Chapter 5
Linear Models

5.1 Introduction

In this chapter we consider linear regression models. These models have received considerable attention because of their mathematical and computational convenience and the relative ease of parameter interpretation. We discuss a number of issues that require consideration in order to perform a successful linear regression analysis. These issues are relevant irrespective of the inferential paradigm adopted and so apply to both frequentist and Bayesian analyses.

The structure of this chapter is as follows. We begin in Sect. 5.2 by describing a motivating example, before laying out the linear model specification in Sect. 5.3. A justification for linear modeling is provided in Sect. 5.4. In Sect. 5.5, we discuss parameter interpretation, and in Sects. 5.6 and 5.7, we describe, respectively, frequentist and Bayesian approaches to inference. In Sect. 5.8, the analysis of variance is briefly discussed. Section 5.9 provides a discussion of the bias-variance trade-off that is encountered when one considers which covariates to include in the mean model. In Sect. 5.10, we examine the robustness of the least squares estimator to model assumptions; this estimator can be motivated from estimating function, likelihood, and Bayesian perspectives. The assessment of assumptions is considered in Sect. 5.11. Section 5.12 returns to the motivating example. Concluding remarks are provided in Sect. 5.13 with references to additional material in Sect. 5.14.

5.2 Motivating Example: Prostate Cancer

Throughout this chapter we use the prostate cancer data of Sect. 1.3.1 to illustrate the main points. These data consist of nine measurements taken on 97 men. Along with the response, the log of prostate-specific antigen (PSA), there are eight covariates. As an illustrative inferential question, we consider estimation of the linear association between log(PSA) and the log of cancer volume, with possible

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_5,

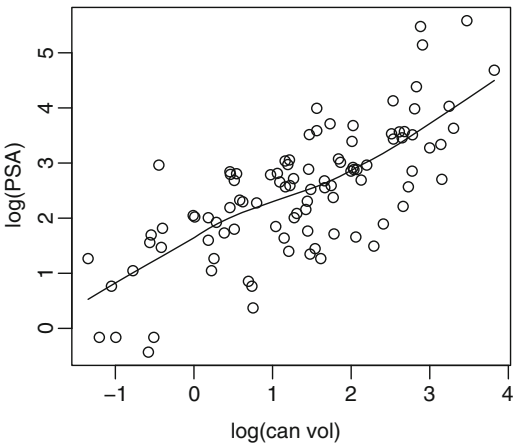

Fig. 5.1 Log of prostate-specific antigen versus log cancer volume, with smoother superimposed

adjustment for other "important" variables. Figure 5.1 plots log(PSA) versus log cancer volume, along with a smoother. The relationship looks linear, but Figs. 1.1 and 1.2 showed that log(PSA) was also associated with a number of the additional seven covariates and that there are strong associations between the eight covariates themselves. Consequently, we might question whether some or all of the other seven variables should be added to the model.

5.3 Model Specification

A multiple linear regression model takes the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \tag{5.1}$$

where we begin by assuming that the error terms are uncorrelated with $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$. In a *simple* linear regression model, $k = 1$ so that we have a single covariate. Linearity here is with respect to the parameters, and so variables may undergo nonlinear transforms from their original scale, before inclusion in (5.1).

In matrix form we write

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.2}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. We will also sometimes write

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{x}_i = [1\ x_{i1}\ \ldots\ x_{ik}]$ for $i = 1, \ldots, n$.

The covariates may be continuous or discrete. Discrete variables with a finite set of values are known as *factors*, with the values being referred to as *levels*. The levels may be ordered, and the ordering may or not be based upon numerical values. For example, dose levels of a drug are associated with numerical values but may be viewed as factor levels. Suppose x represents dose, with levels 0, 1, and 5. There are two alternative models that are immediately suggested for such an x variable. First, we may use a simple linear model in x:

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 x. \tag{5.3}$$

Second, we may adopt the model

$$\mathrm{E}[Y \mid x] = \alpha_0 \times I(x = 0) + \alpha_1 \times I(x = 1) + \alpha_2 \times I(x = 5), \tag{5.4}$$

where the indicator function

$$I(x = \widetilde{x}) = \begin{cases} 0 & \text{if } x \neq \widetilde{x} \\ 1 & \text{if } x = \widetilde{x} \end{cases}$$

and ensures that the appropriate level of x is picked. The mean function (5.4) allows for nonlinearity in the modeled association between Y and the observed x values, but does not allow interpolation to unobserved values of x. In contrast, (5.3) allows interpolation but imposes linearity. For an *ordinal* variable, the order of categories matters, but there are not specific values associated with each level (though values will be assigned as labels for computation). An example of an ordinal value is a pain score with categories none/mild/medium/severe. Alternatively, the levels may be nominal (such as female/male). The coding of factors is discussed in Sect. 5.5.2. Covariates may be of inherent were specific interest or may be included in the model in order to control for sources of variability or, more specifically, confounding; Sect. 5.9 provides more discussion.

The lower-/uppercase notation adopted here explicitly emphasizes that the covariates $\boldsymbol{x}$ are viewed as fixed while the responses $\boldsymbol{Y}$ are random variables. This is true regardless of whether the covariates were fixed by design or were random with respect to the sampling scheme. In the latter case it is assumed that the distribution of $\boldsymbol{x}$ does not carry information concerning $\boldsymbol{\beta}$ or σ^2, so that it is *ancillary* (Appendix F). Specifically, letting $\boldsymbol{\gamma}$ denote parameters associated with a model for $\boldsymbol{x}$, we assume that

$$p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{x} \mid \boldsymbol{\gamma}), \tag{5.5}$$

so that conditioning on $\boldsymbol{x}$ does not incur a loss in information with respect to $\boldsymbol{\beta}$. Hence, we can ignore the second term on the right-hand side of (5.5).

Random covariates, as just discussed, should be distinguished from inaccurately measured covariates. We will assume throughout that the $\boldsymbol{x}$ values are measured *without error*, an assumption that must always be critically assessed. In an observational setting in particular, it is common for elements of $\boldsymbol{x}$ to be measured with at least some error, but, informally speaking, we hope that these errors are small relative to the ranges; if this is not the case, then we must consider so-called *errors-in-variables* models; methods for addressing this problem are extensively discussed in Carroll et al. (2006).

5.4 A Justification for Linear Modeling

In this section we discuss the assumption of linearity. In general, there is no reason to expect the effects of continuous covariates to be causally linear,[1] but if we have a "true" model, $\mathrm{E}[Y \mid x] = f(x)$, then a first-order Taylor series expansion about a point x_0 gives

$$\begin{aligned} f(x) &\approx f(x_0) + \left.\frac{df}{dx}\right|_{x_0} (x - x_0) \\ &= \beta_0 + \beta_1(x - x_0) \end{aligned}$$

so that, at least for x values close to x_0, we have an approximately linear relationship.

As an example, Fig. 5.2 shows the height of 50 children plotted against their age. The true nonlinear form from which these data were generated is the so-called Jenss curve:

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 x - \exp(\beta_2 + \beta_3 x),$$

where Y is the height of the child at year x. This model was studied by Jenss and Bayley (1937), and the parameter values for the simulation were taken from Dwyer et al. (1983). The solid line on Fig. 5.2 is the curve from which these data were simulated, and the dotted and dashed lines are the least squares fits using data from ages less than 1.5 years only and greater than 4.5 years only, respectively. At younger ages, the association is approximately linear, and similarly for older ages, but a single linear curve does not provide a good description over the complete age range.

[1] In fact, as illustrated in Example 1.3.4, many physical phenomena are driven by differential equations with nonlinear models arising as solutions to these equations.

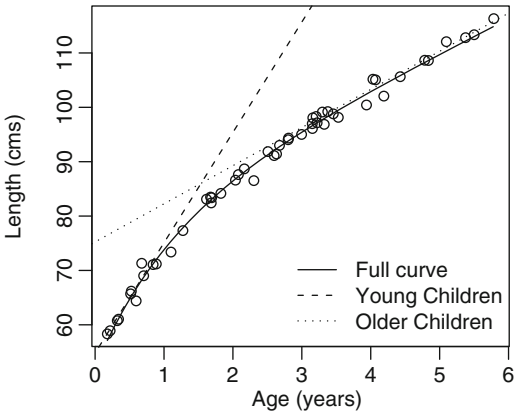

Fig. 5.2 Illustration of linear approximations to a nonlinear growth curve model

5.5 Parameter Interpretation

Before considering inference, we discuss parameter interpretation for the linear model. This topic is of vital importance in many settings, in order to report analyses in a meaningful manner. Interpretation is of far less concern in situations in which we simply wish to carry out prediction; methods for this endeavor are described in Chaps. 10–12. In a Bayesian analysis the specification of informative prior distributions requires a clear understanding of the meaning of parameters.

5.5.1 Causation Versus Association

We begin with the simple linear regression model

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 x. \tag{5.6}$$

Here we have explicitly conditioned upon x which is an important distinction since, for example, the models

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}(Y \mid x)] = \beta_0 \tag{5.7}$$

and

$$\mathrm{E}[Y \mid x] = \beta_0, \tag{5.8}$$

are very different. In (5.7) no assumptions are made, and we are simply saying that there is an average response in the population. However, (5.8) states that the expected response does not vary with x, which is a very strong assumption. Consequently, care should be taken to understand which situation is being considered.

We first consider the intercept parameter β_0 in (5.6), which is the expected response at $x = 0$. The latter expectation may make little sense (e.g., suppose the

response is blood pressure and the covariate is weight), and there are a number of reasons to instead use the model

$$\mathrm{E}[Y \mid x] = \beta_0^\star + \beta_1(x - x^\star), \tag{5.9}$$

within which $\beta_0^\star$ is the expected response at $x = x^\star$. By choosing $x^\star$ to be a meaningful value, we will, for example, be able to specify a prior for $\beta_0^\star$ more easily in a Bayesian analysis (see Sect. 3.4.2 for further discussion). Choosing $x^\star = \overline{x}$ is dataset specific (which does not allow simple comparison of estimates across studies) but provides a number of statistical advantages. Of course, models (5.6) and (5.9) provide identical inference since they are simply two parameterizations of the same model.

In both (5.6) and (5.9), the *mathematical* interpretation of the parameter β_1 is that it represents the additive change in the expected response for a unit increase in x. Notice that the interpretation of β_1 depends on the scales of measurement of both x and Y. More generally, $c\beta_1$ represents the additive change in the expected response for a c unit change in x. A difficulty with such interpretations is that it is inviting to think that if we were to provide an intervention and, for example, increase x by one unit for every individual in a population, then the expected response would change by β_1. The latter is a *causal* interpretation and is not appropriate in most situations, and never in observational studies, because unmeasured variables that are associated with both Y and x will be contributing to the observed association, $\widehat{\beta}_1$, between Y and x. In a designed experiment in which everything proceeds as planned, x is randomly assigned to each individual, and we may interpret β_1 as the expected change in the response for an individual following an intervention in which x were increased by one unit. Even in this ideal situation we need to know that the randomization was successfully implemented. It is also preferable to have large sample sizes so that any chance imbalance in variables between groups (as defined by different x values) is small.

We illustrate the problems with a simple idealized example. Suppose the "true" model is

$$\mathrm{E}[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 z, \tag{5.10}$$

and let

$$\mathrm{E}[Z \mid x] = a + bx, \tag{5.11}$$

describe the linear association between x and z. Then, if Y is regressed on x, only

$$\begin{aligned} \mathrm{E}[Y \mid x] &= \mathrm{E}_{Z \mid x}[\mathrm{E}(Y \mid x, Z)] \\ &= \beta_0 + \beta_1 x + \beta_2 \mathrm{E}[Z \mid x] \\ &= \beta_0^\star + \beta_1^\star x \end{aligned} \tag{5.12}$$

where

$$\begin{aligned} \beta_0^\star &= \beta_0 + a\beta_2 \\ \beta_1^\star &= \beta_1 + b\beta_2 \end{aligned} \tag{5.13}$$

showing that, when we observe x only and fit model (5.12), our estimate of $\beta_1^\star$ reflects not just the effect of x but, in addition, the effect of z mediated through its association with x. If $b = 0$, so that X and Z are uncorrelated, or if $\beta_2 = 0$, so that Z does not affect Y, then there will be no bias. Here "bias" refers to estimation of β_1, and not to $\beta_1^\star$. So for bias to occur in a linear model, Z must be related to both Y and X which, roughly speaking, is the definition of a *confounder*. The simulation at the end of Sect. 4.8 illustrated this phenomenon. A major problem in observational studies is that unmeasured confounders can always distort the true association. This argument reveals the beauty of randomization in which, by construction, there cannot be systematic differences between groups of units randomized to different x levels.

To rehearse this argument in a particular context, suppose Y represents the proportion of individuals with lung cancer in a population of individuals with smoking level (e.g., pack years) x. We know that alcohol consumption, z, is also associated with lung cancer, but it is unmeasured. In addition, X and Z are positively correlated. If we fit model (5.12), that is, regress Y on x only, then the resultant $\widehat{\beta}_1^\star$ is reflecting not only the effect of smoking but that of alcohol also through its association with smoking. Specifically, since $b > 0$ (individuals who smoke are more likely to have increased alcohol consumption), then (5.13) indicates that $\widehat{\beta}_1^\star$ will overestimate the true smoking effect β_1. If we were to intervene in our study population and (somehow) decrease smoking levels by one unit, then we would not expect the lung cancer incidence to decrease by $\beta_1^\star$ because alcohol consumption in the population has remained constant (assuming the imposed reduction does not change alcohol patterns). Rather, from (5.10), the expected decrease in the fraction with lung cancer will be β_1 if there were no other confounders (which of course is not the case). The interpretation of β_1 is the following. If we were to examine two groups of individuals within the study population with levels of smoking of $x+1$ and x, then we would expect lung cancer incidence to be $\widehat{\beta}_1^\star$ higher in the group with the higher level of smoking.

To summarize, great care must be taken with parameter interpretation in observational studies because we are estimating associations and not causal relationships. The parameter estimate associated with x reflects not only the "true" effect of x but also the effects of all other unmeasured variables that are related to both x and Y.

5.5.2 *Multiple Parameters*

In the model

$$\mathrm{E}[Y \mid x_1, \ldots, x_k] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

the parameter β_j is the additive change in the average response associated with a unit change in x_j, with all other variables held constant.

In some situations the parameters of a model may be very difficult to interpret. Consider the quadratic model:

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 x + \beta_2 x^2.$$

In this model, interpretation of β_1 (and β_2) is difficult because we cannot change x by one unit and simultaneously hold x^2 constant. An alternative parameterization that is easier to interpret is $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \gamma_2]$, where $\gamma_0 = \beta_0$, $\gamma_1 = -\beta_1/2\beta_2$, and $\gamma_2 = \beta_0 - \beta_1^2/4\beta_2$. Here γ_1 is the x value representing the turning point of the quadratic, and γ_2 is the expected value of the curve at this point.

We now discuss parameterizations that may be adopted when coding factors. We begin with a simple example in which we examine the association between a response Y and a two-level factor x_1, which we refer to as gender, and code as $x_1 = 0/1$, for female/male. The obvious formulation of the model is

$$\mathrm{E}[Y \mid x_1] = \begin{cases} \beta_0' + \beta_1' & \text{if } x_1 = 0 \text{ (female)}, \\ \beta_0' + \beta_2' & \text{if } x_1 = 1 \text{ (male)}. \end{cases}$$

The parameters in this model are clearly not *identifiable*; the data may be summarized as two means, but the model contains three parameters. This nonidentifiability is sometimes referred to as (intrinsic) *aliasing*, and the solution is to place a constraint on the parameters.

In the *sum-to-zero* parameterization, we impose the constraint $\beta_1' + \beta_2' = 0$, to give the model

$$\mathrm{E}[Y \mid x_1] = \begin{cases} \beta_0'' - \beta_1'' & \text{if } x_1 = 0 \text{ (female)}, \\ \beta_0'' + \beta_1'' & \text{if } x_1 = 1 \text{ (male)}. \end{cases}$$

In this case $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}''$, where the rows of the design matrix are $\boldsymbol{x} = [1, -1]$ if female and $\boldsymbol{x} = [1, 1]$ if male. We write

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[Y \mid x_1 = 0] \times p_0 + \mathrm{E}[Y \mid x_1 = 1] \times (1 - p_0) \\ &= \beta_0'' + \beta_1''(1 - 2p_0), \end{aligned}$$

where p_0 is the proportion of females in the population. We therefore see that β_0'' is the expected response if $p_0 = 1/2$, and

$$\mathrm{E}[Y \mid x_1 = 1] - \mathrm{E}[Y \mid x_1 = 0] = 2\beta_1'',$$

is the expected difference in responses between males and females.

An alternative parameterization imposes the *corner-point* constraint and assigns $\beta_1' = 0$ so that

$$\mathrm{E}[Y \mid x_1] = \begin{cases} \beta_0 & \text{if } x_1 = 0 \text{ (female)}, \\ \beta_0 + \beta_1 & \text{if } x_1 = 1 \text{ (male)}. \end{cases}$$

For this parameterization, $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$, where $\boldsymbol{x} = [1, 0]$ if female and $\boldsymbol{x} = [1, 1]$ if male. In this model, β_0 is the expected response for females, and β_1 is the additive change in the expected response for males, as compared to females.

