drug treatment received (drug A, drug B or placebo, for example). Somewhat confusingly, the groups of a factor variable are referred to as *levels* although the groups generally have no natural ordering, and even if they do, the model structure ignores it.

To understand the construction of $\mathbf{X}$ it helps to consider an example. Suppose that along with y_i we have metric predictor variables x_i and z_i and factor variable g_i, which contains labels dividing y_i into three groups. Suppose further that we believe the following model to be appropriate:

$$y_i = \gamma_{g_i} + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 z_i^2 + \alpha_4 z_i x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where there is a different γ parameter for each of the three levels of g_i. Collecting the γ and α parameters into one vector, $\boldsymbol{\beta}$, we can rewrite the model in matrix-vector form as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x_1 & z_1 & z_1^2 & z_1x_1 \\ 1 & 0 & 0 & x_2 & z_2 & z_2^2 & z_2x_2 \\ 1 & 0 & 0 & x_3 & z_3 & z_3^2 & z_3x_3 \\ 0 & 1 & 0 & x_4 & z_4 & z_4^2 & z_4x_4 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 1 & x_n & z_n & z_n^2 & z_nx_n \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

where y_1 - y_3 are in group 1, y_4 is in group 2, and y_n is in group 3 of the factor g. Notice how the factor levels/groups each get a dummy indicator column in the model matrix, with elements showing whether the corresponding y_i belongs to the group or not. Notice also how the metric variables can enter the model nonlinearly: the model is linear in the parameters and error term, but not necessarily in the predictors.

7.1 The theory of linear models

This section shows how the parameters, $\boldsymbol{\beta}$, of the linear model

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{I}_n\sigma^2) \tag{7.1}$$

can be estimated by least squares. It is assumed that $\mathbf{X}$ is a matrix, with n rows, p columns and rank p ($n > p$). It is also shown that the resulting estimator, $\hat{\boldsymbol{\beta}}$, is unbiased and that, given the normality of the data, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2)$. Results are also derived for setting confidence limits

on parameters and for testing hypotheses about parameters: in particular the hypothesis that several elements of $\boldsymbol{\beta}$ are simultaneously zero.

In this section it is important not to confuse the *length* of a vector with its *dimension*. For example $(1, 1, 1)^{\mathrm{T}}$ has dimension 3 and length $\sqrt{3}$. Also note that no distinction has been made notationally between random variables and particular observations of those random variables: it is usually clear from the context which is meant.

7.1.1 Least squares estimation of $\boldsymbol{\beta}$

Point estimates of the linear model parameters, $\boldsymbol{\beta}$, can be obtained by the method of least squares; that is, by minimising the residual sum of squares

$$\mathcal{S} = \sum_{i=1}^{n} (y_i - \mu_i)^2,$$

with respect to $\boldsymbol{\beta}$, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. This fitting objective follows directly from the log likelihood for the model, but even without the assumption of normality, the *Gauss-Markov theorem* says that minimising $\mathcal{S}$ w.r.t. $\boldsymbol{\beta}$ will produce the minimum variance linear unbiased estimator of $\boldsymbol{\beta}$.

To use least squares with a linear model, written in general matrix-vector form, first recall the link between the Euclidean length of a vector and the sum of squares of its elements. If $\mathbf{v}$ is any vector of dimension, n, then $\|\mathbf{v}\|^2 \equiv \mathbf{v}^{\mathrm{T}}\mathbf{v} \equiv \sum_{i=1}^{n} v_i^2$. Hence

$$\mathcal{S} = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Since $\mathcal{S}$ is simply the squared (Euclidian) length of the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, its value will be unchanged if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is rotated or reflected. This observation is the basis for a practical method for finding $\hat{\boldsymbol{\beta}}$ and for developing the distributional results required to use linear models.

