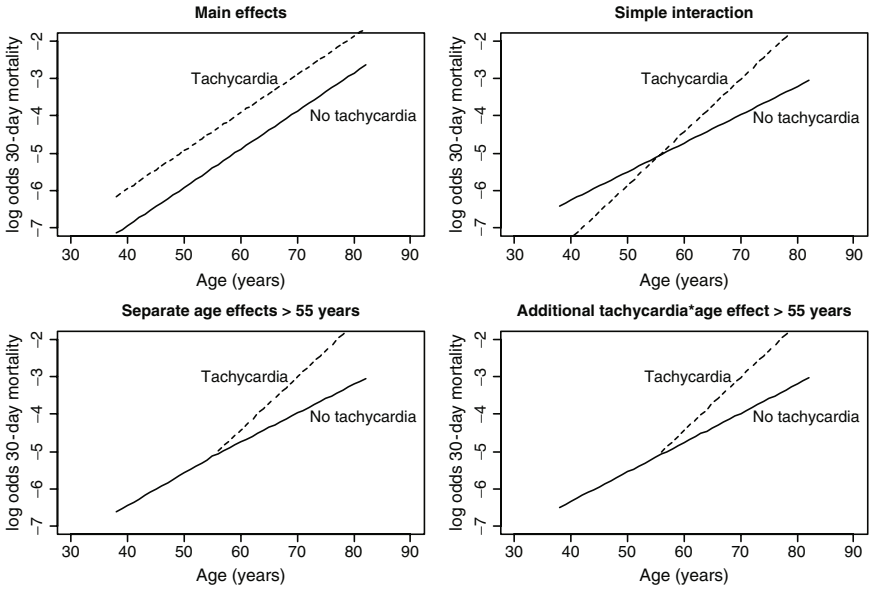

Fig. 12.2 Age by tachycardia relationships to 30-day mortality in a large subsample of GUSTO-I (n=785, 52 deaths). Panel (**a**) main effects only; panel (**b**) simple positive interaction; panel (**c**) separate effects for (no) tachycardia over age 55; panel (**d**) one age effect and an additional effect of tachycardia over age 55 years. The difference between panel **c** and **d** is barely notable, but in panel **c**, three age effects are estimated, while in panel **d** two age effects are estimated

12.2.3 Better Prediction with Interaction Terms?

We may wonder we predict better with the AGE×HRT interaction (Table 12.2). We hereto test the models as shown in Fig. 12.2 in a large, independent part of GUSTO-I (n=20,318). Surprisingly, we find that a model with the AGE×HRT interaction (Fig. 12.2b), performs worse than a model without this interaction term. The models without the counterintuitive effect of tachycardia below age 55 perform similar, both at apparent validation and at external validation in n=20,318. The explanation for this remarkable finding is in Fig. 12.3: the interaction between tachycardia and age was positive in the subsample, but negative in the independent validation part of GUSTO-I (less effect of tachycardia at older ages). This example illustrates that considering interaction in an unstructured way can damage predictive ability of a model.

Table 12.2 Performance of models developed in a subsample of GUSTO-I (n=785) in an independent part of GUSTO-I (n=20,318). The model with main effects contained eight dichotomized predictors

Model	*df*	Apparent (n = 785)	Validation (n=20,318)
Main effects	8	0.828	0.805
Main effects+AGE×HRT interaction	9	0.831	0.796
One age effect <55, 2 age effects ≥ 55	9	0.832	0.798
HRT effect only for age >55 years	8	0.832	0.798