A final model is

$$\mathrm{E}[Y \mid x_1] = \begin{cases} \beta_0^\dagger & \text{if } x_1 = 0 \text{ (female)} \\ \beta_1^\dagger & \text{if } x_1 = 1 \text{ (male).} \end{cases}$$

In this case $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}^\dagger$ where $x = [1, 0]$ if female and $x = [0, 1]$ if male so that $\beta_0^\star$ is the expected response for a female and $\beta_1^\star$ is the expected response for a male. We stress that inference for each of the formulations is identical; all that changes is parameter interpretation.

The benefits or otherwise of alternative parameterizations should be considered in the light of their extension to the case of more than two levels and to multiple factors. For example, the $[\beta_0^\dagger, \beta_1^\dagger]$ parameterization does not generalize well to a situation in which there are multiple factors and we do not wish to assume a unique mean for each combination of factors (i.e., a non-saturated model). It is obviously important to determine the default parameterization adopted in any particular statistical package so that parameter interpretation can be accurately carried out.

In this book we adopt the corner-point parameterization. Unlike the sum-to-zero constraint, this parameterization is not symmetric, since the first level of each factor is afforded special status, but parameter interpretation is relatively straightforward. If possible, one should define the factors so that the first level is the most natural "baseline." We illustrate the use of this parameterization with an example concerning two factors, x_1 and x_2, with x_1 having 3 levels, coded as 0, 1, 2, and x_2 having 4 levels coded as 0, 1, 2, 3. The coding for the no interaction (main effects[2] only) model is

$$\mathrm{E}[Y \mid x_1, x_2] = \begin{cases} \mu & \text{if } x_1 = 0, x_2 = 0, \\ \mu + \alpha_j & \text{if } x_1 = j, j = 1, 2, x_2 = 0, \\ \mu + \beta_k & \text{if } x_1 = 0, x_2 = k, k = 1, 2, 3, \\ \mu + \alpha_j + \beta_k & \text{if } x_1 = j, j = 1, 2, x_2 = k, k = 1, 2, 3. \end{cases}$$

As shorthand, we write this model as

$$\mathrm{E}[Y \mid x_1 = j, x_2 = k] = \mu + \alpha_j \times I(x_1 = j) + \beta_k \times I(x_2 = k),$$

for $j = 0, 1, 2$, $k = 0, 1, 2, 3$, with $\alpha_0 = \beta_0 = 0$.

[2]This terminology is potentially deceptive since "effects" invite a causal interpretation.

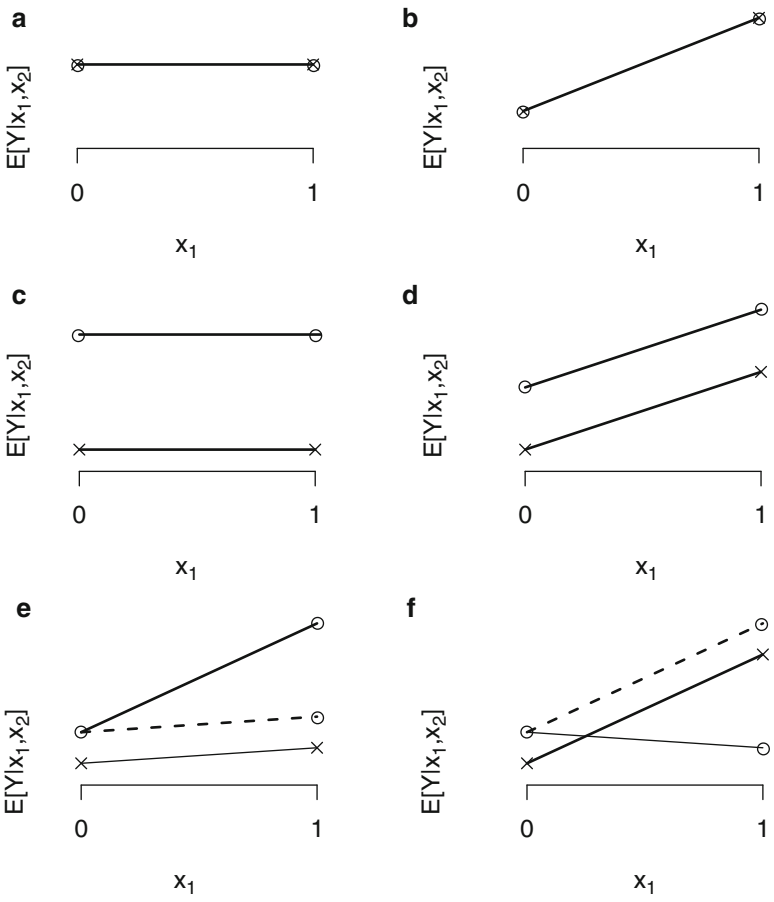

Fig. 5.3 Expected values for various models with two binary factors x_1 and x_2, "$\times$" represents $x_2 = 0$ and "$\circ$" $x_2 = 1$: (**a**) Null model, (**b**) x_1 main effect only, (**c**) x_2 main effect only, (**d**) x_1 and x_2 main effects, (**e**) interaction model 1, (**f**) interaction model 2. The *dashed lines* in panels (**e**) and (**f**) denote the expected response under the main effects only model

When one or more of the covariates are factors, interest may focus on interactions. To illustrate, suppose first we have two binary factors, x_1 and x_2 each coded as $0, 1$. The most general form for the mean is the saturated model

$$\mathrm{E}[Y \mid x_1, x_2] = \mu + \alpha_1 \times I(x_1 = 1) + \beta_1 \times I(x_2 = 1) + \gamma_{11} \times I(x_1 = 1, x_2 = 1) \tag{5.14}$$

where we have four unknown parameters and the responses may be summarized as four mean values. Figure 5.3 shows a variety of scenarios that may occur with this model. Panel (a) shows the null model in which the response does not depend

Table 5.1 Corner-point notation for two-factor model with interaction

		x_2			
		0	1	2	3
x_1	0	μ	$\mu+\beta_1$	$\mu+\beta_2$	$\mu+\beta_3$
	1	$\mu+\alpha_1$	$\mu+\alpha_1+\beta_1+\gamma_{11}$	$\mu+\alpha_1+\beta_2+\gamma_{12}$	$\mu+\alpha_1+\beta_3+\gamma_{13}$
	2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_1+\gamma_{21}$	$\mu+\alpha_2+\beta_2+\gamma_{22}$	$\mu+\alpha_2+\beta_3+\gamma_{23}$

on either variable, and panels (b) and (c) main effects due to x_1 only and to x_2 only, respectively. In panel (d) the response depends on both factors in a simple additive main effects only fashion (which is characterized by the parallel lines on the plot). The association with x_2 is the same for both levels of x_1 and $\gamma_{11} = 0$ in (5.14). Panels (e) and (f) show two different interaction scenarios. In panel (e), when $x_1 = 1$ and $x_2 = 1$ simultaneously, the expected response is greater than that predicted by the main effects only model (which is shown as a dashed line). In panel (f), the effect of the interaction is to reduce the association due to x_2. For the $x_1 = 0$ population, individuals with $x_2 = 1$ have an increased expected response over individuals with $x_2 = 0$. In the $x_1 = 1$ population, this association is reversed. In the saturated model (5.14), γ_{11} is measuring the difference between the average in the $x_1 = 1, x_2 = 1$ population and that predicted by the main effects only model. In the saturated model, α_1 is the expected change in the response between the $x_1 = 1$ and the $x_1 = 0$ populations when $x_2 = 0$, $\alpha_1 + \gamma_{11}$ is this same comparison when $x_2 = 1$.

In this example we have a *two-way* (also known as *first-order*) interaction (a terminology that extends in an obvious fashion to three or more factor). If an interaction exists in a model, then all main effects that are involved in the interaction will often be included in the model, which is known as the *hierarchy principle* (see Sect. 4.8 for further discussion). Following this principle aids in interpretation, but there are situations in which one would not restrict oneself to this subset of models. For example, in a prediction setting (Chaps. 12–10), we may ignore the hierarchy principle.

Table 5.1 illustrates the corner-point parameterization for the case in which there are two factors with three and four levels and all two-way interactions are present. The main effects model is obtained by setting $\gamma_{jk} = 0$ for $j = 1, 2, k = 1, 2, 3$. This notation extends to generalized linear models, as we see in Chap. 6.

5.5.3 Data Transformations

Model (5.1) assumes uncorrelated errors with constant variance. If there is evidence of nonconstant variance, the response may be transformed to achieve constant variance, though this changes other characteristics of the model. Historically, this was a popular approach due to the lack of easily implemented alternatives to the linear model with constant variance, and it is still useful in some instances.

For example, for positive data taking the log transform and fitting linear models is a common strategy. An alternative approach that is often preferable is to retain the mean–variance relationship and model on the original scale of the response (using a generalized linear model, for example see Chap. 6).

Suppose we have

$$\mathrm{E}[Y] = \mu_y$$

and

$$\mathrm{var}(Y) = \sigma_y^2 g(\mu_y),$$

so that the mean–variance relationship is determined by $g(\cdot)$, which is assumed known, at least approximately. Consider the transformed random variable, $Z = h(Y)$. Taking the approximation

$$Z \approx h(\mu_y) + (Y - \mu_y)h'(\mu_y),$$

where $h'(\mu_y) = \frac{dh}{dy}\big|_{\mu_y}$, produces

$$\mathrm{E}[Z] \approx h(\mu_y),$$

and

$$\mathrm{var}(Z) \approx \sigma_y^2 g(\mu_y) h'(\mu_y)^2.$$

To obtain independence between the variance and the mean, we therefore require

$$h(\cdot) = \int g(y)^{-1/2}\, dy. \tag{5.15}$$

For example, a commonly encountered relationship for positive responses is $\mathrm{var}(Y) = \sigma_y^2 \mu_y^2$, so that the coefficient of variation (which is the standard deviation divided by the mean) is constant. In this case, the suggested transformation, from (5.15), is $Z = \log Y$. As a second example, if $\mathrm{var}(Y) = \sigma_y^2 \mu_y$, the recommended transformation is $Z = \sqrt{Y}$.

Transformations of Y, and/or covariates, may also be taken in order to obtain an approximately linear association, though it is advisable to do this before seeing the scatterplot of y versus x, since data dredging is a bad idea, as discussed in Sect. 4.10.

Parameter interpretation is usually less straightforward if we have transformed the response and/or the covariates, as we illustrate with a series of examples. In this section, for clarity, we explicitly state the base of the logarithm. Suppose we fit the model

$$\log_e Y = \beta_0 + \beta_1 x + \epsilon, \tag{5.16}$$

or equivalently

$$Y = \exp(\beta_0 + \beta_1 x + \epsilon) = \exp(\beta_0 + \beta_1 x)\delta, \tag{5.17}$$

where $\delta = \exp(\epsilon)$. The expectation of Y depends on the distribution of ϵ, but the median of $Y \mid x$ is $\exp(\beta_0 + \beta_1 x)$, so long as the median of ϵ is zero. It will often be more appropriate to report associations in terms of the median for a positive response; $\exp(\beta_0)$ is the median response when $x = 0$, and $\exp(\beta_1)$ is the ratio of median responses corresponding to a unit increase in x. We may interpret the intercept in terms of the expected value for specific distributional choices for ϵ. For example, if $\epsilon \mid x \sim \mathrm{N}(0, \sigma^2)$, then since Y is lognormal (Appendix D),

$$\mathrm{E}[Y \mid x] = \exp(\beta_0 + \beta_1 x + \sigma^2/2),$$

giving $\mathrm{E}[Y \mid x = 0] = \exp(\beta_0 + \sigma^2/2)$ and

$$\frac{\mathrm{E}[Y \mid x + 1]}{\mathrm{E}[Y \mid x]} = \exp(\beta_1), \tag{5.18}$$

so that $\exp(\beta_1)$ can be interpreted as the *ratio* of expected responses between subpopulations whose x values differ by one unit. The interpretation (5.18) is true for other distributions, so long as $\mathrm{E}[\exp(\epsilon) \mid x]$ does not depend on x. In general, if (5.18) holds, $\exp(c\beta_1)$ is the ratio of expected responses between subpopulations with covariate values $x + c$ and x. An alternative interpretation follows from observing that

$$\frac{d}{dx}\mathrm{E}[Y \mid x] = \beta_1 \mathrm{E}[Y \mid x],$$

so that the rate of change of the mean function with respect to x is proportional to the mean, with proportionality constant β_1.

Model (5.16), with the assumption of normal errors, is useful if the standard deviation on the original scale is proportional to the mean (to give a constant coefficient of variation) since, evaluating the variance of a lognormal distribution (Appendix D),

$$\mathrm{var}(Y \mid x) = \mathrm{E}[Y \mid x]^2 \left[\exp(\sigma^2) - 1\right],$$

and if σ^2 is small, $\exp(\sigma^2) \approx 1 + \sigma^2$, and so

$$\mathrm{var}(Y \mid x) \approx \mathrm{E}[Y \mid x]^2 \sigma^2,$$

showing that for this model we have, approximately, a constant coefficient of variation. Hence, log transformation of the response is often useful for strictly positive responses, which ties in with the example following (5.15).

A model that looks similar to (5.17) is

$$Y = \mathrm{E}[Y \mid x] + \epsilon = \exp(\beta_0 + \beta_1 x) + \epsilon. \tag{5.19}$$

In this model we have additive errors, whereas in the previous case, the errors were multiplicative. For the additive model, $\exp(\beta_0)$ is the expected value at $x = 0$, and $\exp(\beta_1)$ is the ratio of expected responses between subpopulations whose x values

differ by one unit, regardless of the error distribution (so long as it has zero mean). In model (5.19), we may question whether additive errors are reasonable given that the mean function is always positive, though if the responses are well away from zero, there may not be a problem. Model (5.19) is nonlinear in the parameters, whereas (5.16) is linear, which has implications for inference and computation, as discussed in Chap. 6.

We now consider the model

$$Y = \beta_0 + \beta_1 \log_{10} x + \epsilon \tag{5.20}$$

which can be useful if linearity of the mean is reasonable on a log scale. For example, if we have dose levels of a drug x of 1, 10, 100, and 1,000, then we would be very surprised if changing x from 1 to 2 produces the same change in the expected response as increasing x from 1,000 to 1,001. Modeling on the original scale might also result in extreme x values that are overly influential, though the appropriateness of the description of the relationship between Y and x should drive the decision as to which scale to model on. For model (5.20), the obvious mathematical interpretation is that β_1 represents the difference in the expected response for individuals whose $\log_{10} x$ values differ by one unit. A more substantive interpretation follows from observing that

$$\mathrm{E}[Y \mid cx] - \mathrm{E}[Y \mid x] = \beta_1 \log_{10} c$$

so that for a $c = 10$-fold increase in x, the expected responses differ by β_1. Therefore, taking $\log_{10} x$ gives an associated coefficient that gives the same change in the average when going from 1 to 10, as when going from 100 to 1,000.

Similarly, if we consider a linear model in $\log_2 x$, then $k\beta_1$ is the additive difference between the expected response for two subpopulations with covariates $2^k x$ and x. For example, if one subpopulation has twice the covariate of another, the difference in the expected response is β_1. In general, if we reparameterize via $\log_a x$ (to give β_1 as the change corresponding to an a-fold change), then the effect of a b-fold change is $\beta_1 \log_a b$. As an example, if we initially assume the model

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 \log_e x,$$

then $\beta_1 \log_e 10 = 2.30 \times \beta_1$ is the expected change for a 10-fold change in x.

We now consider the model with both Y and x transformed

$$\log_e Y = \beta_0 + \beta_1 \log_{10} x + \epsilon.$$

Under this specification, $\exp(\beta_1)$ represents the multiplicative change in the median response corresponding to a 10-fold increase in x.

Example: Prostate Cancer

For the prostate data, a simple linear regression model that does not adjust for additional variables is

$$\log(\text{PSA}) = \beta_0 + \beta_1 \times \log_e(\text{can vol}) + \epsilon$$

where the errors ϵ are uncorrelated with $\text{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. In this model, $\exp(\beta_1)$ is the multiplicative change in median PSA associated with an e-fold change in cancer volume. Perhaps more usefully, $2.30 \times \beta_1$ is the multiplicative change in median PSA associated with a 10-fold increase in cancer volume.

5.6 Frequentist Inference

5.6.1 Likelihood

Consider the model

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with $\boldsymbol{\epsilon} \sim \text{N}_n(\boldsymbol{0}, \sigma^2\mathbf{I}_n)$, $\boldsymbol{x} = [1, x_1, \ldots, x_k]$, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\text{T}}$. The complete parameter vector is $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma]$ and is of dimension $p \times 1$ where $p = k + 2$. The likelihood function is

$$L(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right],$$

with log likelihood

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}),$$

which yields the score equations (estimating functions)

$$\boldsymbol{S}_1(\boldsymbol{\theta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2}\boldsymbol{x}^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}) \tag{5.21}$$

$$S_2(\boldsymbol{\theta}) = \frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}). \tag{5.22}$$

Setting (5.21) and (5.22) to zero (and assuming $\boldsymbol{x}^{\text{T}}\boldsymbol{x}$ is of full rank) gives the MLEs

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{Y}$$

$$\widehat{\sigma} = \left[\frac{1}{n}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})\right]^{1/2}.$$

We now examine the properties of these estimators, beginning with $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}\mathrm{E}[\widehat{\boldsymbol{\beta}}] &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{E}[\boldsymbol{Y}] \\ &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

so that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator for all n. Though S_2 is an unbiased estimating function, $\widehat{\sigma}$ is a nonlinear function of S_2 and so has finite sample bias (but is asymptotically unbiased).

Asymptotic variance estimators are obtained from the information matrix:

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\theta}}\right] = \begin{bmatrix}\boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & I_{22}\end{bmatrix}$$

where $\boldsymbol{S} = [\boldsymbol{S}_1, S_2]^{\mathrm{T}}$, and

$$\begin{aligned}\boldsymbol{I}_{11} &= \frac{\partial \boldsymbol{S}_1}{\partial \boldsymbol{\beta}} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{\sigma^2} \\ \boldsymbol{I}_{12} &= \boldsymbol{I}_{21}^{\mathrm{T}} = \frac{\partial \boldsymbol{S}_1}{\partial \sigma} = \boldsymbol{0} \\ I_{22} &= \frac{\partial S_2}{\partial \sigma} = \frac{2n}{\sigma^2}.\end{aligned}$$

Taking $\mathrm{var}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{I}(\boldsymbol{\theta})^{-1}$ gives

$$\begin{aligned}\mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ \mathrm{var}(\widehat{\sigma}) &= \frac{\sigma^2}{2n}.\end{aligned}$$

In practice, σ^2 is replaced by its estimator to give

$$\begin{aligned}\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) &= \widehat{\sigma}^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ \widehat{\mathrm{var}}(\widehat{\sigma}) &= \frac{\widehat{\sigma}^2}{2n}.\end{aligned}$$

For $\widehat{\boldsymbol{\beta}}$ to be unbiased, we need only assume $\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$, while for $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$, we require $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\mathbf{I}_n$, but *not* normality of errors. The expression for the variance is also exact for finite n.

The asymptotic distribution of the MLE based on n observations, $\widehat{\boldsymbol{\beta}}_n$, is

$$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\rightarrow_d\; \mathrm{N}_{k+1}(\boldsymbol{0}, \sigma^2\mathbf{I}_{k+1}), \tag{5.23}$$

and (by Slutsky's theorem, Appendix G) is still valid if σ is replaced by a consistent estimator. It should be stressed that normality of Y is not required, just n sufficiently large for the central limit theorem to apply. Since $\widehat{\boldsymbol{\beta}}_n$ is a linear combination of independent observations, the central limit theorem may be directly applied. Another way of viewing this asymptotic derivation is of replacing the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\beta})$ by $p(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{\beta})$.

For $\widehat{\sigma}$ to be asymptotically unbiased, we require $\text{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2 \mathbf{I}_n$, so that the estimating function for σ, (5.22), is unbiased. For $\widehat{\text{var}}(\widehat{\sigma}) = \widehat{\sigma}^2/2n$ to hold, we need the third and fourth moments to be correct and equal to zero and σ^2, respectively, as with the normal distribution. The dependence on higher-order moments results in inference for σ being intrinsically more hazardous than inference for $\boldsymbol{\beta}$.

Intervals for β_j, the jth components of $\boldsymbol{\beta}$, are based upon the statistic

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\text{s.e.}}(\widehat{\beta}_j)},$$

where the standard error in the denominator is $\widehat{\sigma}$ times the square root of the (j, j)th element of $(\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}$. The robustness to non-normality of the data is in part due to the standardization via the estimated standard error. In particular, we only require $\widehat{\sigma} \to_p \sigma$. An asymptotic $100 \times (1 - \alpha)\%$ confidence interval for β_j is

$$\widehat{\beta}_j \pm z_{\alpha/2} \times \widehat{\text{s.e.}}(\widehat{\beta}_j)$$

where $z_{\alpha/2} = \Phi(\alpha/2)$.

If we wish to make inference about σ^2, then we might be tempted to construct a confidence interval for σ^2 by leaning on $\epsilon_i \mid \sigma^2 \sim_{iid} \text{N}(0, \sigma^2)$. This leads to

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-k-1}, \tag{5.24}$$

where $\text{RSS} = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2$ is the residual sum of squares. Intervals obtained in this manner are extremely non-robust to departures from normality; however, see van der Vaart (1998, p. 27). The chi-square statistic does not standardize in any way, and any attempt to do so would require an estimate of the fourth moment of the error distribution, an endeavor that will be difficult due to the inherent variability in an estimate of the kurtosis (for a normal distribution, the kurtosis is zero, and so we do not require an estimate). Consequently, an interval (or test) based on (5.24) should not be used in practice unless we have strong evidence to suggest normality (or close to normality) of errors.

If the errors are such that $\boldsymbol{\epsilon} \mid \sigma^2 \sim \text{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then combining (5.23) with (5.24) gives, using (E.2) of Appendix E, the distribution

$$\widehat{\boldsymbol{\beta}} \sim \text{T}_{k+1}\left[\boldsymbol{\beta}, s^2 (\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}, n - k - 1\right], \tag{5.25}$$

a $(k+1)$-dimensional Student's t distribution with location $\boldsymbol{\beta}$, scale matrix $s^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$, and $n-k-1$ degrees of freedom (Sect. D). A $100\times(1-\alpha)\%$ confidence interval for β_j follows as

$$\widehat{\beta}_j \pm t_{\alpha/2}^{n-k-1} \times \widehat{\text{s.e.}}(\widehat{\beta}_j)$$

where $t_{\alpha/2}^{n-k-1}$ is the $\alpha/2$ percentage point of a standard t random variable with $n-k-1$ degrees of freedom. A more reliable approach to the construction of confidence intervals for elements of $\boldsymbol{\beta}$ is to use the bootstrap or sandwich estimation, though if n is small, the latter are likely to be unstable. For small n, a Bayesian approach may be taken, though there is no way that the distributional assumption made for the data (i.e., the likelihood) can be reliably assessed.

We have just discussed the non-robustness of (5.24) to normality. It is perhaps surprising then that confidence intervals constructed from (5.25) are used, since they are derived directly from (5.24). However, the resultant intervals are conservative in the sense that they are wider than those constructed from (5.23), explaining their widespread use.

For a test of $H_0 : \beta_j = c$, $j = 1, \ldots, k$, we may derive a t-test. Under H_0,

$$T = \frac{\widehat{\beta}_j - c}{S_j^{1/2}\widehat{\sigma}} \sim T_{n-k-1}, \tag{5.26}$$

where S_j is the (j,j)th element of $(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$ and T_{n-k-1} denotes the univariate t distribution with $n-k-1$ degrees of freedom, location $\widehat{\beta}_j$, and scale $S_j\widehat{\sigma}^2$. Although $\widehat{\sigma}$ can be very unstable, (5.26) it is an example of a self-normalized sum and so is asymptotically normal (Giné et al. 1997). The test with $c = 0$ is equivalent to the partial F statistic

$$F = \frac{\text{FSS}(\beta_j \mid \beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)/1}{\text{RSS}(\boldsymbol{\beta})/(n-k-1)},$$

where $\text{RSS}(\boldsymbol{\beta})$ is the residual sum of squares given the regression model $\text{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$ and the fitted sum of squares

$$\text{FSS}(\beta_j \mid \beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k) = \text{RSS}(\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k) - \text{RSS}(\boldsymbol{\beta}),$$

is equal to the change in residual sum of squares when β_j is dropped from the model. The "partial" here refers to the occurrence of β_l, $l \neq j$ in the model. Under H_0, $F \sim F_{1,n-k-1}$. The link with (5.26) is that $F = T^2$ with T evaluated at $c = 0$.

Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ be a partition with $\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q]$ and $\boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k]$, with $0 \leq q < k$. Interest may focus on simultaneously testing whether a set of parameters is equal to zero, via a test of the null

$$H_0 : \boldsymbol{\beta}_1 \text{ unrestricted, } \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \mathbf{0}].$$

Under H_0, the partial F statistic

$$F = \frac{\text{FSS}(\beta_{q+1}, \ldots, \beta_k \mid \beta_0, \beta_1, \ldots, \beta_q)/(k-q)}{\text{RSS}/(n-k-1)} = \frac{\text{FSS}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/(k-q)}{\text{RSS}/(n-k-1)} \tag{5.27}$$

is distributed as $F_{k-q,n-k-1}$ (Appendix D). Note that

$$\text{FSS}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) \neq \text{FSS}(\boldsymbol{\beta}_2),$$

unless $[x_1, \ldots, x_q]$ is orthogonal to $[x_{q+1}, \ldots, x_k]$. Such derivations are crucial to the mechanics of analysis of variance models, which we describe in Sect. 5.8.

Extending the above with $q = -1$ so that all $k+1$ parameters are being considered, the $100 \times (1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$ is the ellipsoid

$$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\text{T}} \boldsymbol{x}^{\text{T}} \boldsymbol{x} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq (k+1) s^2 F_{k+1,n-k-1}(1-\alpha) \tag{5.28}$$

where $s^2 = \text{RSS}/(n-k-1)$ and $F_{k+1,n-k-1}(1-\alpha)$ is the $1-\alpha$ point of the F distribution with $k+1, n-k-1$ degrees of freedom. The total sum of squares (TSS) may be partitioned as

$$\begin{aligned}
\text{TSS} &= (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\text{T}} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\text{T}} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\text{T}} \boldsymbol{x}^{\text{T}} \boldsymbol{x} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\
&= \text{RSS} + \text{FSS}.
\end{aligned}$$

Such expressions are specific to the linear model.

We now consider prediction of both an expected and an observed response. The latter require consideration of what we term *measurement error*, though we recognize that the errors in the model in general represent not only discrepancies arising from the measurement instrument but all manner of additional errors and sources of model misspecification. For inference concerning the *expected* response at covariate vector $\boldsymbol{x}_0$, we define $\theta = \boldsymbol{x}_0 \boldsymbol{\beta}$. Then $\widehat{\theta} = \boldsymbol{x}_0 \widehat{\boldsymbol{\beta}}$ and under correct first and second moment specification and via the central limit theorem:

$$[\boldsymbol{x}_0 (\boldsymbol{x}^{\text{T}} \boldsymbol{x})^{-1} \boldsymbol{x}_0^{\text{T}}]^{-1/2} (\widehat{\theta}_n - \theta) \;\rightarrow_d\; \text{N}(0, \sigma^2) \tag{5.29}$$

from which confidence intervals may be constructed. For prediction of an *observed* response at $\boldsymbol{x}_0$, we define $\phi = \boldsymbol{x}_0 \boldsymbol{\beta} + \epsilon$ with estimator $\widehat{\phi} = \boldsymbol{x}_0 \widehat{\boldsymbol{\beta}} + \widehat{\epsilon}$. It is now crucial to make a distributional assumption for the errors. Under $\epsilon \sim \text{N}(0, \sigma^2)$,

$$[1 + \boldsymbol{x}_0 (\boldsymbol{x}^{\text{T}} \boldsymbol{x})^{-1} \boldsymbol{x}_0^{\text{T}}]^{-1/2} (\widehat{\phi} - \phi) \sim \text{N}(0, \sigma^2). \tag{5.30}$$

The accuracy of intervals based on this form will be extremely sensitive to the normality assumption.

5.6.2 Least Squares Estimation

We describe an intuitive method of estimation with a long history and attractive properties. In *ordinary* least squares, the estimator is chosen to minimize the residual sum of squares

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}).$$

Differentiation (and scaling for convenience) gives

$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}}\text{RSS} = \boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}) \tag{5.31}$$

with solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{Y},$$

so long as $\boldsymbol{x}^{\text{T}}\boldsymbol{x}$ is of full rank. If we assume $\text{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$, then $\text{E}[\boldsymbol{G}(\boldsymbol{\beta})]=\boldsymbol{0}$ and so (5.31) corresponds to an estimating equation, and we may apply the nonidentically distributed version of Result 2.1, summarized in (2.13), with

$$\boldsymbol{A}_n = \text{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\beta}}\right] = -\boldsymbol{x}^{\text{T}}\boldsymbol{x}$$

$$\boldsymbol{B}_n = \text{var}(\boldsymbol{G}) = \boldsymbol{x}^{\text{T}}\text{var}(\boldsymbol{Y})\boldsymbol{x}.$$

Consequently, to obtain the variance–covariance matrix of $\widehat{\boldsymbol{\beta}}$, we need to specify $\text{var}(\boldsymbol{Y})$. Assuming $\text{var}(\boldsymbol{Y}) = \sigma^2\mathbf{I}_n$ gives $\boldsymbol{B} = \sigma^2\boldsymbol{x}^{\text{T}}\boldsymbol{x}$ and

$$(\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\rightarrow_d\; \text{N}_{k+1}(\boldsymbol{0}, \sigma^2\mathbf{I}_{k+1}).$$

More generally, sandwich estimation may be applied, as we discuss in Sect. 5.6.4.

In the method of *generalized least squares*, we assume $\text{E}[\boldsymbol{Y} \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$ and $\text{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a known matrix (*weighted least squares* corresponds to diagonal $\boldsymbol{V}$) and consider the function

$$\text{RSS}_G(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}).$$

Minimization of $\text{RSS}_G(\beta)$ yields the estimating function

$$\boldsymbol{G}_G(\boldsymbol{\beta}) = \boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

and corresponding estimator

$$\widehat{\boldsymbol{\beta}}_G = (\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{Y},$$

with asymptotic distribution

$$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_{k+1}(\boldsymbol{0}, \sigma^2\mathbf{I}_{k+1}). \tag{5.32}$$

This estimator also arises from a likelihood with $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{V})$ with $\boldsymbol{V} = \mathbf{I}_n$ giving the ordinary least squares estimator, as expected. An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2_G = \frac{1}{n-k-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}), \tag{5.33}$$

(see Exercise 5.1) and may be substituted for σ^2 in (5.32).

Given a particular dataset with n cases, a natural question is as follows: What is the practical significance of a central limit theorem and the associated regularity conditions? In the simple linear regression context, we require

$$\max_{1\le i\le n}(x_i - \overline{x})^2 / \sum_{j=1}^{n}(x_j - \overline{x})^2 \rightarrow 0, \tag{5.34}$$

as $n \rightarrow \infty$. Intuitively, the imaginary way in which the number of data points is going to infinity is such that no single x value can dominate. In Sect. 5.10 we will present a number of simulations showing the behavior of the least squares estimator as a function of n, the distribution of the errors, and the distribution of the x values. Such simulations give one an indication of when asymptotic normality "kicks in." The required conditions indicate the sorts of x distributions that are more or less desirable for valid asymptotic inference. A crucial observation is that reliable asymptotic inference via (5.32) requires the mean–variance relationship to be correctly specified. We now present a theorem that provides one justification for the use of the least squares estimator.

5.6.3 *The Gauss–Markov Theorem*

Definition. The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$:

- Is a linear function of $\boldsymbol{Y}$, so that the estimator can be written $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$, for an $n \times (k+1)$ matrix $\boldsymbol{B}$
- Is unbiased so that $\mathrm{E}[\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}] = \boldsymbol{\beta}$
- Has the smallest variance among all linear estimators

We now state and prove a celebrated theorem.

The Gauss–Markov Theorem: Consider the linear model $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{x}\boldsymbol{\beta}$, where $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{x}$ is $n \times (k+1)$, and $\boldsymbol{\beta}$ is $(k+1) \times 1$. Suppose further that $\mathrm{cov}(\boldsymbol{Y}) = \sigma^2\mathbf{I}_n$.

Then $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}$.

Proof. The estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ is clearly linear, and we have already shown it is unbiased. We therefore only need to show the variance is smallest among linear unbiased estimators.

Let $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$ be another linear unbiased estimator with $\boldsymbol{A}$ a $(k+1) \times n$ matrix. Since the estimator is unbiased, $\mathrm{E}[\widetilde{\boldsymbol{\beta}}] = \boldsymbol{A}\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{x}\boldsymbol{\beta}$ for any $\boldsymbol{\beta}$, which implies $\boldsymbol{A}\boldsymbol{x} = \mathbf{I}_{k+1}$. Now

$$\begin{aligned}\mathrm{var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{A}\sigma^2\mathbf{I}_{k+1}\boldsymbol{A}^{\mathrm{T}} - \sigma^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ &= \sigma^2\left[\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{A}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right].\end{aligned}$$

At this point we define $\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}$, which is known as the *hat* matrix (see Sect. 5.11.2). The hat matrix is symmetric and idempotent so that $\boldsymbol{h}^{\mathrm{T}} = \boldsymbol{h}$ and $\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}} = \boldsymbol{h}$. Further, $\mathbf{I}_n - \boldsymbol{h}$ inherits these properties. Using these facts, we can write

$$\begin{aligned}\mathrm{var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{A}(\mathbf{I}_n - \boldsymbol{h})\boldsymbol{A}^{\mathrm{T}} \\ &= \sigma^2\boldsymbol{A}(\mathbf{I}_n - \boldsymbol{h})(\mathbf{I}_n - \boldsymbol{h})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\end{aligned}$$

and this $(k+1) \times (k+1)$ matrix is positive definite, establishing that $\widehat{\boldsymbol{\beta}}$ has the smallest variance among linear unbiased estimators. □

This result shows that $\widehat{\boldsymbol{\beta}}$, which is the least squares estimate, the maximum likelihood estimate with a normal model, and the Bayesian posterior mean with normal model and improper prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ (as we show in Sect. 5.7), is optimal among linear estimators. We emphasize that, in the above theorem, only first and second moment assumptions were used with no distributional assumptions being required.