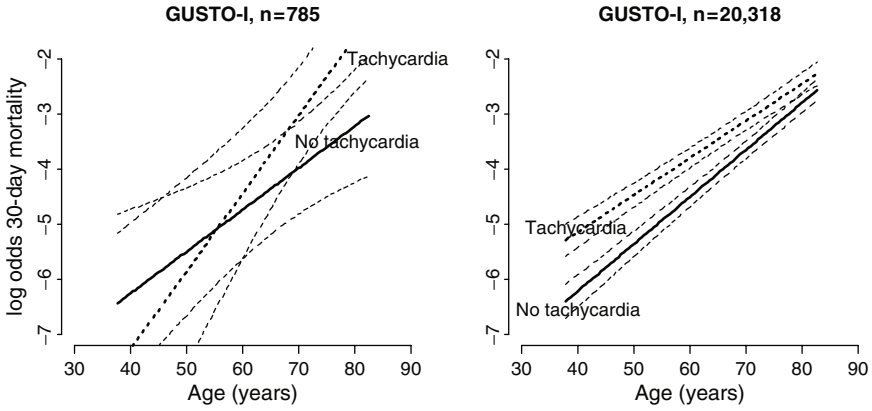

Fig. 12.3 AGE×HRT interactions in GUSTO-I. *Left panel*: positive interaction in a subsample (n=785, 52 deaths, p-value for interaction 0.10), negative interaction in an independent validation part of the GUSTO-I data set (n=20,318, 1,428 deaths, p-value for interaction 0.002). 95% confidence intervals are given around each line

12.2.4 Summary Points

- An interaction term indicates that the effect of a predictor depends on values of another predictor
- Interaction terms to consider in a prediction model depend on the context, but some types of interactions may warrant specific consideration
- For better interpretation, we may use a smart coding of interactions, and eliminate counterintuitive effects, e.g. that a predictor becomes protective for some patients
- The performance of a prediction model does not necessarily increase by including an interaction term
- Pre-specification of some interaction terms for a model may be preferable to exploratory determination of which terms to include

12.3 Non-linearity in Multivariable Analysis

We discussed the assessment of continuous predictor variables in Chap. 9 for the univariate case, where each predictor is considered separately. Harrell advocates to use restricted cubic spline functions to define transformations of continuous variables.[174,177] An RCS function consists of pieced-together cubic splines (containing x 3 terms) that are restricted to be linear in the tails. These functions have many favourable properties, such as appropriate flexibility combined with stability at the tails of the function. We can also consider multivariable modelling with fractional polynomials,[367] and with smoothing spline transformations (in multivariable generalized additive models ("GAM"), Table 12.3). The flexibility of a smoothing spline transformation in a GAM is determined by penalty terms, which relate to the effective degrees of freedom (*df*). There are presently two variants of GAM available with respect to choosing the effective *df* in a multivariable context. One variant is that the effective *df* are set by the analyst.[180] Alternatively, a generalized cross-validation (GCV) procedure can be used to define statistically optimal transformations for multiple continuous predictors in a GAM.[490] We discuss these approaches in more detail below.

12.3.1 Multivariable Restricted Cubic Splines (RCS)

An RCS requires the specification of knots, which can well be based on the distribution of the predictor variable.[174] The key issue is the choice of the number of knots: 5 knots implies a function with 4 *df*, 4 knots 3 *df*, and 3 knots 2 *df*. Although 5 knots are sufficient to capture many non-linear patterns, it may not be wise to include 5 knots for each continuous predictor in a multivariable model. Too much

Table 12.3 Approaches to non-linearity in multivariable clinical prediction models

Approach	Characteristic	Multivariable strategy	R implementation
Restricted cubic splines	Cubic splines, with restriction in shape at the ends of the predictor distribution	Keep complexity as defined a priori or based on findings in univariate/multivariable analysis	`rcs` in Design package
Fractional polynomials	Combine one or two polynomials	Search iteratively for optimal transformations	`fp` and `mfp` in `mfp` package
Splines in GAM	Spline functions with smoothing depending on effective degrees of freedom	Degrees of freedom set by analyst or from a generalized cross-validation (GCV) procedure	`gam` and `mgcv` package

flexibility would lead to overfitting. One strategy is to define a priori how much flexibility will be allowed for each predictor, i.e. how many *df* will be spent. In smaller data sets, we may for example choose to use only linear terms or splines with 3 knots (2 *df*), especially if no strong prior information suggests that a non-linear function is necessary.[174] Alternatively, we might examine different RCS transformations (5, 4, 3 knots) in univariate and/or multivariable analysis, and choose an appropriate number of knots for each predictor based on the findings in the data. It might be reasonable to choose the complexity of non-linear functions based on the χ^2 statistic of each predictor, with more flexibility for stronger predictors.