5.6.4 Sandwich Estimation

We have already examined the properties of the ordinary least squares/maximum likelihood estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{Y}$ and have seen that $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$, *if* $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\mathbf{I}_n$. Suppose that the correct variance model is $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{V}$ so that the model from which the estimator was derived was incorrect. Then the estimator is still unbiased, but the appropriate variance estimator is

$$\begin{aligned}\mathrm{var}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x})\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2, \qquad (5.35)\end{aligned}$$

Expression (5.35) can also be derived directly from the estimating function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

since we know

$$(\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{\mathrm{T}\,-1})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \;\rightarrow_d\; \mathrm{N}_{k+1}(\mathbf{0}_n, \mathbf{I}_n),$$

where

$$\boldsymbol{B}_n = \mathrm{var}(\boldsymbol{G}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}\sigma^2$$
$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\beta}}\right] = -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x},$$

to give

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2.$$

We now describe a sandwich estimator of the variance that relaxes the constant variance assumption but assumes uncorrelated responses. When the variance is not constant, the ordinary least squares estimator is consistent (since the mean specification is correct), but the usual standard errors will be inappropriate.

Consider the estimating function $\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})$. The "bread" of the sandwich $\boldsymbol{A}^{-1}$ remains unchanged since $\boldsymbol{A}$ does not depend on Y. The "filling" becomes

$$\boldsymbol{B} = \mathrm{var}(\boldsymbol{G}) = \boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y})\boldsymbol{x} = \sum_{i=1}^{n}\sigma_i^2\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i, \tag{5.36}$$

where $\sigma_i^2 = \mathrm{var}(Y_i)$ and we have assumed that the data are uncorrelated. Unfortunately, σ_i^2 is unknown, but various simple estimation techniques are available. An obvious estimator stems from setting $\widehat{\sigma}_i^2 = (Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2$ to give

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2, \tag{5.37}$$

and its use provides a consistent estimator of (5.36). However, this variance estimator has finite sample downward bias.

For linear regression, the MLE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}_i^2,$$

is downwardly biased (as we saw in Sect. 5.6.1), with bias $-(k+1)\sigma^2/n$, which suggests using

$$\widehat{\boldsymbol{B}}_n = \frac{n}{n-k-1}\sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2. \tag{5.38}$$

This simple correction provides an estimator of the variance that has finite bias, since the bias in $\widehat{\sigma}^2$ changes as a function of the design points $\boldsymbol{x}_i$, but will often improve

on (5.37). In linear regression, if $\text{var}(Y_i) = \sigma^2$, then $\text{E}[(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2] = \sigma^2(1 - h_{ii})$ where h_{ii} is the ith diagonal element of the hat matrix $\boldsymbol{x}(\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}$ (we derive this result in Sect. 5.11.2). Therefore, another suggested correction is

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\text{T}}\boldsymbol{x}_i \frac{(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2}{(1 - h_{ii})}. \tag{5.39}$$

For each of (5.37), (5.38), and (5.39), the variance of the estimator $\widehat{\boldsymbol{\beta}}$ is consistently estimated by $\widehat{\boldsymbol{A}}_n^{-1}\widehat{\boldsymbol{B}}_n\widehat{\boldsymbol{A}}_n^{-1}$.

We report the results of a small simulation study, in which we examine the performance of the sandwich estimator as a function of n, the distribution of x, and the variance estimator. We carry out six sets of simulations with the x distribution either uniform on (0,1) or exponential with rate parameter 1, and $\text{var}(Y \mid x) = \text{E}[Y \mid x]^q \times \sigma^2$ with $q = 0, 1, 2$, so that the variance of the errors is constant, increases in proportion to the mean, or increases in proportion to the square of the mean. The errors are normally distributed and uncorrelated in all cases (Sect. 5.10 considers the impact of other forms of model misspecification).

In Table 5.2, we see that, as expected, confidence intervals obtained directly from the usual variance of the ordinary least squares estimator, that is, $(\boldsymbol{x}^{\text{T}}\boldsymbol{x})^{-1}\widehat{\sigma}^2$, give accurate coverage when the error variance is constant. When the x distribution is uniform, the coverage is accurate even under variance model misspecification. There is poor coverage for the exponential distribution, however, which worsens with increasing n. The coverage of the sandwich estimator confidence intervals requires large samples to obtain accurate coverage for the exponential x model. There is a clear efficiency loss when using sandwich estimation, if the variance of the errors is constant. The downward bias of the sandwich estimator based on the unadjusted residuals is apparent, though this bias decreases with increasing n. Working with residuals standardized by $n/(n - k - 1)$, (5.38), improves the coverage, while the use of the hat matrix version, (5.39), improves performance further.

If the errors are correlated, the sandwich estimators of the variance considered here will not be consistent. Chapter 8 provides a description of sandwich estimators for the correlated data situation that may be used when there is replication across "clusters."

Example: Prostate Cancer

We fit the model

$$\log y_i = \beta_0 + \beta_1 \log_{10}(x_i) + \epsilon_i \tag{5.40}$$

where y_i is PSA and x_i is the cancer volume for individual i and ϵ_i are assumed uncorrelated with constant variance σ^2.

Table 5.2 Confidence interval coverage of nominal 95% intervals under a model-based variance estimator in which the variance is assumed independent of the mean and under three sandwich estimators given by (5.37)–(5.39)

n	Model-based	Sandwich 1	Sandwich 2	Sandwich 3
5	95	84	90	93
10	95	88	91	92
25	94	92	93	94
50	95	94	94	94
100	95	95	95	95
250	95	95	95	95
$\text{var}(Y \mid x) = \sigma^2$, x uniform				
5	95	82	88	92
10	95	85	88	91
25	95	89	91	92
50	95	91	92	93
100	95	93	93	94
250	95	94	94	94
$\text{var}(Y \mid x) = \sigma^2$, x exponential				
5	95	83	89	92
10	95	89	92	93
25	95	92	94	94
50	95	94	95	95
100	95	95	95	95
250	95	95	95	95
$\text{var}(Y \mid x) = \text{E}[Y \mid x] \times \sigma^2$, x uniform				
5	92	76	83	89
10	90	77	82	87
25	87	83	85	88
50	85	87	88	90
100	85	90	91	92
250	83	93	93	93
$\text{var}(Y \mid x) = \text{E}[Y \mid x] \times \sigma^2$, x exponential				
5	95	83	89	92
10	95	89	92	93
25	95	92	93	94
50	94	94	94	94
100	95	94	95	95
250	95	95	95	95
$\text{var}(Y \mid x) = \text{E}[Y \mid x]^2 \times \sigma^2$, x uniform				
5	89	70	78	86
10	81	71	75	82
25	75	78	80	85
50	73	85	86	88
100	71	89	90	91
250	68	92	92	93
$\text{var}(Y \mid x) = \text{E}[Y \mid x]^2 \times \sigma^2$, x exponential				

The true values are $\beta_0 = 1$, $\beta_1 = 1$, and all results are based on 10,000 simulations. In all cases, the errors are normally distributed and uncorrelated. The true variance model and distribution of x are given in the last line of each block

Table 5.3 Least squares/maximum likelihood parameter estimates and model-based and sandwich estimates of the standard errors, for the prostate cancer data

Parameter	Estimate	Model-based standard error	Sandwich standard error
β_0	1.51	0.122	0.123
β_1	0.719	0.0682	0.0728

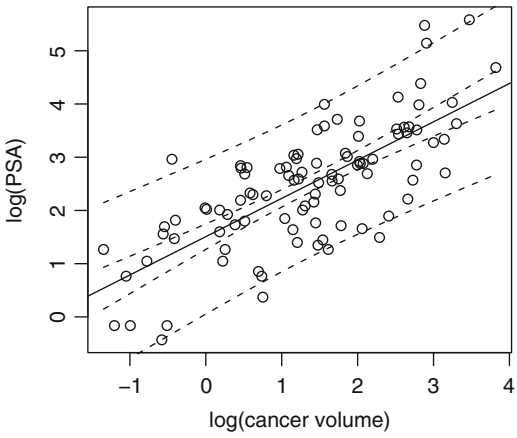

Fig. 5.4 Log of prostate-specific antigen versus log of cancer volume, along with the least squares/maximum likelihood fit, and 95% pointwise confidence intervals for the expected linear association (*narrow bands*) and for a new observation (*wide bands*)

where y_i is PSA and x_i is cancer volume for individual i and ϵ_i are assumed uncorrelated with constant variance σ^2. Table 5.3 gives summaries of the linear association under model-based and sandwich variance estimates. The point estimates and model-based standard error estimates arise from either ML estimation (assuming normality of errors) or ordinary least squares estimation of $\boldsymbol{\beta}$. The sandwich estimates of the standard errors relax the constancy of variance assumption but assume uncorrelated errors. The standard error of the intercept is essentially unchanged under sandwich estimation, when compared to the model-based version, while that for the slope is slightly increased. The sample size of $n = 97$ is large enough to guarantee asymptotic normality of the estimator. For a 10-fold increase in cancer volume (in cc), there is a $\exp(\widehat{\beta}_1) = 2.1$ increase in PSA concentration.

Figure 5.4 plots the log of PSA versus the log of cancer volume and superimposes the estimated linear association, along with pointwise 95% confidence intervals for the expected linear association and for a new observation (assuming normally distributed data). There does not appear to be any deviation in random scatter of the data around the line (a residual plot would give a clearer way of assessing the nonconstant variance assumption, as we will see in Sect. 5.10). In Fig. 5.5(a), we plot PSA versus log cancer volume and clearly see the variance of PSA increasing with increasing cancer volume on this scale. Figure 5.5(b) plots PSA versus cancer volume. It is very difficult to assess the goodness of fit of the fitted relationship

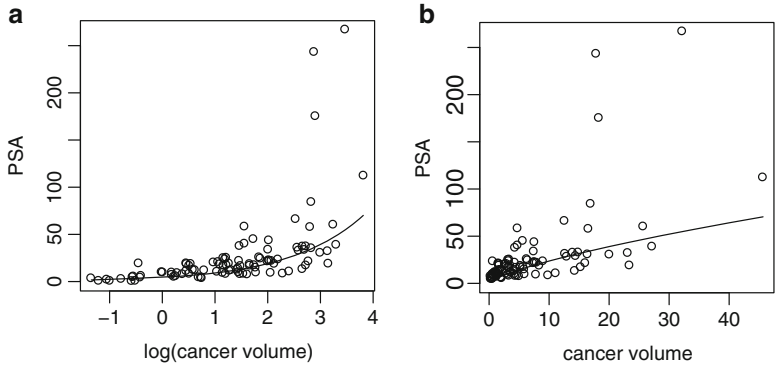

Fig. 5.5 (**a**) Prostate-specific antigen versus log cancer volume, (**b**) Prostate-specific antigen versus cancer volume. In each case, the least squares/maximum likelihood fit is included

or assumptions concerning the mean–variance relationship when the response and covariate are on their original scales. In both plots, the fitted line is from the fitting of model (5.40).

5.7 Bayesian Inference

We now consider Bayesian inference for the linear model. As with likelihood inference, we are required to specify the probability of the data and we assume $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$. The posterior distribution is

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto L(\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}, \sigma^2). \tag{5.41}$$

Closed-form posterior distributions for $\boldsymbol{\beta}$ and σ^2 are only available under restricted prior distributions. In particular, consider the improper prior distribution

$$\pi(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) \times p(\sigma^2) \propto \sigma^{-2} \tag{5.42}$$

Under this prior and likelihood combination, the posterior is, up to proportionality,

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right]. \tag{5.43}$$

To derive $p(\boldsymbol{\beta} \mid \boldsymbol{y})$, we need to integrate σ^2 from the joint distribution. To achieve this, it is useful to use an equality derived earlier, (2.23):

$$(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) = s^2(n - k - 1) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $\widehat{\boldsymbol{\beta}}$ is the ML/LS estimate. Substitution into (5.43) gives

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \int (\sigma^2)^{-(n+2)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[s^2(n-k-1) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]\right\} d\sigma^2.$$

The integrand here is the kernel of an inverse gamma distribution (Appendix D) for σ^2 and so has a known normalizing constant, the substitution of which gives

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \Gamma\left(\frac{n}{2}\right)\left[\frac{s^2(n-k-1) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{2}\right]^{-n/2}$$
$$\propto \left[1 + \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}s^{-2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{n-k-1}\right]^{-(n-k-1+k+1)/2}$$

after some simplification. By inspection we recognize that this expression is the kernel of a $(k+1)$-dimensional t distribution (Appendix D) with location $\widehat{\boldsymbol{\beta}}$, scale matrix $s^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$, and $n-k-1$ degrees of freedom, that is,

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim \mathrm{T}_{k+1}\left[(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}, (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}s^2, n-k-1\right]. \tag{5.44}$$

Consequently, under the prior (5.42), the Bayesian posterior mean $\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{y}]$ corresponds to the MLE, and $100(1-\alpha)\%$ credible intervals are identical to $100(1-\alpha)\%$ confidence intervals, though of course the two intervals have very different interpretations.

Asymptotically, as with likelihood estimation, it is the covariance model $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x})$ that is most important for valid inference, and normality of the error terms is unimportant. One way of thinking about this is as replacing $\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2$ by

$$\widehat{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \widehat{\sigma}^2 \sim \mathrm{N}_{k+1}\left[\boldsymbol{\beta}, \widehat{\sigma}^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1/2}\right].$$

We may derive the marginal posterior distribution of σ^2 as

$$\sigma^2 \mid \boldsymbol{y} \sim (n-p-1)s^2 \times \chi^{-2}_{n-k-1}, \tag{5.45}$$

a scaled inverse chi-squared distribution. As in the frequentist development, inference for σ^2 is likely to be highly sensitive to the normality assumption.

Although we can obtain analytic forms for $p(\boldsymbol{\beta} \mid \boldsymbol{y})$ and $p(\sigma^2 \mid \boldsymbol{y})$ under the prior (5.42), closed forms will not be available for general functions of interest. Direct sampling from the posterior may be utilized for inference in this case though. A sample from the *joint* distribution $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})$ can be generated using the composition method (Sect. 3.8.4) via the factorization

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = p(\sigma^2 \mid \boldsymbol{y}) \times p(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}),$$

where $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim \mathrm{N}_{k+1}\left[\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\right]$, and $\sigma^2 \mid \boldsymbol{y}$ is given by (5.45). Independent samples are generated via the pair of distributions

$$\begin{aligned} \sigma^{2(t)} &\sim p(\sigma^2 \mid \boldsymbol{y}) \\ \boldsymbol{\beta}^{(t)} &\sim p(\boldsymbol{\beta} \mid \sigma^{2(t)}, \boldsymbol{y}), \end{aligned}$$

for $t = 1, \ldots, T$. Samples for functions of interest $\phi = g(\boldsymbol{\beta}, \sigma^2)$ are then available as $\phi^{(t)} = g(\boldsymbol{\beta}^{(t)}, \sigma^{2(t)})$.

The conjugate prior (Sect. 3.7.1) here takes the form $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta} \mid \sigma^2)\pi(\sigma^2)$ with $\boldsymbol{\beta} \mid \sigma^2 \sim \mathrm{N}_{k+1}(\boldsymbol{m}, \sigma^2\boldsymbol{V})$ and $\sigma^{-2} \sim \mathrm{Ga}(a, b)$. However, this specification is not that useful in practice since the prior for $\boldsymbol{\beta}$ depends on σ^2. In particular, for smaller and smaller σ^2, the prior for $\boldsymbol{\beta}$ becomes increasingly concentrated about $\boldsymbol{m}$ which would not seem realistic in many contexts.

Under other prior distributions, analytic/numerical approximations or sampling-based techniques are required. An obvious prior choice is

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{V}), \quad \sigma^{-2} \sim \mathrm{Ga}(a, b)$$

which gives the posterior

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto l(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})\pi(\sigma^2)$$

which is intractable, unless $\boldsymbol{V}^{-1}$ is the $(k+1) \times (k+1)$ matrix of zeroes, which is the improper prior case, (5.42), already considered. Although the posterior is not available in closed form under this prior, it is straightforward to construct a blocked Gibbs sampling algorithm (Sect. 3.8.4). Specifically, letting $L(\boldsymbol{\beta}, \sigma^2)$ denote the likelihood, one iterates between the pair of conditional distributions:

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2) &\propto L(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}) \\ &\sim \mathrm{N}(\boldsymbol{m}^\star, \boldsymbol{V}^\star) \qquad (5.46) \\ p(\sigma^{-2} \mid \boldsymbol{y}, \boldsymbol{\beta}) &\propto L(\boldsymbol{\beta}, \sigma^2)\pi(\sigma^{-2}) \\ &\sim \mathrm{Ga}\left(a + \frac{n}{2}, b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right) \qquad (5.47) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{m}^\star &= \boldsymbol{W} \times \widehat{\boldsymbol{\beta}} + (\mathbf{I}_{k+1} - \boldsymbol{W}) \times \boldsymbol{m} \\ \boldsymbol{V}^\star &= \boldsymbol{W} \times \mathrm{var}(\widehat{\boldsymbol{\beta}}) \end{aligned}$$

and

$$\boldsymbol{W} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{V}^{-1}\sigma^2)^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}).$$

Conditional conjugacy is exploited in this derivation; for details, see Exercise 5.4. For general prior distributions, the Gibbs sampler is less convenient because the conditional distributions will be of unrecognizable form, but Metropolis–Hastings steps (Sect. 3.8.2) for $\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2$ and $\sigma^{-2} \mid \boldsymbol{y}, \boldsymbol{\beta}$ are straightforward to construct.

5.8 Analysis of Variance

The analysis of variance, or ANOVA, is a method by which the variability in the response is partitioned into components due to the various classifying variables and due to error. At one level, the ANOVA model is just a special case of a multiple linear regression model, but ANOVA does not simply have a role as an "outgrowth" of linear models. Rather Cox and Reid (2000, p. 245) state that ANOVA has a role "in clarifying the structure of sets of data, especially relatively complicated mixtures of crossed and nested data. This indicates what contrasts can be estimated and the relevant basis for estimating error. From this viewpoint the analysis of variance table comes first, then the linear model, not *vice-versa*." A study of the analysis of variance is intrinsically linked to the study of the design of experiments. Numerous books exist on ANOVA and the design of experiments; here we only give a brief discussion and introduce the main concepts. Specifically, we distinguish between *crossed* and *nested* (or hierarchical) designs and *fixed* and *random* effects modeling.

5.8.1 One-Way ANOVA

Consider the data in Table 5.4, taken from Davies (1967), which consist of the yield (in grams) from six randomly chosen batches of raw material, with five replicates each. The aim of this experiment was to find out to what extent batch-to-batch variation is responsible for variation in the final product yield.

Data such as these correspond to the simplest situation in which we have a single factor and a one-way classification. We may model the yield Y_{ij} in the jth sample from batch i as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{5.48}$$

Table 5.4 Yield of dyestuff in grams of standard color, in each of six batches

Replicate observation	Batch					
	1	2	3	4	5	6
1	1,545	1,540	1,595	1,445	1,595	1,520
2	1,440	1,555	1,550	1,440	1,630	1,455
3	1,440	1,490	1,605	1,595	1,515	1,450
4	1,520	1,560	1,510	1,465	1,635	1,480
5	1,580	1,495	1,560	1,545	1,625	1,445

with $\epsilon_{ij} \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, a$, $j = 1, \ldots, n$. We need a constraint to prevent aliasing (Sect. 5.5.2), with two possibilities being the sum-to-zero constraint, $\sum_{i=1}^{a} \alpha_i = 0$, and corner-point constraint: $\alpha_1 = 0$. Model (5.48) is an example of a multiple linear regression with mean

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$$

in which

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{2n} \\ \vdots \\ Y_{a1} \\ \vdots \\ Y_{an} \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & 0 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{bmatrix},$$

and where we adopt the corner-point constraint. Suppose we are interested in whether there are differences between the strengths from different looms. No differences correspond to the null hypothesis:

$$H_0 : \alpha_1 = \ldots = \alpha_a = 0. \tag{5.49}$$

Carrying out $a(a-1)/2$ t-tests leads to multiple testing problems (Sect. 4.5). Viewing this problem from a frequentist perspective and with $a = 6$ batches, we have 15 tests of pairs of batches, and with an individual type I error of 0.05, this gives an overall type I error of $1 - 0.95^{10} = 0.54$. As an alternative, we may test (5.49) using an F test (Sect. 5.6.1). Specifically, the F statistic is given by

$$F = \frac{\mathrm{FSS}(\boldsymbol{\alpha} \mid \mu)/(a-1)}{\mathrm{RSS}(\boldsymbol{\alpha})/a(n-1)} \tag{5.50}$$

where

$$\mathrm{FSS}(\boldsymbol{\alpha} \mid \mu) = \mathrm{RSS}(\mu) - \mathrm{RSS}(\mu, \boldsymbol{\alpha})$$

is the fitted sum of squares that results when $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_a]$ is added to the model containing μ only. In (5.50), the F statistic is the ratio of two so-called mean squares, which are average sum of squares, and under H_0, since the contributions in numerator and denominator are independent, $F \sim F_{a-1,a(n-1)}$. The ANOVA table associated with the test is given in Table 5.5. This table lays out the quantities that require calculation and shows the decomposition of the total sum of squares into

Table 5.5 ANOVA table for the one-way classification. The F statistic is for a test of $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$; DF is short for degrees of freedom and EMS for the expected mean square, which is E[SS/DF]

Source	Sum of squares	DF	EMS	F statistic
Between batches	$\text{SS}_1 = n\sum_{i=1}^{a}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$	$a-1$	$\sigma^2 + n\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$	$\frac{\text{SS}_1/(a-1)}{\text{SS}_2/a(n-1)}$
Error	$\text{SS}_2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$a(n-1)$	σ^2	
Total	$\text{SS}_\text{T} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$an-1$		

Table 5.6 One-way ANOVA table for the dyestuff data; DF is shorthand for degrees of freedom

Source	Sum of squares	DF	Mean square	F statistic
Between batches	56, 358	5	11,272	4.60 (0.0044)
Error	58,830	24	2,451	
Total	115,188	29		

The quantity in brackets in the final column is the p-value

that due to groups (batches in this example) and that due to error. The intuition behind the F test is that if there are no group effects, then the average sum of squares corresponding to the groups will, in expectation, equal the error variance. Consequently, we see in Table 5.5 that the expected mean square is simply σ^2 when $\alpha_1 = \ldots = \alpha_a = 0$. The success of the F test depends on the fact that we may decompose the overall sum of squares into the sum of the constituent parts corresponding to different components, and these follow independent χ^2 random variables.

Table 5.6 gives the numerical values for the dyestuff data of Table 5.4 and results in a very small p-value. As discussed in Sect. 4.2, the calibration of p-values is difficult, but for this relatively small sample size, a p-value of 0.0044 strongly suggests that the null is very unlikely to be true, and we would conclude that there are significant differences between batch means for these data. A Bayesian approach to testing may be based on Bayes factors. In this linear modeling context, there are close links between the Bayes factor and the F statistic (O'Hagan 1994, Sect. 9.34), though as usual the interpretations of the two quantities differ considerably. It is straightforward to extend the F test to the case of different sample sizes within looms, that is, to the case of general n_i, $i = 1, \ldots, a$.

If we are interested in the overall average yield, we would not want to ignore batch effects if present (even if they are not of explicit interest), because a model with no batch effects would not allow for the positive correlations that are induced between yields within the same batch. This issue is discussed in far greater detail in Chap. 8.

Table 5.7 Data on clotting times (in minutes) for eight subjects, each of whom receives four treatments

	Treatment				
Subject	1	2	3	4	Mean
1	8.4	9.4	9.8	12.2	9.95
2	12.8	15.2	12.9	14.4	13.82
3	9.6	9.1	11.2	9.8	9.92
4	9.8	8.8	9.9	12.0	10.12
5	8.4	8.2	8.5	8.5	8.40
6	8.6	9.9	9.8	10.9	9.80
7	8.9	9.0	9.2	10.4	9.38
8	7.9	8.1	8.2	10.0	8.55
Mean	9.30	9.71	9.94	11.02	9.99

5.8.2 *Crossed Designs*

We now consider two factors, which we label A and B, with a and b levels, respectively. If each level of A is *crossed* with each level of B, we have a *factorial* design. Suppose that there are n replicates within each of the ab cells. The interaction model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n$. This model contains $1 + a + b + ab$ parameters, while the data supply only ab sample means. Therefore, it is clear that constraints on the parameters are required. In the corner-point parameterization (Sect. 5.5.2), the $1 + a + b$ constraints are

$$\alpha_1 = \beta_1 = \gamma_{11} = \ldots = \gamma_{1b} = \gamma_{21} = \ldots \gamma_{a1} = 0.$$

Alternatively, we may adopt the sum-to-zero constraints:

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$$

Table 5.7 reproduces data from Armitage and Berry (1994) in which clotting times of plasma are analyzed. These data are from a crossed design in which each of $a = 8$ subjects received $b = 4$ treatments. The design is crossed since each patient receives each of the treatments. These data also provide an example of a *randomized block design* in which the aim is to provide a more homogeneous experimental setting within which to compare the treatments. Ignoring the blocking factor increases the unexplained variability and reduces efficiency. Section 8.3 provides further discussion.

Table 5.8 ANOVA table for the two-way crossed classification with one observation per cell; DF is short for degrees of freedom and EMS for the expected mean square

Source	Sum of squares	DF	EMS	F statistic
Factor A	$\text{SS}_\text{A} = b\sum_{i=1}^{a}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$a-1$	$\frac{\text{SS}_\text{A}}{a-1}$	$\frac{\sigma^2 + b\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Factor B	$\text{SS}_\text{B} = a\sum_{j=1}^{b}(\overline{Y}_{.j} - \overline{Y}_{..})^2$	$b-1$	$\frac{\text{SS}_\text{B}}{b-1}$	$\frac{\sigma^2 + a\sum_{j=1}^{b}\beta_j^2}{b-1}$
Error	$\text{SS}_\text{E} = \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$	$(a-1)(b-1)$	$\frac{\text{SS}_\text{E}}{(a-1)}$	σ^2
Total	$\text{SS}_\text{T} = \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{..})^2$	$ab-1$		

Table 5.9 ANOVA table for the plasma clotting time data in Table 5.7; DF is short for degrees of freedom. The quantity in brackets in the final column is the p-value

Source of variation	Sum of squares	DF	Mean square	F statistic
Treatment	13.0	3	4.34	6.62 (0.0026)
Subjects	79.0	7	11.3	17.2 (2.2×10^{-7})
Error	13.8	21	0.656	
Total	105.8	31		

There are no replicates within each of the 8×4 cells in Table 5.7, and so it is not possible to examine interactions between subjects and treatments. Consequently, we concentrate on the main effects only model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{5.51}$$

for $i = 1, \ldots, 4; j = 1, \ldots, 8$ and with $\epsilon_{ij} \mid \sigma^2 \sim_{iid} \text{N}(0, \sigma^2)$. Here we adopt the corner-point parameterization with $\alpha_1 = 0$ and $\beta_1 = 0$. Table 5.8 gives the generic ANOVA table for a two-way classification with no replicates, and Table 5.9 gives the numerical values for the plasma data. For these data, primary interest is in treatment effects (the α_i's), and Table 5.9 shows the steps to obtaining a p-value of 0.0026 for the null of $H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$ which, for this small sample size, points strongly towards the null being unlikely. In passing, we note that there are large between-subject differences for these data, so that the crossed design is very efficient.

We now examine treatment differences using estimation. Under the improper prior

$$p(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

interval estimates obtained from Bayesian, likelihood, and least squares analyses are identical. We take a Bayesian stance and report the posterior distribution for each of the treatment effects. We let $\boldsymbol{\theta} = [\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ where $\boldsymbol{\alpha} = [\alpha_2, \alpha_3, \alpha_4]$ and $\boldsymbol{\beta} = [\beta_2, \ldots, \beta_8]$. The joint posterior for $\boldsymbol{\theta}$ is multivariate Student's t, with $n - k - 1 = 32 - 11 = 21$ degrees of freedom, posterior mean $\widehat{\boldsymbol{\theta}}$ (the least squares estimate) and posterior scale matrix, $(\boldsymbol{x}^\text{T}\boldsymbol{x})^{-1}\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the usual

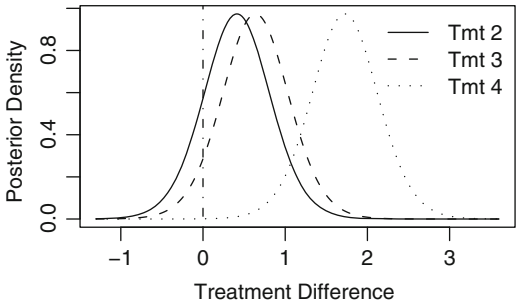

Fig. 5.6 Marginal posterior distributions for the treatment contrasts, with treatment 1 as the baseline, for the plasma clotting time data in Table 5.7

unbiased estimator of the residual error variance. Since treatment 1 is the reference we examine treatment differences with respect to this baseline group. Figure 5.6 gives the posterior distributions for $\alpha_2, \alpha_3, \alpha_4$. The posterior probabilities that the average responses under treatments 2, 3, and 4 are greater than zero are 0.16, 0.065, and 0.00017, respectively. Consequently, we conclude that there is strong evidence that treatment 4 differs from treatment 1, with decreasingly lesser evidence of differences between treatment 1 and treatments 3 and 2.

5.8.3 Nested Designs

For a design with two factors, suppose that Y_{ijk} denotes a response at level i of factor A and level j of factor B, with replication indexed by k. In a nested design, in contrast to a crossed design, $j = 1$ in level 1 of factor A has no meaningful connection with $j = 1$ in level 2 of factor A. In the context of the previous example, suppose each of eight patients received a single treatment each, but with k replicate measurements. In this case, we again have two factors, treatments and patients, but the patient effects are *nested* within treatments. A nested model for two factors is

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk},$$

with $i = 1, \ldots, a$ indexing factor A and $j = 1, \ldots, b$ factor B. In the nested patient/treatment example, A represents treatment and B patient, and so $\beta_{j(i)}$ represents the change in expected response for patient j within level i of treatment. Notice that there is no interaction in the model, because factor B is nested within factor A, and not crossed, and so there is no way of estimating the usual interactions. In a sense, $\beta_{j(i)}$ is an interaction parameter since it is the patient effect specific to a particular treatment. Table 5.10 gives the ANOVA table for this design.

Table 5.10 ANOVA table for the two-way nested classification; DF is short for degrees of freedom and EMS for the expected mean square

Source	Sum of squares	DF	EMS	F statistic
Factor A	$\text{SS}_A = bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$	$a-1$	$\frac{\text{SS}_A}{a-1}$	$\frac{\sigma^2+bn\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Factor B (within A)	$\text{SS}_B = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2$	$a(b-1)$	$\frac{\text{SS}_B}{(a(b-1)}$	$\frac{\sigma^2+a\sum_{j=1}^{b}\beta_j^2}{b-1}$
Error	$\text{SS}_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$ab(n-1)$	$\frac{\text{SS}_E}{(a-1)(b-1)}$	σ^2
Total	$\text{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n}(Y_{ijk} - \overline{Y}_{\cdots})^2$	$abn-1$		

5.8.4 *Random and Mixed Effects Models*

The examples we have presented so far are known, in the frequentist literature, as *fixed effects* ANOVA models since the parameters, for example, the α_i's in the one-way classification, are viewed as nonrandom. An alternative *random effects* approach is to view these parameters as a sample from a probability distribution, with the usual choice being $\alpha_i \mid \sigma_\alpha^2 \sim_{iid} \text{N}(0, \sigma_\alpha^2)$. From a frequentist perspective, the choice is based on whether the units that are selected can be viewed as being a random sample from some larger distribution of effects. Often, patients in a trial may be regarded as a random sample from some population, while treatment effects may be regarded as fixed effects. In this case, we have a *mixed effects* model. Model (5.51) was used for the data in Table 5.7 with the α_i and β_j being treated as fixed effects. Alternatively, we could use a mixed effects model with the individual effects α_i being treated as random effects and the β_j, representing treatment effects, being seen as fixed effects.

From a Bayesian perspective, the distinction being fixed and random effects is less distinct since all unknowns are viewed as random variables. However, the prior choice reflects the distinction. For example, in model (5.51), the "fixed effects" corresponding to treatments may be assigned independent prior distributions $\beta_j \sim \text{N}(0, V)$ where V is fixed, while the "random effects" corresponding to patients may be assigned the prior $\alpha_i \mid \sigma_\alpha^2 \sim_{iid} \text{N}(0, \sigma_\alpha^2)$ with σ_α^2 assigned a prior and estimated from the data.

A full description of estimation for random and mixed effects models will be postponed until Chap. 8, though here we briefly describe likelihood-based inference for the one-way model (5.48). Readers who have not previously encountered random effects models may wish to skip the remainder of this section and return after consulting Chap. 8. The one-way model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

Table 5.11 ANOVA table for test of $H_0 : \sigma_\alpha^2 = 0$; DF is short for degrees of freedom and EMS for the expected mean square

Source	Sum of squares	DF	EMS
Between batches	$n\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$a-1$	$\sigma^2 + n\sigma_\alpha^2$
Error	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$a(n-1)$	σ^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$an-1$	

where we have the usual assumption $\epsilon_{ij} \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $j = 1, \ldots, n$, and add $\alpha_i \mid \sigma_\alpha^2 \sim_{iid} \mathrm{N}(0, \sigma_\alpha^2)$, $i = 1, \ldots, a$ as the random effects distribution. We no longer need a constraint on the α_i's in the random effects model since these parameters are "tied together" via the normality assumption. A primary question of interest is often whether there are between-unit differences, and this can be examined via the hypothesis $H_0 : \sigma_\alpha^2 = 0$. In the one-way classification, this test turns out to be equivalent to the F test given previously in Sect. 5.8.1, though this equivalence is not true for more complex models. The ANOVA table given in Table 5.11 is very similar to that for the fixed effects model form in Table 5.5, though we highlight the difference in the final column.