12.3.2 Multivariable Fractional Polynomials (FP)

As discussed in Chap. 9, fractional polynomials are formulated as a power transformation of a predictor x: x^p, where p is chosen from the set −2, −1, −0.5, 0, 0.5, 1, 2, 3. This defines 8 transformations, including inverse (x^{-1}), log (x^0), square root ($x^{0.5}$), linear (x^1), squared (x^2) and cubic transformations (x^3). In addition to these 8 FP1 functions, 28 FP2 functions can be considered of the form $x^{p1} + x^{p2}$; when $p1{=}p2$ one defines another 8 FP2 functions as $x^p + x^p \log x$, for a total of 36 FP2 functions.[367] FP1 and FP2 have 2 and 4 *df*, respectively.

Estimation algorithms have been developed for various software packages, including R.[366] The `mfp` algorithm applies a special type of backward stepwise selection procedure for the determination of reasonable functional forms for each continuous predictor. The algorithm starts with a full model including all predictors, with all continuous predictors in linear form. The predictors are considered in order of decreasing statistical significance, such that relatively important predictors are considered before unimportant ones.

For a particular continuous predictor, we may search within the 44 FP2 transformations for a best fitting function. The best transformation is compared to deleting the predictor. This procedure uses 4 *df* to test for inclusion of the continuous predictor, as having "any effect." If this test is significant, we may continue with a test for non-linearity: FP2 vs. linear, using 3 *df*. Finally, we tests an FP2 vs. FP1 transformation as a test of a more complex function against simpler one (2 *df* test for "simplification"). The functional form for this predictor is kept, and the process is repeated for each other predictor. The first iteration concludes when all the variables have been processed. The next cycle is similar, except that the functional forms from the initial cycle are retained for all variables excepting the one currently being processed. Updating of FP functions and selection of variables continues until the functions and variables included in the model do not change.[367]

This test procedure aims to preserve the overall type I error. The price is that we are slightly conservative if the true predictor–outcome relationship is linear, i.e. a straight line. This is because in step 1, we test for overall effect with 4 *df*, leading to lower statistical significance in case of a true linear relationship.

12.3.3 Multivariable Splines in GAM

In a GAM, flexible, smooth functions are defined for continuous predictors. The smooth functions can be defined by splines or other "basis functions."[490] To avoid overfitting we statistically penalize lack of smoothness ("wiggliness") using a smoothing parameter. The penalization reduces the effective degrees of freedom used by each continuous predictor. The optimal smoothness can be determined with prediction error criteria, e.g. in a Generalized cross-validation (GCV) procedure. Further details are provided by Wood[490] and Hastie.[181]

In multivariable modelling, splines in a GAM may well serve as a reference standard for comparison of simpler, parametric transformations, such as FP (or RCS) functions.[353] We compare several approaches in a case study below. In practice, one would not have to perform all of these transformations but choose one approach that one is familiar with.

*12.4 Example: Non-Linearity in Testicular Cancer Case Study

We aim to predict the presence of benign tissue only ("necrosis") in patients treated with chemotherapy for testicular cancer. We consider six predictors, of which three are binary (Teratoma, pre-chemotherapy elevated AFP, pre-chemotherapy elevated HCG), and three continuous (pre-chemotherapy LDH, reduction in mass size during chemotherapy, post-chemotherapy size). The LDH values were standardized by dividing by the upper limit of the local upper normal value ("LDHst" variable).