Estimation via a likelihood approach proceeds by integrating the α_i from the model to give the marginal distribution

$$p(y_i \mid \mu, \sigma^2, \sigma_\alpha^2) = \int p(y_i \mid \mu, \alpha_i, \sigma^2) \times p(\alpha_i \mid \sigma_\alpha^2)\, d\alpha_i,$$

and results in

$$y_i \mid \mu, \sigma^2, \sigma_\alpha^2 \sim_{iid} \mathrm{N}(\mu \mathbf{1}_r, \sigma^2 \mathbf{I}_r + \sigma_\alpha^2 \mathbf{J}_r),$$

where $\mathbf{1}_r$ is the $r \times 1$ vector of 1's, $\mathbf{I}_r$ is the $r \times r$ identity matrix, and $\mathbf{J}_r$ is the $r \times r$ matrix of 1's. This likelihood can be maximized with respect to $\mu, \sigma_\alpha^2, \sigma^2$, and asymptotic standard errors may be calculated from the information matrix. A Bayesian approach combines the marginal likelihood with a prior $\pi(\mu, \sigma_\alpha^2, \sigma^2)$.

5.9 Bias-Variance Trade-Off

Chapter 4 gave an extended discussion of model formulation and model selection, and the example at the end of Sect. 4.8 acted as a prelude to this section in which we describe the bias-variance trade-off that is encountered when we consider which variables to include in a model.

Suppose the true model is

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{x}$ is $n \times (k+1)$, $\boldsymbol{\beta}$ is $(k+1) \times 1$, and the errors are such that $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$. We have seen that the estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y},$$

arises from ordinary least squares, likelihood (with normal errors, or large n), and Bayesian (with normal errors and prior (5.42), or large n) considerations. Asymptotically,

$$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{N}_{k+1}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where we assume $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ is of full rank. Since $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ is positive definite (all proper variance–covariance matrices are positive definite), we can find a unique Cholesky decomposition that is an upper-triangular matrix $\boldsymbol{U}$ such that $(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$. Proofs of the matrix results in this section may be found in Schott (1997, p.139–140). This decomposition leads to

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_j) = \sigma^2 \sum_{l=1}^{k+1} U_{jl}^2,$$

with $U_{jl} = 0$ if $j > l$.

We now split the collection of predictors into two groups, $\boldsymbol{x} = [\boldsymbol{x}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{B}}]$, and examine the implications of regressing on a subset of predictors. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_{\mathrm{A}}, \boldsymbol{\beta}_{\mathrm{B}}]^{\mathrm{T}}$ where $\boldsymbol{x}_{\mathrm{A}}$ is $n \times (q+1)$ with $q < k$ and $\boldsymbol{\beta}_{\mathrm{A}}$ is $(q+1) \times 1$. Now suppose we fit the model

$$\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{B}}] = \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}_{\mathrm{A}}^{\star}$$

where we distinguish between $\boldsymbol{\beta}_{\mathrm{A}}^{\star}$ and $\boldsymbol{\beta}_{\mathrm{A}}$ since the interpretation of the two sets of parameters differs. In particular, each coefficient in $\boldsymbol{\beta}_{\mathrm{A}}$ has an interpretation as the linear association of the corresponding variable, controlling for all of the other variables in $\boldsymbol{x}$. For coefficients in $\boldsymbol{\beta}_{\mathrm{A}}^{\star}$, control is only for variables in $\boldsymbol{x}_{\mathrm{A}}$. The estimator in the reduced model is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{A}}^{\star} = (\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{Y},$$

and

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\mathrm{A}}^{\star}] &= (\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\mathrm{E}[Y] \\
&= (\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta}_{\mathrm{B}}) \\
&= \boldsymbol{\beta}_{\mathrm{A}} + (\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta}_{\mathrm{B}}, \qquad (5.52)
\end{aligned}$$

so that the second term is the bias arising from omission of the last $k - q$ covariates. This defines the quantity that is being consistently estimated by $\widehat{\boldsymbol{\beta}}^{\star}_{\mathrm{A}}$. An alternative, less direct, derivation follows from the results of Sect. 2.4.3 in which we showed that the Kullback–Leibler distance between the true model and the reduced (assumed) model is that which is being minimized.

From (5.52), we see that the bias is zero if $\boldsymbol{x}_{\mathrm{A}}$ and $\boldsymbol{x}_{\mathrm{B}}$ are orthogonal, or if $\boldsymbol{\beta}_{\mathrm{B}} = 0$. Consequently, for bias to result, we need $\boldsymbol{x}_{\mathrm{B}}$ to be associated with both the response Y and at least one of the variables in $\boldsymbol{x}_{\mathrm{A}}$. These requirements, roughly speaking, are the conditions for $\boldsymbol{x}_{\mathrm{B}}$ to be considered a confounder. More precisely, Rothman and Greenland (1998) give the following criteria for a confounder:

1. A confounding variable must be associated with the response.
2. A confounding variable must be associated with the variable of interest in the population from which the data are sampled.
3. A confounding variable must not be affected by the variable of interest or the response. In particular it cannot be an intermediate step in the causal path between the variable of interest and the response.

At first sight, this result suggests that we should include as many variables as possible in the mean model, since this will reduce bias. But the splitting of the mean squared error of an estimator into the sum of the squared bias and the variance shows that this is only half of the story. Unfortunately, including variables that are not associated (or have a weak association only) with Y can increase the variance of the estimator (or equivalently, the posterior variance), as we now demonstrate.

We write

$$(\boldsymbol{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{A}})^{-1} = \boldsymbol{U}_{\mathrm{A}}\boldsymbol{U}_{\mathrm{A}}^{\mathrm{T}}$$

where $\boldsymbol{U}_{\mathrm{A}}$ is upper-triangular and consists of the first $q + 1$ rows and columns of $\boldsymbol{U}$. Denoting the jth element of the estimators from the reduced and full models as $\widehat{\beta}^{\star}_{\mathrm{A}j}$ and $\widehat{\beta}_{\mathrm{A}j}$, retrospectively, we have

$$\begin{aligned}\mathrm{var}(\widehat{\beta}^{\star}_{\mathrm{A}j}) &= \sigma^2 \sum_{l=1}^{q+1} U_{jl}^2 \\ &\leq \mathrm{var}(\widehat{\beta}_{\mathrm{A}j}),\end{aligned}$$

for $j = 0, 1, \ldots, q$, with equality if and only if $\boldsymbol{x}_{\mathrm{A}}$ and $\boldsymbol{x}_{\mathrm{B}}$ are orthogonal.

Hence, if σ^2 is fixed across analyses, we conclude that adding covariates decreases precision. Intuitively this is because there is only so much information within a dataset, and if we add in variables that are related to Y and are not orthogonal to existing variables, the associations are not so accurately estimated since there are now competing explanations for the data.

Another layer of complexity is added when we take into account estimation of σ^2 since the *estimated* standard errors of the estimator now depend on $\widehat{\sigma}^2$. The usual unbiased estimator is given by the residual sum of squares divided by the degrees of

freedom. The former is nonincreasing as covariates are added to the model, and the latter is decreasing. Consequently, as variables are entered into the model in terms of their "significance," a typical pattern is for $\widehat{\sigma}^2$ to decrease with the addition of important covariates, with an increase then occurring as variables that are almost unrelated are added (due to the decrease in the denominator of the estimator).

To expand on this further, consider the "true" model in which we assume for simplicity that β_{B} is univariate so that $\boldsymbol{x}_{\text{B}}$ is $n \times 1$:

$$\boldsymbol{y} = \boldsymbol{x}_{\text{A}}\boldsymbol{\beta}_{\text{A}} + \boldsymbol{x}_{\text{B}}\boldsymbol{\beta}_{\text{B}} + \boldsymbol{\epsilon}$$

where $\text{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. We now fit the model

$$Y = \boldsymbol{x}_{\text{A}}\boldsymbol{\beta}_{\text{A}}^{\star} + \epsilon^{\star},$$

so that $\boldsymbol{x}_{\text{B}}$ is omitted. Then, viewing $\boldsymbol{X}_{\text{B}}$ as random (since it is unobserved), we obtain

$$\text{var}(\boldsymbol{Y} \mid \boldsymbol{x}_{\text{A}}) = \sigma^2\mathbf{I}_n + \beta_{\text{B}}^2\text{var}(\boldsymbol{X}_{\text{B}} \mid \boldsymbol{x}_{\text{A}}),$$

showing the form of the increase in residual variance (unless $\beta_{\text{B}} = 0$) when variables related to the response are added to the model. If $\boldsymbol{x}_{\text{A}}$ and $\boldsymbol{x}_{\text{B}}$ are collinear, the variance of $\boldsymbol{X}_{\text{B}}$ does not depend on $\boldsymbol{x}_{\text{A}}$.

We expand on the development of this section, with a slight change of notation, via the "true" model

$$Y_i = \beta_0 + \beta_{\text{A}}(x_i - \overline{x}) + \beta_{\text{B}}(z_i - \overline{z}) + \epsilon_i,$$

and fitted model

$$Y_i = \beta_0^{\star} + \beta_{\text{A}}^{\star}(x_i - \overline{x}) + \epsilon_i.$$

Then, $\widehat{\beta}_0 = \widehat{\beta}_0^{\star} = \overline{Y}$ (since the covariates are centered in each model), and so each is an unbiased estimator of the intercept:

$$\text{E}[\widehat{\beta}_0] = \text{E}[\widehat{\beta}_0^{\star}] = \beta_0.$$

From (5.52),

$$\begin{aligned} \text{E}[\widehat{\beta}_{\text{A}}^{\star}] &= \beta_{\text{A}} + \beta_{\text{B}} \times \frac{S_{xz}}{S_{xx}} \\ &= \beta_{\text{A}} + \beta_{\text{B}} \times \rho_{xz}\left(\frac{S_{xz}}{S_{xx}}\right)^{1/2} \end{aligned} \tag{5.53}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{xz} = \sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z}) \quad S_{zz} = \sum_{i=1}^{n}(z_i - \overline{z})^2$$

and

$$\rho_{xz} = \frac{S_{xz}}{(S_{xx}S_{zz})^{1/2}}.$$

We have seen (5.53) before in a slightly different form, namely (5.11) in the context of confounding. In the full model we have

$$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & S_{zz}/D & -S_{xz}/D \\ 0 & -S_{xz}/D & S_{xx}/D \end{bmatrix},$$

where $D = S_{xx}S_{zz} - S_{xz}^2$, giving

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_{\mathrm{A}}) &= \frac{\sigma^2}{S_{xx} - S_{xz}^2/S_{zz}} \\ &\geq \frac{\sigma^2}{S_{xx}} = \mathrm{var}(\widehat{\beta}_{\mathrm{A}}^{\star}), \end{aligned}$$

with equality if and only if $S_{xz} = 0$ (so that X and Z are orthogonal), assuming that σ^2 is known.

When deciding upon the number of covariates for inclusion in the mean model, there are therefore competing factors to consider. The bias in the estimator cannot increase as more variables are added, but the precision of the estimator may increase or decrease, depending on the strength of the associations of the variables that are candidates for inclusion. The unexplained variation in the data (measured through $\widehat{\sigma}^2$) may be reduced, but the uncertainty in which of the covariates to assign the variation in the response to is increased. If the number of potential additional variables is large, the loss of precision may be considerable.

Section 4.8 described and critiqued various approaches to variable selection, emphasizing that the strategy taken is highly dependent on the context and in particular whether the aim is exploratory, confirmatory, or predictive. Chapter 12 considers the latter case in detail.

Example: Prostate Cancer

In this section we briefly illustrate the ideas of the previous section using two covariates from the PSA dataset, log(can vol) which we denote x_2 and log(cap pen) which we denote x_1. Let $\boldsymbol{x} = [x_1, x_2]$ and recall Y is log(PSA). Figure 5.7(a)

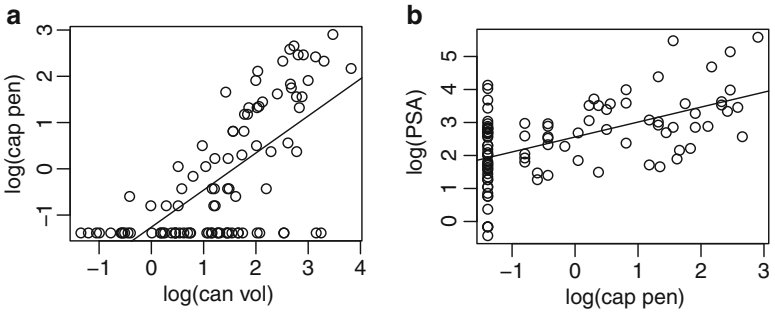

Fig. 5.7 (**a**) Association between log capsular penetration and log cancer volume, with fitted line, (**b**) association between log prostate-specific antigen and log capsular penetration, with fitted line

plots x_2 versus x_1 and illustrates the strong association between these variables. Figure 5.7(b) plots Y versus x_1, and we see an association here too. We obtain the following estimates:

$$\begin{aligned} \mathrm{E}[Y \mid \boldsymbol{x}] &= \beta_0^\star + \beta_1^\star x_1 && (5.54)\\ &= 1.51 + 0.72 \times x_1 && (5.55)\\ \mathrm{E}[Y \mid \boldsymbol{x}] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 && (5.56)\\ &= 1.61 + 0.66 \times x_1 + 0.080 \times x_2 && (5.57)\\ \mathrm{E}[x_2 \mid x_1] &= a + b x_1 \\ &= -12.6 + 0.80 \times x_1 && (5.58) \end{aligned}$$

We first confirm, using (5.12) and (5.11), that the estimate associated with log(can vol) in model (5.54) combines the effect of this variable and log(cap pen):

$$\begin{aligned} \widehat{\beta}_1^\star &= \widehat{\beta}_1 + \widehat{b} \times \widehat{\beta}_2 \\ &= 0.66 + 0.80 \times 0.08 = 0.72, \end{aligned}$$

with $\widehat{b}$ from (5.58), to give the estimate appearing in (5.55). The standard error associated with x_1 in model (5.54) is 0.068, while in the full model (5.56), it increases to 0.092 due to the association observed in Fig. 5.7a between x_1 and x_2.

5.10 Robustness to Assumptions

In this section we investigate the behavior of the estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^\mathrm{T}\boldsymbol{x})^{-1}\boldsymbol{x}^\mathrm{T}\boldsymbol{Y},$$

under departures from the assumptions that lead to

$$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\rightarrow_d\; \mathrm{N}_{k+1}(\mathbf{0}_{k+1}, \sigma^2\mathbf{I}_{k+1}).$$

Correct inference arises from normality of the estimator, and the error terms should have constant variance and absence of correlation. Normality of the estimator occurs with a sufficiently large sample size or if the error terms are normal. Judging when the sample size is large enough can be assessed through simulation, and there is an interplay between sample size and the closeness of the error distribution to normality. We present results examining the effect of departures on confidence interval coverage, but these are identical to Bayesian credible intervals under the improper prior (5.42). Regardless of the distribution of the errors and the mean–variance relationship, we always obtain an unbiased estimator, hence the emphasis on confidence interval coverage.

5.10.1 Distribution of Errors

We begin by examining the effect of non-normality of the errors and simulate data from a linear model with errors that are uncorrelated and with constant variance. The distribution of the errors is taken as either normal, Laplacian, Student's t with 3 degrees of freedom, or lognormal. We examine the behavior of the least squares estimator for β_1, with $n = 5$ and $n = 20$, and two distributions for the covariate, either $x_i \sim_{iid} \mathrm{U}(0,1)$ or $x_i \sim_{iid} \mathrm{Ga}(1,1)$ (an exponential distribution), for $i = 1, \ldots, n$. The latter was chosen to examine the effects of a skewed covariate distribution.

Table 5.12 presents the 95% confidence interval coverage for β_1; based on 10,000 simulations, the true value is $\beta_1 = 0$. For the normal error distributions, the coverage should be exactly 95%, but we include simulation-based results to give an indication of the Monte Carlo error. In all cases the coverage probabilities are good, showing the robustness of inference in this simple scenario. When the number of covariates, k is large relative to n, more care is required, especially if the distributions of the covariate are very skewed. Lumley et al. (2002) discuss the validity of the least squares estimator when the data are not normal.

5.10.2 Nonconstant Variance

We have already considered the robustness of inference to nonconstant error variance in Sect. 5.6.4, in the context of sandwich estimation. Table 5.2 showed that confidence interval coverage will be poor when an incorrect mean–variance relationship is assumed. Sandwich estimation provides a good frequentist alternative estimation strategy, so long as the sample size is large enough for the variance of

Table 5.12 Coverage of 95% confidence intervals for β_1 for various error distributions, distributions of the covariate, and sample sizes n. The entries are based on 10,000 simulations

Error distribution	Distribution of x	n	Coverage
Normal $\mathrm{N}(0,1)$	Uniform	5	95
Normal $\mathrm{N}(0,1)$	Uniform	20	94
Normal $\mathrm{N}(0,1)$	Exponential	5	95
Normal $\mathrm{N}(0,1)$	Exponential	20	95
Laplacian $\mathrm{Lap}(0,1)$	Uniform	5	95
Laplacian $\mathrm{Lap}(0,1)$	Uniform	20	95
Laplacian $\mathrm{Lap}(0,1)$	Exponential	5	94
Laplacian $\mathrm{Lap}(0,1)$	Exponential	20	95
Student $T(0,1,3)$	Uniform	5	95
Student $T(0,1,3)$	Uniform	20	95
Student $T(0,1,3)$	Exponential	5	95
Student $T(0,1,3)$	Exponential	20	95
Lognormal $\mathrm{LN}(0,1)$	Uniform	5	95
Lognormal $\mathrm{LN}(0,1)$	Uniform	20	96
Lognormal $\mathrm{LN}(0,1)$	Exponential	5	94
Lognormal $\mathrm{LN}(0,1)$	Exponential	20	95

the estimator to be reliably estimated. The bootstrap (Sect. 2.7) provides another method for reliable variance estimation, again when the sample size is not small.

5.10.3 Correlated Errors

Finally we investigate the effect on coverage of correlated error terms. A simple scenario to imagine is (x, y) pairs collected on consecutive days. We assume an AR(1) autoregression model of order 1 (Sect. 8.4.2) which results in $\boldsymbol{\epsilon} \mid \sigma^2 \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{V})$, where $\boldsymbol{V}$ is the $n \times n$ matrix

$$\boldsymbol{V} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho & \cdots & 1 \end{bmatrix}$$

and with ρ the correlation between errors on successive days. Table 5.13 gives the 95% confidence interval coverage (arising from a model in which the errors are assumed uncorrelated) as a function of sample size, the distribution of x (uniform or exponential), and strength of correlation. The table clearly shows that correlated errors can drastically impact confidence interval coverage, with the coverage becoming increasingly bad as the sample size increases.

Table 5.13 95% confidence interval for the slope parameter β_1 as a function of the autocorrelation parameter ρ and the sample size n. The entries are based upon 10,000 simulations and are calculated under a model in which the errors are assumed uncorrelated

Distribution of x	Correlation ρ	n	Coverage
Uniform	0.1	5	94
Uniform	0.1	20	93
Uniform	0.1	50	92
Uniform	0.5	5	89
Uniform	0.5	20	76
Uniform	0.5	50	75
Uniform	0.95	5	79
Uniform	0.95	20	36
Uniform	0.95	50	26
Exponential	0.1	5	94
Exponential	0.1	20	93
Exponential	0.1	50	93
Exponential	0.5	5	89
Exponential	0.5	20	79
Exponential	0.5	50	77
Exponential	0.95	5	81
Exponential	0.95	20	41
Exponential	0.95	50	32

Intuitively, one might expect that in this situation the standard errors based on $(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$ would always underestimate the true standard error of the estimator. In the scenario described above, the effect of correlated errors depends critically upon the correlation among the x variables across time, however. If the x-variable is slowly varying over time, then the standard errors will be underestimated, but if the variable is changing rapidly, then the true standard errors may be smaller than those reported. This is because if there is high positive correlation, then the difference in the error terms on consecutive days is small, and so if Y changes, it must be due to changes in x. For further discussion, see Sect. 8.3.

5.11 Assessment of Assumptions

In this section we will describe a number of approaches for assessing the assumptions required for valid inference.

5.11.1 Review of Assumptions

We consider the linear regression model:

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{x}$ is $n \times (k+1)$, $\boldsymbol{\beta}$ is $(k+1) \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$, with $\boldsymbol{\epsilon} \mid \sigma^2 \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Under these assumptions, we have seen that the estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$, with $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$, emerges from likelihood, least squares, and Bayesian approaches. The standard errors and confidence intervals we report are valid if:

- The error terms have constant variance. If sandwich estimation is used, then this assumption may be relaxed, so long as we have a large sample size.
- The error terms are uncorrelated.
- The estimator is normally distributed, so that we can effectively replace the likelihood $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2$ by $\widehat{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \sim \mathrm{N}_p\left[\boldsymbol{\beta}, (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\widehat{\sigma}^2\right]$. This occurs if the error terms are normally distributed and/or the sample size n is sufficiently large for the central limit theorem to ensure that the estimator is normally distributed.

As we saw in Sect. 5.10, confidence interval coverage can be very poor if the error variance is nonconstant and/or the errors are correlated. Normality of errors is not a big issue with the linear model with respect to estimation (which explains the popularity of least squares), unless the sample size is very small (relative to the number of predictors) or the distribution of the x values is very skewed. For validity of a predictive interval for an observable, we need to make a further assumption concerning the distribution of the error terms, however. This interval is given by (5.30) under the assumption of normal errors.

From a frequentist perspective and given the assumed mean model, $\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$, the estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. For example, in simple linear regression, $\widehat{\beta}_1$ is an unbiased estimator of the linear association in a population, regardless of the true relationship between response and covariate. The assumed mean model may be a poor description, however, and we will usually wish to examine the appropriateness of the model to decide on whether linearity holds.

Another aspect of model checking is scrutinizing the data for *outlying* or *influential* points. It is difficult to define exactly what is meant by an outlier, and we content ourselves with a fuzzy description of an outlier as "a data point that is unusual relative to the others." Single outlying observations may stand out in the plots described below. The presence of multiple outliers is more troublesome due to *masking*, in which the presence of an outlier is hidden by other outliers.

5.11.2 Residuals and Influence

In general, model checking may be carried out *locally*, using informal techniques such as residual plots, or *globally* using formal testing procedures; we concentrate on the former. The *observed error* is given by

$$e_i = Y_i - \widehat{Y}_i, \tag{5.59}$$

where $\widehat{Y}_i = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$, while the *true error* is

$$\epsilon_i = Y_i - \mathrm{E}[Y_i \mid \boldsymbol{x}_i].$$

In *residual analysis* we examine the observed residuals for discrepancies from the assumed model. We define *residuals* as

$$\boldsymbol{e} = [e_1, \ldots, e_n]^{\mathrm{T}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\mathbf{I}_n - \boldsymbol{h})\boldsymbol{Y}, \tag{5.60}$$

where $\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}$ is the *hat* (or projection) matrix encountered in Sect. 5.6.3. The hat matrix is symmetric, $\boldsymbol{h}^{\mathrm{T}} = \boldsymbol{h}$, and idempotent, $\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}} = \boldsymbol{h}$. We want to examine the relationship between $\boldsymbol{e}$ and $\boldsymbol{\epsilon}$ so we can use the former to assess whether assumptions concerning the latter hold.

Substitution of

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

into (5.60) gives

$$\boldsymbol{e} = (\mathbf{I}_n - \boldsymbol{h})\boldsymbol{\epsilon}, \tag{5.61}$$

or

$$e_i = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j, \tag{5.62}$$

showing that the estimated residuals differ from the true residuals, complicating residual analysis.

We examine the moments of the error terms. The residuals $\boldsymbol{e}$ are random variables since they are a function of the random variables $\boldsymbol{\epsilon}$. We have

$$\mathrm{E}[\boldsymbol{e}] = (\mathbf{I}_n - \boldsymbol{h})\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}_n$$

and the variance–covariance matrix is

$$\mathrm{var}(\boldsymbol{e}) = (\mathbf{I}_n - \boldsymbol{h})(\mathbf{I}_n - \boldsymbol{h})^{\mathrm{T}}\sigma^2 = (\mathbf{I}_n - \boldsymbol{h})\sigma^2,$$

so that fitting the model has induced dependence in the residuals. In particular,

$$\mathrm{var}(e_i) = (1 - h_{ii})\sigma^2,$$

since for a symmetric and idempotent matrix $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$ (see Schott 1997, p. 374), and

$$\mathrm{cov}(e_i, e_j) = -h_{ij},$$

showing that the observed errors have correlation given by

$$\text{corr}(e_i, e_j) = -\frac{h_{ij}}{[(1-h_{ii})(1-h_{jj})]^{1/2}}.$$

Consequently, even if the model is correctly specified, the residuals have nonconstant variance and are correlated. We may write

$$\widehat{Y}_i = h_{ii}Y_i + \sum_{j=1, j\neq i}^{n} h_{ij}Y_j, \tag{5.63}$$

so that if h_{ii} is large relative to the other elements in the ith row of $\boldsymbol{h}$, then the ith fitted value will be largely influenced by Y_i; h_{ii} is known as the *leverage*. Note that the leverage depends on the design matrix (i.e., the $\boldsymbol{x}$'s) only. Exercise 5.8 shows that $\text{tr}(\boldsymbol{h}) = k+1$ so the average leverage is at least $(k+1)/n$. If $h_{ii} = 1$, $\widehat{y}_i = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$ and the ith observation is fitted exactly, using a single degree of freedom for this point alone, which is not desirable.

Based on these results we may define *standardized residuals*:

$$e_i^\star = \frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}(1-h_{ii})^{1/2}}, \tag{5.64}$$

for $i = 1, \ldots, n$, and where $\widehat{\sigma}$ is an unbiased estimator of σ. These residuals have mean $\text{E}[\widehat{\sigma}e_i^*] = 0$ and variance $\text{var}(\widehat{\sigma}e_i^*) = \sigma^2$, but they are not independent since they are based on $n-k-1$ independent quantities. Often the $(1-h_{ii})^{1/2}$ terms in the denominator of (5.64) are ignored.

For the simple linear regression model,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

and

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}.$$

Therefore, with respect to (5.63), we see that an extreme x_i value produces a fitted value $\widehat{Y}_i$ that is more heavily influenced by the observed value of Y_i. Such x_i values also influence other fitted values, particularly those with x values not close to $\overline{x}$. The two constraints on the model are

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - \widehat{Y}_i = 0$$

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)x_i = 0$$

which induces correlation in the e_i's.

5.11.3 Using the Residuals

The constancy of variance assumption may be assessed by plotting the residuals, e_i versus the fitted values $\widehat{Y}_i$ with a random scatter suggesting no cause for concern. Examination may be simpler if squared residuals e_i^2 or absolute values of the residuals $|e_i|$ are plotted versus the fitted values $\widehat{Y}_i$. These plots are useful since departures from constant variance often correspond to a mean–variance relationship which, given sufficient data and range of the mean function, will hopefully reveal itself in these plots. If the variance increases with the mean, plotting e_i versus $\widehat{Y}_i$ will reveal a funnel shape with the wider end of the funnel to the right of the plot. For the plots using the squared or absolute residuals, interpretation may be improved with the addition of a smoother.

When one of the columns of $\boldsymbol{x}$ represents time, we may plot the residuals versus time and assess dependence between error terms. Dependence may also be detected using scatterplots of lagged residuals, for example, by plotting e_i versus e_{i-1} for $i = 2, \ldots, n$. Independent residuals should produce a plot with a random scatter of points. The autocorrelation at different lags may also be estimated for equally spaced data in time, while for unequally spaced data, a semi-variogram may be constructed. The latter is described in the context of longitudinal data in Sect. 8.8.

To assess normality of the residuals, we may construct a normal QQ plot. We first order the residuals and call these $e_{(i)}$, $i = 1, \ldots, n$. The expected order statistic of size n from a normal distribution is given (approximately) by

$$f_{(i)} = \Phi^{-1}\left(\frac{i - 0.5}{n}\right), i = 1, \ldots, n,$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, that is, if $Z \sim \mathrm{N}(0, 1)$ then $\Phi(z) = \Pr(Z < z)$. We then plot $e_{(i)}$ versus $f_{(i)}$. If the normality assumption is reasonable, the points should lie approximately on a straight line. If we plot the ordered standardized residuals $e^*_{(i)}$ versus $f_{(i)}$, then, in addition, the line should have slope one. Deciding on whether the points are suitably close to linear is difficult and may be aided by simulating multiple datasets from which intervals may be derived for each i. Care must be taken in interpretation as (5.62) shows that the observed residuals are a linear combination of the error terms and hence may exhibit *supernormality*, that is, even if ϵ_i is not normal, $\sum_{j=1}^{n} h_{ij}\epsilon_j$ may tend toward normality (and dominate the first term, ϵ_i).

Figure 5.8 shows what we might expect to see under various distributional assumptions. QQ normal plots for normal. Laplacian, Student's t_3, and lognormal error distributions are displayed in the four rows, with sample sizes of $n = 10, 25, 50, 200$ across columns. The characteristic skewed shape of the lognormal distribution is revealed for all sample sizes, but it is difficult to distinguish between the Laplacian and the normal, even for a large sample size. For small n, interpretation is very difficult.

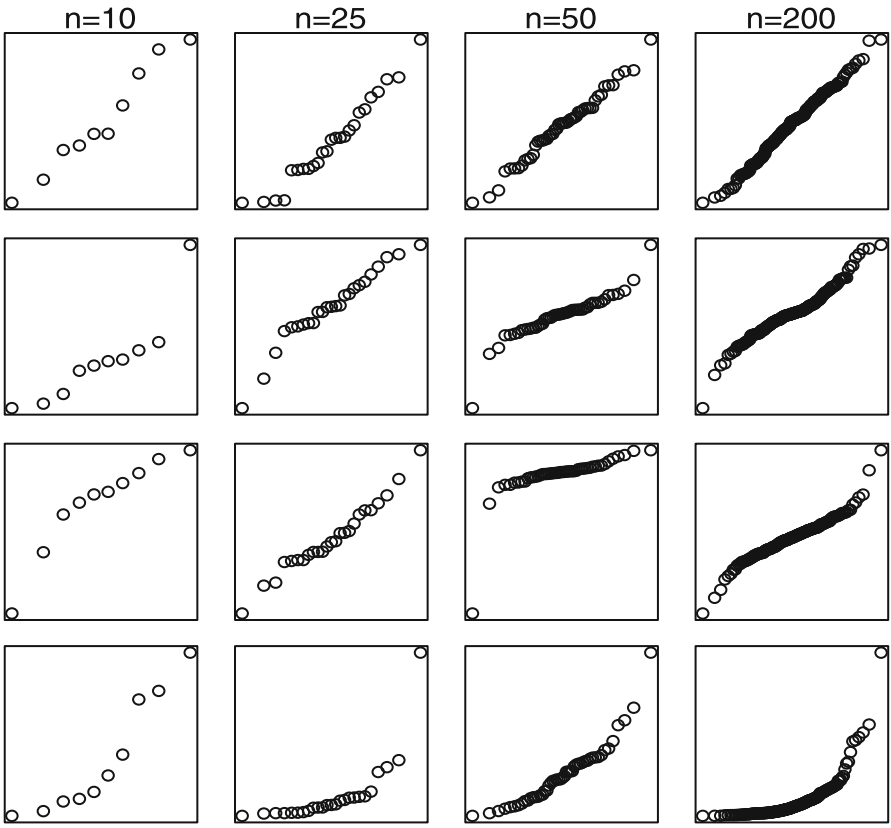

Fig. 5.8 Normal scores plot for various distributions and sample sizes. Columns 1–4 represent sample sizes of 10, 25, 50, and 200. Rows 1–4 correspond to errors generated from normal, Laplacian, Student's t_3, and lognormal distributions, respectively. In each plot, the expected residuals are plotted on the x-axis, and the observed ordered residuals on the y-axis

In general, simulation may be used to examine the behavior of plots when the model is true. QQ plots may be constructed to assess any distributional assumption, by an appropriate choice of $f_{(i)}$. The Bayesian approach to inference allow alternative likelihoods to the normal to be fitted relatively easily under an MCMC implementation. We have concentrated on frequentist residuals, but all of the above plots may be based on Bayesian residuals. For example, we can obtain samples from the posterior distribution of $\boldsymbol{\beta}$ and σ and then substitute these samples into

$$e_i^\star = \frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma(1-h_{ii})^{1/2}}, \tag{5.65}$$

to produce samples from the posterior distribution of the residuals. The posterior mean or median of the $e_i^\star$ can then be calculated and examined. More simply, one

Table 5.14 Parameter estimates and standard errors (model-based and sandwich) for the prostate cancer data

Variable	Estimate	Standard error	
		Model-based	Sandwich
log(can vol)	0.59	0.088	0.077
log(weight)	0.45	0.17	0.19
age	−0.020	0.011	0.0094
log(BPH)	0.11	0.058	0.057
SVI	0.77	0.24	0.21
log(cap pen)	−0.11	0.091	0.079
gleason	0.045	0.16	0.13
PGS45	0.0045	0.0044	0.0042
$\widehat{\sigma}$	0.78	–	–

could substitute the posterior means or medians of $\boldsymbol{\beta}$ and σ into (5.65). An early use of Bayesian residuals analysis was provided by Chaloner and Brant (1988).

A major problem with residual analysis, unless one is in purely exploratory mode, is that if the assumptions are found wanting and we change the model, what are the frequentist properties in terms of bias, the coverage of intervals, and the α level of tests? Recall the discussion of Chap. 4. To avoid changing the model, including transforming x and/or y, one should try and think as much as possible about a suitable model, *before* the data are analyzed. As always the exact procedure followed should be reported, so that inferential summaries can be more easily interpreted. The same problems exist for a Bayesian analysis, since one should specify a priori all models that one envisages fitting (which may not be feasible in advance), with subsequent averaging across models (Sect. 3.6).