In initial univariate analyses, we used RCS functions to study non-linearity in the effects of the continuous predictors. Subsequently, we used simple parametric transformations, mainly based on visual assessment of the univariate RCS functions.[425] The chosen transformations were logarithmic for LDHst; linear for reduction in size; and square root for post-chemotherapy size (Fig. 12.4). We now explore the transformations chosen with other modelling strategies, including fractional polynomials and smoothing splines in generalized additive models.

We compare RCS, FP, and GAM functions with two bendings: FP2 transformations, RCS with 4 knots (3 *df*), and GAM splines with 3 effective *df*. For LDH, the transformations vary to quite some extent. The relationship of LDH to necrosis is rather different for a logarithmic transformation compared to other transformations. A simple linear term might also have been reasonable. This is supported by the FP procedure (Table 12.4). LDH has an effect (p-value for "any effect" = 0.02), but non-linearity was non-significant (p=0.48). For postchemotherapy size, the RCS, FP2, and GAM transformations agree much better visually (Fig. 12.4), and the square root transformation looks reasonable. The FP procedure indicates significant non-linearity (p=0.0002), and non-significant improvement by an FP2 function over an FP1 function (p=0.46). The chosen FP1 function is logarithmic rather than the square root. Finally, reduction in mass size seems to be fit adequately with a linear term. The RCS, FP2, and GAM transformations fluctuate around this straight line, with the most wiggly pattern for the GAM. The FP procedure confirms that there is no reason to include non-linear terms (p=0.64). The R code for these analyses is available at the book's web site.

Fractional polynomials were considered in univariate logistic regression analysis, and subsequently in three multivariable logistic regression models. A full model included three binary predictors (teratoma (yes/no, 1 *df*), elevated AFP (yes/no, 1 *df*), elevated HCG (yes/no, 1 *df*)), and three continuous predictors with FP2 functions (LDH standardized, reduction in size, post-chemotherapy size).

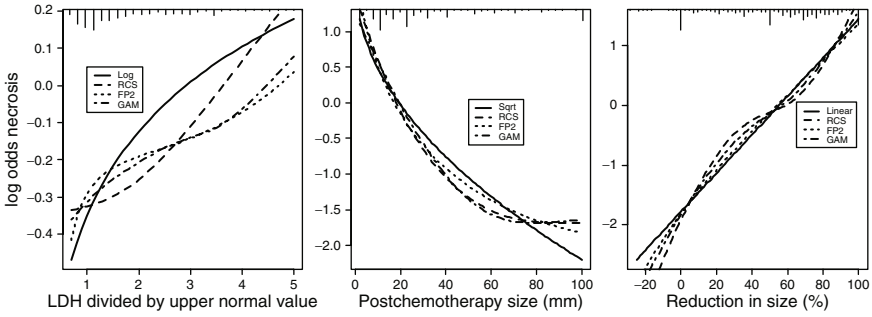

Fig. 12.4 Non-linearity in univariate analysis of LDH, post-chemotherapy size, and reduction in mass size. Curves are shown for a parametric approximation (log, sqrt, linear), restricted cubic spline functions with 4 knots (3 *df*), a fractional polynomial (4 *df*), and a generalized additive model with spline smoother (3 *df*). The distributions of values are shown at the top of the graphs

*12.4.1 Details of Multivariable FP and GAM Analyses

Multivariable fractional polynomials were fitted without selection ("full model," 3 *df* for dichotomous + 3×4 = 12 *df* for continuous predictors, in total 15 *df*), and with a variant of a backward stepwise selection algorithm (Table 12.4). The FP2 transformations were log(LDHst)+LDHst^3; 1/reduction+1/sqrt(reduction); and sqrt(postsize) + sqrt(postsize)×log(postsize). A multivariable FP procedure with $p<0.05$ for selection led to a model with linear terms for the three continuous predictors and three binary predictors (each of the six predictors $p<0.01$). All tests for non-linearity were non-significant (Table 12.4). Selection with $p<0.20$ led to a linear term for LDHst, 1/reduction, and log(postsize) in FP1 transformations. Post-chemotherapy size and reduction in size had *p*-values for non-linearity of 0.03 and 0.08, but FP2 transformations were not much better than FP1 transformations (*p*-values 0.46 and 0.27 respectively, Table 12.4).

*12.4.2 GAM in Univariate and Multivariable Analysis

For comparison, we examine the smooth functions selected as optimal with a GCV procedure (Fig 12.5). In univariate analysis, a (near) linear term is optimal for LDH and reduction in size (1.1 and 1 effective *df*). Post-chemotherapy size

Table 12.4 Fractional polynomial analysis of three continuous predictors in the testicular cancer data set (n=544)

	Predictor	*P*-value "any effect" (FP2 vs. no effect, 4 *df*)	*P*-value "non-linearity" (FP2 vs. linear, 3 *df*)	*P*-value "FP2" (FP2 vs. FP1, 2 *df*)	FP1	FP2
Univariate	LDH (standardized)	0.021	0.48	0.59	2	−2, 3
Full model		<0.0001	0.18	0.73	0 (=log)	**0 (=log), 3**
Stepwise $p<0.05$		0.0003	0.46	0.62	0.5	0 (=log), 3
Stepwise $p<0.20$		<0.0001	0.28	0.66	0 (=log)	0 (=log), 3
Univariate	Post-chemotherapy size (mm)	<0.0001	0.0002	0.46	0 (=log)	0.5, 1
Full model		0.0004	0.004	0.45	0 (=log)	**0.5, 0.5**
Stepwise $p<0.05$		0.012	0.086	0.30	0 (=log)	−0.5, −0.5
Stepwise $p<0.20$		0.0005	0.034	0.46	**0 (=log)**	−0.5, −0.5
Univariate	Reduction in size (%)	<0.0001	0.64	0.63	0 (=log)	−1, 3
Full model		0.0005	0.06	0.16	−1	**−1, −0.5**
Stepwise $p<0.05$		0.0002	0.64	0.78	−1	−1, 3
Stepwise $p<0.20$		0.0009	0.08	0.27	**−1**	−1. −0.5