5.12 Example: Prostate Cancer

We return to the PSA data and provide a more comprehensive analysis. We fit the full (main effects only) model

$$\begin{aligned}\log \text{PSA} = \beta_0 &+ \beta_1 \times \log(\text{can vol}) + \beta_2 \times \log(\text{weight}) + \beta_3 \times \text{age} + \beta_4 \times \log(\text{bph}) \\ &+ \beta_5 \times \text{svi} + \beta_6 \times \log(\text{cap pen}) + \beta_7 \times \text{gleason} + \beta_8 \times \text{PGS45} + \epsilon,\end{aligned}$$

with $\epsilon \,|\, \sigma^2 \sim_{iid} \text{N}(0, \sigma^2)$. The resultant least squares parameter estimates and standard errors are given in Table 5.14. This table includes the sandwich standard errors, to address the possibility of nonconstant variance error terms. These are virtually identical to the model-based standard errors. This is not surprising given Fig. 5.9(a), which plots the absolute value of the residuals against the fitted values, and indicates that the constant variance assumption appears reasonable.

With $n-k-1 = 88$, we do not require normality of errors, but for illustration we include a QQ normal plot in Fig. 5.9(b) and see that the errors are close to normal. Figures 5.9(c) and (d) plot the residuals versus two of the more important covariates,

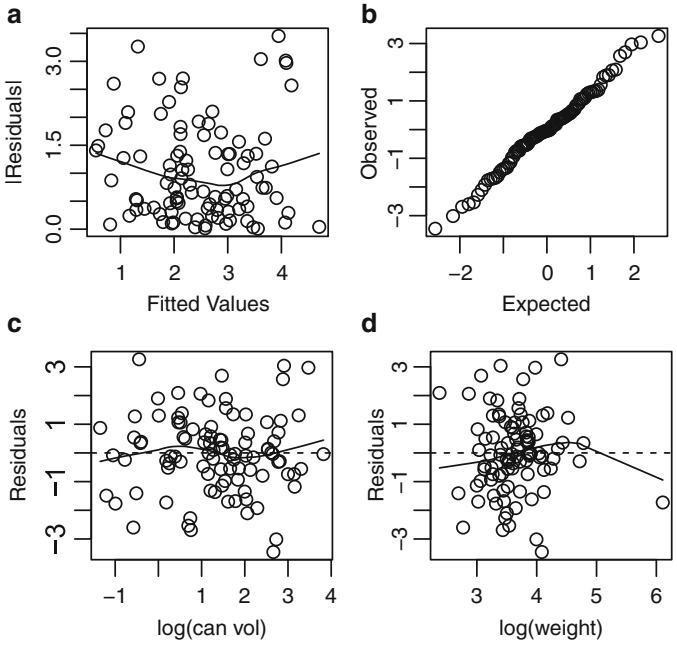

Fig. 5.9 Diagnostic plots in the prostate cancer study: (**a**) absolute values of residuals versus fitted values, with smoother, (**b**) normal QQ plot of residuals; (**c**) residuals versus log cancer volume, with smoother, (**d**) residuals versus log weight, with smoother

log cancer volume and log weight, with smoothers added. In each case, we see no strong evidence of nonlinearity.

We now discuss a Bayesian analysis of these data. With the improper prior (5.42), we saw in Sect. 5.7 that inference was identical with the frequentist approach so that the estimates and (model-based) standard errors in Table 5.14 are also posterior means and posterior standard deviations. Figure 5.10 displays the marginal posterior densities (which are located and scaled Student's t distributions with 88 degrees of freedom) for the eight coefficients. In this plot, for comparability, we scale each of the x variables to lie on the range (0,1).

Turning now to an informative prior distribution, without more specific knowledge, we let $\boldsymbol{\beta}^\star = [\beta_0^\star, \ldots, \beta_8^\star]^{\mathrm{T}}$ represent the vector of coefficients associated with the standardized covariates on (0,1). The prior is taken as $\pi(\boldsymbol{\beta}^\star)\pi(\sigma^2)$ with

$$\pi(\boldsymbol{\beta}^\star) = \prod_{j=0}^{8} \pi(\beta_j^\star) \tag{5.66}$$

and $\pi(\beta_0^\star) \propto 1$ (an improper prior). For the regression coefficients $\beta_j^\star \sim_{iid} \mathrm{N}(0, V)$ with the standard deviations, $\sqrt{V}$, chosen in the following way. For the prostate

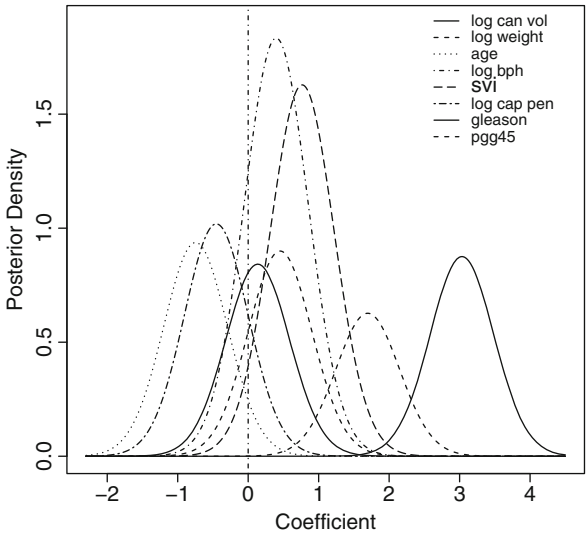

Fig. 5.10 Marginal posterior distributions of regression coefficients associated with the eight (standardized) covariates, for the prostate cancer data

data, we believe that it is unlikely that any of the standardized covariates, over the range (0,1), will change the median PSA by more than 10 units. The way we include this information in the prior is by assuming that the $1.96 \times \sqrt{V}$ point of the prior corresponds to the maximum value we believe is a priori plausible, that is, we set $\beta_j^\star = \log(10)$ equal to this point. For σ^2, we assume the improper choice $\pi(\sigma^2) \propto \sigma^{-2}$.

Figure 5.11 shows the 95% credible intervals under the flat and informative priors, and we see the general shrinkage towards zero (the prior mean). On average there is around a 10% reduction in the posterior standard deviations (and hence the credible intervals) under the informative prior, which shows how the use of informative priors can aid in the bias-variance trade-off. The above analysis is closely related to *ridge regression*, as will be discussed in Sect. 10.5.1.

5.13 Concluding Remarks

In this chapter we have concentrated on the linear model

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ is $n \times (k+1)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. Although the range of models that are routinely available for fitting has expanded greatly (see Chaps. 6 and 7), the linear model continues to be popular. There are good reasons for this, since parameter interpretation is straightforward and the estimators commonly used are linear in the data and therefore possess desirable robustness properties.

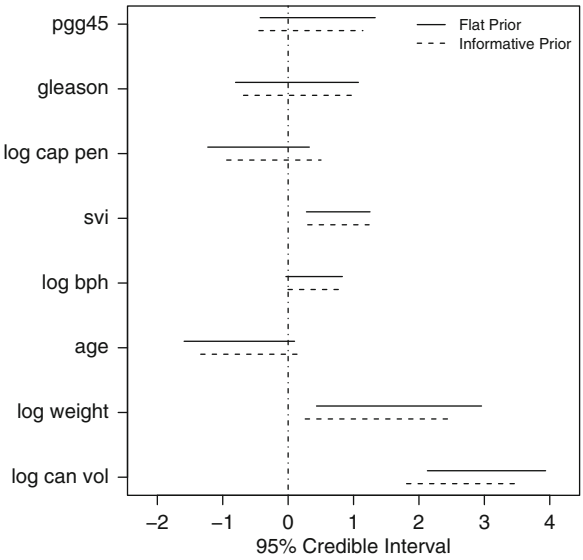

Fig. 5.11 95% credible intervals for regression coefficients corresponding to standardized covariates, under flat and informative priors, for the prostate cancer data

Unless n is not large, or there is substantial prior information, the point estimate

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$$

and $100(1-\alpha)\%$ interval estimate

$$\widehat{\beta}_j \pm t^{n-k-1}_{1-\alpha/2} \times \widehat{\text{s.e.}}(\widehat{\beta}_j),$$

where $t^{n-k-1}_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ point of a Student's t distribution with $n-k-1$ degrees of freedom, emerges from likelihood, ordinary least squares, and Bayesian analyses. These summaries are robust to a range of distributions for the error terms, so long as n is large. Nonconstant error variance and correlated errors can both seriously damage the appropriateness of the interval estimate, however. With larger sample sizes, sandwich estimation provides a good approach for guarding against nonconstant error variance.

5.14 Bibliographic Notes

McCullagh and Nelder (1989, Chap. 3) provide an extended discussion on parameterization issues, including aliasing, and the interpretation of parameters. For more discussion of conditions for asymptotic normality for simple linear regression, see (van der Vaart 1998, p.21). Firth (1987) discusses the loss of precision when the data are not normally distributed and shows that the skewness of the true distribution of the errors is an important factor. The theory presented in Lehmann (1986,

p. 209–211) indicates that dependence in the residuals can cause real problems for estimation of appropriate standard errors. Further details of residual analysis may be found in Cook and Weisberg (1982).

The classic frequentist text on the analysis of variance is Scheffé (1959), while Searle et al. (1992) provide a more recent treatment. An interesting discussion, from a Bayesian slant, is provided by Gelman and Hill (2007, Chap. 22).

Numerous texts have been written on the linear model; see, for example, Ravishanker and Dey (2002) and Seber and Lee (2003) for the theory and Faraway (2004) for a more practical slant.

5.15 Exercises

5.1 Consider the model

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of responses, $\boldsymbol{x}$ is the $n \times (k+1)$ design matrix, $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_k]$, and $\mathrm{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$, $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V}$ is a *known* correlation matrix $\boldsymbol{V}$.

(a) By considering the sum of squares,

$$\mathrm{RSS}_{\mathrm{v}} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}).$$

show that the *generalized* least squares estimator is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{v}} = (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{Y},$$

provided the necessary inverse exists.

(b) Derive the distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{v}}$.

(c) Show that $\widehat{\sigma}^2_{\mathrm{v}}$, as defined in (5.33), is an unbiased estimator of σ^2.

5.2 Suppose $\widehat{\boldsymbol{\beta}}_1 \neq \widehat{\boldsymbol{\beta}}_2$ are two different least squares estimates of $\boldsymbol{\beta}$. Show there are infinitely many least squares estimates of $\boldsymbol{\beta}$.

5.3 Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, where $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$ and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. Prove that the least squares estimates of β_0 and β_1 are uncorrelated if and only if $\overline{x} = 0$.

5.4 Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

with $\epsilon_i \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Suppose the prior distribution is of the form

$$\pi(\beta_0, \beta_1, \sigma^2) = \pi(\beta_0, \beta_1) \times \sigma^{-2}, \tag{5.67}$$

where the prior for $[\beta_0, \beta_1]$ is

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} m_0 \\ m_1 \end{bmatrix}, \begin{bmatrix} v_{00} & v_{01} \\ v_{01} & v_{11} \end{bmatrix} \right).$$

In this exercise the conditional distributions required for Gibbs sampling (Sect. 3.8.4) will be derived.

(a) Write down the form of the posterior distribution (up to proportionality) and derive the conditional distributions $p(\beta_0 \mid \beta_1, \sigma^2, \boldsymbol{y})$, $p(\beta_1 \mid \beta_0, \sigma^2, \boldsymbol{y})$, and $p(\sigma^2 \mid \beta_0, \beta_1, \boldsymbol{y})$. Hence, give details of the Gibbs sampling algorithm.
(b) Another blocked Gibbs sampling algorithm (Sect. 3.8.6) would simulate from the distributions $p(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y})$ and $p(\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{y})$. Derive these distributions, given in (5.46) and (5.47), and hence describe the form of the Gibbs sampling algorithm.

5.5 The algorithm derived in Exercise 5.4(b) will now be implemented for the prostate cancer data of Sect. 1.3.1. These data are available in the `R` package `lasso2` and are named `Prostate`. Take Y as log prostate specific antigen and x as log cancer volume. Implement the blocked Gibbs sampling algorithm using the prior (5.67), with $m_0 = m_1 = 0$, $v_{00} = v_{11} = 2$, and $v_{01} = 0$. Run two chains, one with starting values corresponding to the unbiased estimates of the parameters and one starting from a point randomly generated from the prior $\pi(\beta_0, \beta_1)$. Report:

(a) Histogram representations of the univariate marginal distributions $p(\beta_0 \mid \boldsymbol{y})$, $p(\beta_1 \mid \boldsymbol{y})$, and $p(\sigma \mid \boldsymbol{y})$ and scatterplots of the bivariate marginal distributions $p(\beta_0, \beta_1 \mid \boldsymbol{y})$, $p(\beta_0, \sigma \mid \boldsymbol{y})$, and $p(\beta_1, \sigma \mid \boldsymbol{y})$.
(b) The posterior means, standard deviations, and 10%, 50%, 90% quantiles for β_0, β_1, and σ.
(c) $\Pr(\beta_1 > 0.5 \mid \boldsymbol{y})$.
(d) Justify your choice of "burn-in" period (Sect. 3.8.6). For example, you may present the trace plots $\beta_0^{(t)}, \beta_0^{(t)}, \log \sigma^{2(t)}$ versus t.
(e) Confirm the results you have obtained using `INLA` or `WinBUGS`.

5.6 In this question, parameter interpretation will be considered. Consider a continuous univariate response y, with two potential covariates, a continuous variable x_1, and a binary factor x_2. The x variables will be referred to as age and gender, respectively. Consider the four models:

Model A

$$y = \begin{cases} \theta_0 + \epsilon, & \text{for men } (x_2 = 0) \\ \theta_1 + \epsilon, & \text{for women } (x_2 = 1). \end{cases}$$

Model B

$$y = \theta_0 + \theta_1 x_1 + \epsilon.$$

Model C

$$y = \begin{cases} \theta_0 + \theta_1 x_1 + \epsilon, & \text{for men } (x_2 = 0) \\ \theta_2 + \theta_1 x_1 + \epsilon, & \text{for women } (x_2 = 1). \end{cases}$$

Model D

$$y = \begin{cases} \theta_0 + \theta_1 x_1 + \epsilon, & \text{for men } (x_2 = 0) \\ (\theta_0 + \phi_0) + \theta_1 x_1 + \epsilon, & \text{for women } (x_2 = 1). \end{cases}$$

Model E

$$y = \begin{cases} \theta_0 + \theta_1 x_1 + \epsilon, & \text{for men } (x_2 = 0), \text{ and} \\ \theta_0 + \theta_2 x_1 + \epsilon, & \text{for women } (x_2 = 1). \end{cases}$$

Model F

$$y = \begin{cases} \theta_0 + \theta_1 x_1 + \epsilon, & \text{for men } (x_2 = 0), \\ \theta_2 + \theta_3 x_1 + \epsilon, & \text{for women } (x_2 = 1). \end{cases}$$

For each model, the error terms ϵ are assumed to have zero mean.

(a) For each model, provide a careful interpretation of the parameters and give a description of the assumed form of the relationship.
(b) Which of the above models are equivalent?

5.7 Let $Y_1, \ldots, Y_n$ be distributed as $Y_i \mid \theta, \sigma^2 \sim_{ind} \mathrm{N}(i\theta, i^2\sigma^2)$ for $i = 1, \ldots, n$. Find the generalized least squares estimate of θ and prove that the variance of this estimate is σ^2/n.

5.8 Suppose that the design matrix $\boldsymbol{x}$ of dimension $n \times (k+1)$ has rank $k+1$ and let $\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}$ represent the hat matrix. Show that $\mathrm{tr}(\boldsymbol{h}) = (k+1)$.

5.9 Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

for $i = 1, \ldots, n$, where

$$\begin{bmatrix} X_i \\ \epsilon_i \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} \mu_x \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{bmatrix} \right),$$

to give

$$\begin{bmatrix} Y_i \\ X_i \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{bmatrix} \right)$$

where $\sigma_y^2 = \beta_1^2\sigma_x^2 + \sigma_\epsilon^2$, $\mu_y = \beta_0 + \beta_1\mu_x$ and $\sigma_{xy} = \beta_1\sigma_x^2$.

(a) Derive $\mathrm{E}[Y_i \mid x_i]$ and $\mathrm{var}(Y_i \mid x_i)$.

Now suppose one does not observe x_i, $i = 1, \ldots, n$ but instead $w_i = x_i + u_i$ where

$$\begin{bmatrix} X_i \\ \epsilon_i \\ U_i \end{bmatrix} \sim \mathrm{N}_3 \left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{bmatrix} \right).$$

Assume that Y_i is conditionally independent of W_i, that is, $\mathrm{E}[Y_i \mid x_i, u_i] = \mathrm{E}[Y_i \mid x_i]$. Suppose the true model is $\mathrm{E}[Y_i \mid x_i] = \beta_0 + \beta_1 x_i$ but the observed data are $[w_i, y_i]$, $i = 1, \ldots, n$.

(b) Relate $\mathrm{E}[Y_i \mid w_i]$ to $\mathrm{E}[x_i \mid w_i]$.
(c) What is the joint distribution of X_i and W_i and what is $\mathrm{E}[X_i \mid w_i]$?
(d) Using your answers to (b) and (c), show that $\mathrm{E}[Y_i \mid w_i] = \beta_0^\star + \beta_1^\star x_i$.
(e) What is the relationship between $\beta_0^\star, \beta_1^\star$ and β_0, β_1?