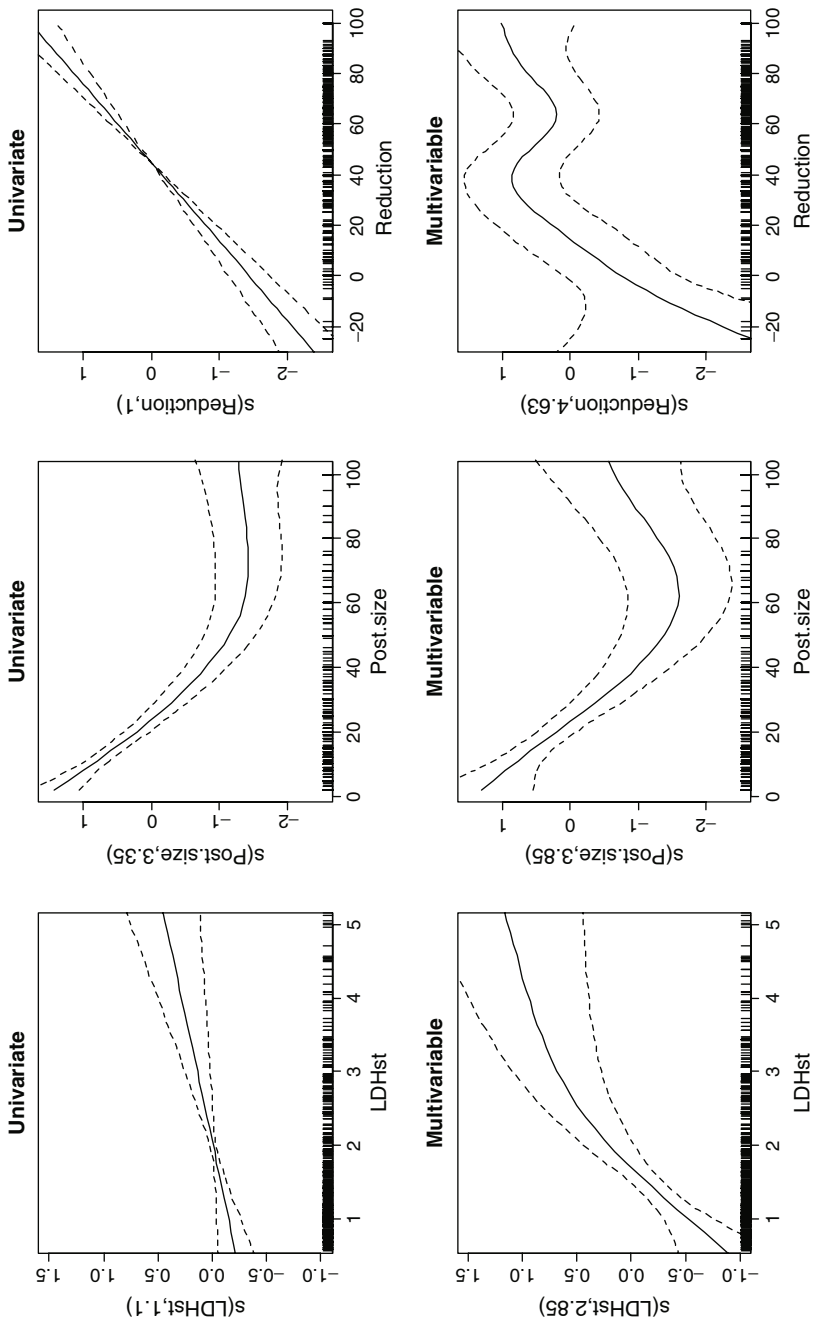

Fig. 12.5 Generalized additive models with optimal smoothing spline transformations according to a generalized cross-validation procedure in the testicular cancer example (n=544). *Top row*: optimal transformation in univariate logistic regression analysis; *bottom row*: multivariable logistic regression analysis with six predictors. The degrees of freedom of the optimal smoothing spline transformation are shown in each y-axis label. The distribution of values if shown at the bottom of the graphs

is modelled with a non-linear function using 3.35 effective *df*. In multivariable analyses, non-linear functions are used for all three continuous predictors, using 2.85, 3.85, and 4.63 effective *df* for LDHst, post-chemotherapy size and reduction, respectively (Fig. 12.5). Hence, more complex transformations were chosen in multivariable than in univariate analyses. The multivariable function for LDH looks much like a log transformation, as chosen previously. For post-chemotherapy size, we note an implausible increase in logodds of necrosis with very large mass sizes, and for reduction in size we note a wiggly shape between 20% and 100%. Hence, the smooth functions might not be smooth enough from a pathophysiological perspective. Further external validation might indicate whether the chosen "optimal" transformations are merely examples of overfitting.

*12.4.3 Predictive Performance

Finally, we study the predictive performance of the alternative non-linear transformations (Table 12.5). With linear terms only, we use 6 *df*, and achieve a model χ^2 of 205 (apparent R^2 41.9%, internally validated R^2 40.3%). We find the same model by applying a multivariable FP procedure with $p<0.05$ for selection; in fact we used more than 6 degrees of freedom in this approach, since we allowed for non-linear terms to be included in the model. If we fit a full FP2 model without selection, we use 15 *df*, and achieve a model χ^2 of 222. The increase by 17 (from 205 to 222) with 9 *df* is of borderline statistical significance (overall χ^2 test, $p=0.049$). If we apply a more liberal *p*-value for non-linearity, we use two FP1 transformations (1/reduction and log(postsize)) for a model χ^2 of 213. Using RCS functions with each 4 knots leads to a better fit than the FP2 functions (231 vs. 222). The increase in model LR (from 205 to 231, +26) is statistically significant (overall χ^2 test, 6 *df*, $p<0.001$). Our previous visual approximation of non-linearity

Table 12.5 Logistic Regression models with alternative codings of three continuous predictors

Strategy	Model	*df*	Model χ^2
Assume linearity (same as FP2, bw p<0.05 selection)	All linear	3+3	205
FP2, no selection	Full FP2	3+12	222
FP2, bw p<0.20 selection	LDHst, 1/reduction, log(postsize)	3+>3	213
RCS, no selection	3 RCS functions, each 4 knots	3+9	231
Visual approximation	log(LDHst) + reduction + sqrt(postsize)	3+>3	212
GAM, pre-specify *df*	3 smooth functions, each 3 *df*	3+9	232
GAM, GCV	3 optimally smoothed functions	3+(2.8+4.6+3.9)	240

in LDH and postsize led to a similar fit as the FP1 functions (model χ^2 212 vs. 213). Smoothing splines were similar in performance as the RCS model when 9 *df* were spent on the continuous predictors (model χ^2 232 vs. 231). With "optimal" transformations (GAM, GCV), more effective degrees of freedom were spent, and the highest model χ^2 or model LR was achieved (240). All model LRs indicate apparent performance. Rigorous internal validation, including all model selection steps, would be desired to indicate any true increase in performance, after correction for optimism. If inclusion of all modelling decisions is impossible, validation in a fully independent validation set may be required (split-sample, or external validation, see Chap. 17).

*12.4.4 R code for Non-Linear Modelling

```
# RCS: multivariable logistic regression with 3 rcs functions,
# each 4 knots
library(Hmisc)
library(Design)
lrm(NEC ~ Teratoma + Pre.AFP + Pre.HCG + rcs(LDHst, 4)+
   rcs(Post.size, 4) + rcs(Reduction, 4), data=n544)
# FP: multivariable fractional polynomial
library(mfp)
mfp(NEC ~ Teratoma + Pre.AFP+Pre.HCG + fp(LDHst) +
   fp(Post.size) + fp(Reduction), alpha=1, data=n544)
# GAM: multivariable gam, 3 effective df for each continuous predictor
library(gam)
gam(NEC ~ Teratoma+Pre.AFP + Pre.HCG + s(LDHst, 3) + s(Post.size, 3) +
        s(Reduction, 3), data=n544, family=binomial)
# multivariable gam, optimal effective df for each continuous predictor
# based on generalized cross-validation (GCV)
library(mgcv)
gam (NEC ~ Teratoma + Pre.AFP+Pre.HCG + s(LDHst) + s(Post.size) +
          s(Reduction), data=n544, family=binomial)
```

12.5 Concluding Remarks

On the one hand, one may see the additivity and linearity assumptions as essential aspects of a regression model. Hence one might argue that we should assess these assumptions thoroughly. When we are interested in the effect of a specific predictor, this may make sense. On the other hand, a thorough assessment of assumptions increases the risk of overfitting if we are primarily interested in obtaining predictions from a model. We will be tempted to adapt the model specification based on findings in the data, i.e. extend the model with interaction terms and/or non-linear terms. The price of striving for such perfection is that we may end up with a model that performs worse for future patients than a parsimonious model without interac-

tion terms or non-linear terms. Instead, we might strive for a "wrong, but useful" model.[51] Such a model should provide well-calibrated and discriminating predictions, despite possibly violating some underlying model assumptions.

In the examples in this chapter, model performance did not increase impressively. Of course, results may be different in other situations, but strong qualitative interaction or U-shaped non-linearity may be relatively rare. In general, it may be sobering to assess the increase in predictive performance by inclusion of interaction terms and non-linear terms; this is often quite modest in medical examples.

Note that prediction modelling techniques deal with interactions differently. A procedure such as Naïve Bayes estimation uses univariate effects of predictors in a multivariable prediction context; additivity is assumed and interactions are not studied. In contrast, tree models assume high-order interaction by default. Similarly, neural networks assume high-order interactions, allowing for their flexibility to fit specific patterns in a data set. To explore interactions we might hence also use a tree model, since it assumes interaction by default. Interactions that stand out could subsequently be considered in a regression model, and assessed for their significance. Shrinkage or penalized estimation may be particularly valuable to reduce interaction effects that were identified among a large set of potential interactions. Penalized ML is discussed in more detail in the next chapter.

12.5.1 Recommendations

Several measures can be taken to prevent the overfitting that may occur by considering additivity and linearity assumptions. First we should balance the number of interaction and non-linear terms to be considered with the effective sample size in the analysis (Table 12.6). We might only consider interactions in studies with relatively large sample sizes, i.e. many events compared to the number of terms considered. In smaller data sets, we may simply have to rely on the additivity assumption to be reasonable. We can also say that we estimate average (or "marginal") effects of predictors across subgroups; we know that we will never be able to exclude that we missed a relevant high-order interaction. For the linearity assumption, we might consider non-linear terms only for predictors with a presumed strong, and likely non-linear, effect. If previous studies have used a non-linear transformation for a predictor, we could also consider this transformation. Subject knowledge should also support the choice for a transformation; plotting the effect of a transformed predictor is essential (e.g. Figs. 12.1–12.5).

The second measure to prevent overfitting is to use overall tests, rather than focus on separate tests for interaction and non-linear terms. Note that based on an overall test, we would not have continued estimation of interaction of age and tachyacardia in the GUSTO-I subsample (Sect. 12.2). We should also note that interaction terms make life a bit more difficult for model presentation, arguing against their inclusion in a model unless their relevance is substantial for the specific prediction problem.

Table 12.6 Approaches to limit overfitting by assessing additivity and linearity assumptions

Approach	Description
Limited number of interaction/non-linear terms	Consider interaction term that are a priori plausible (Table 12.1); Consider non-linear terms only for predictors with a presumed strong, and likely non-linear, effect
Overall testing	Perform overall tests per interacting predictor (e.g. all age interactions)
Compare flexible vs. simple model	Compare the validated performance of a flexible model (e.g. including interactions and non-linearities) with a simple model without interaction and assuming linearity
Extra shrinkage of interaction/non-linear terms	Use a stronger shrinkage factor (<1) or more penalty in a penalized maximum likelihood procedure for interaction and non-linear terms

Third, an extension of this overall testing approach is to compare the performance of a flexible model to a simple model without interaction and non-linear effects (e.g. Table 12.5). The flexible model may for example be a neural network, or a GAM. Both the simple model and the flexible model should be validated, e.g. with bootstrapping, to see the validated rather than apparent improvement that might be achieved with inclusion of interaction and non-linear terms.

Finally, we may use shrinkage techniques to reduce the regression coefficients of selected interaction or non-linear terms. Some extra shrinkage may try to compensate for the "testimation bias" (see Chaps. 5 and 11), which is expected when terms were included in a model because they were relatively large.[174] The search for interactions and non-linear terms makes the effective degrees of freedom of a flexible model larger than the final degrees of freedom of a fitted model. This is recognized by FP transformations, where FP1 is tested with 2 *df*, and FP2 with 4 *df*. It is not included in *p*-values for optimal GAM transformations (according to GCV). *P*-values are then only approximate as a result of ignoring uncertainty in the model specification (e.g. searching for a smoothing parameter in GAM).

Questions

12.1 Additivity and interaction
- (a) Explain the additivity assumption in your own words, and the relevance of the scale for assessing additivity?
- (b) Explain interaction terms in your own words?
- (c) How many interaction terms can be assessed in a model with ten binary predictors?
- (d) How many of these would be expected to be statistically significant at the $p<0.05$ level?

12.2 Assumptions and model performance
- (a) Why would you consider testing of the additivity assumption with interaction terms?
- (b) What key problem can occur when interactions and non-linearities are included in the model? How can this be prevented?
- (c) Model performance increases with more flexible non-linear functions. In Table 12.5, the maximum Model LR is 240. Is this model hence prefered for predicting outcome, or do you think other considerations are also relevant?