Specifically, as with any real matrix, $\mathbf{X}$ can always be decomposed

$$\mathbf{X} = \mathbf{Q}\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_f\mathbf{R}, \tag{7.2}$$

where $\mathbf{R}$ is a $p \times p$ upper triangular matrix,[2] and $\mathbf{Q}$ is an $n \times n$ orthogonal matrix, the first p columns of which form $\mathbf{Q}_f$. Recall that orthogonal matrices rotate/reflect vectors, but do not change their length. Orthogonality also means that $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_n$. Multiplying $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ by $\mathbf{Q}^{\mathrm{T}}$ implies

[2] That is, $R_{i,j} = 0$ if $i > j$. See also Section B.5.

that

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{Q}^{\mathrm{T}}\mathbf{y}-\mathbf{Q}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\|^2 = \left\|\mathbf{Q}^{\mathrm{T}}\mathbf{y}-\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}\right\|^2.$$

Defining p vector $\mathbf{f}$ and $n-p$ vector $\mathbf{r}$ so that $\begin{bmatrix}\mathbf{f}\\ \mathbf{r}\end{bmatrix} \equiv \mathbf{Q}^{\mathrm{T}}\mathbf{y}$, yields[3]

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \left\|\begin{bmatrix}\mathbf{f}\\ \mathbf{r}\end{bmatrix}-\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}\right\|^2 = \|\mathbf{f}-\mathbf{R}\boldsymbol{\beta}\|^2+\|\mathbf{r}\|^2.$$

The length of $\mathbf{r}$ does not depend on $\boldsymbol{\beta}$ and $\|\mathbf{f}-\mathbf{R}\boldsymbol{\beta}\|^2$ can be reduced to zero by choosing $\boldsymbol{\beta}$ so that $\mathbf{R}\boldsymbol{\beta}$ equals $\mathbf{f}$. Hence,

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{f} \tag{7.3}$$

is the least squares estimator of $\boldsymbol{\beta}$. Notice that $\|\mathbf{r}\|^2 = \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2$, the residual sum of squares for the model fit.

7.1.2 The distribution of $\hat{\boldsymbol{\beta}}$

The distribution of the estimator, $\hat{\boldsymbol{\beta}}$, follows from that of $\mathbf{Q}^{\mathrm{T}}\mathbf{y}$. Multivariate normality of $\mathbf{Q}^{\mathrm{T}}\mathbf{y}$ follows from that of $\mathbf{y}$, and since the covariance matrix of $\mathbf{y}$ is $\mathbf{I}_n\sigma^2$, the covariance matrix of $\mathbf{Q}^{\mathrm{T}}\mathbf{y}$ is

$$\mathbf{V}_{\mathbf{Q}^{\mathrm{T}}\mathbf{y}} = \mathbf{Q}^{\mathrm{T}}\mathbf{I}_n\mathbf{Q}\sigma^2 = \mathbf{I}_n\sigma^2.$$

Furthermore,

$$E\begin{bmatrix}\mathbf{f}\\ \mathbf{r}\end{bmatrix} = E(\mathbf{Q}^{\mathrm{T}}\mathbf{y}) = \mathbf{Q}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}$$

$$\Rightarrow E(\mathbf{f}) = \mathbf{R}\boldsymbol{\beta} \text{ and } E(\mathbf{r}) = \mathbf{0}.$$

So we have that

$$\mathbf{f} \sim N(\mathbf{R}\boldsymbol{\beta}, \mathbf{I}_p\sigma^2) \text{ and } \mathbf{r} \sim N(\mathbf{0}, \mathbf{I}_{n-p}\sigma^2)$$

with both vectors independent of each other.

Turning to the properties of $\hat{\boldsymbol{\beta}}$ itself, unbiasedness follows immediately:

$$E(\hat{\boldsymbol{\beta}}) = \mathbf{R}^{-1}E(\mathbf{f}) = \mathbf{R}^{-1}\mathbf{R}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

[3] If the final equality is not obvious recall that $\|\mathbf{x}\|^2 = \sum_i x_i^2$, so if $\mathbf{x} = \begin{bmatrix}\mathbf{v}\\ \mathbf{w}\end{bmatrix}$, $\|\mathbf{x}\|^2 = \sum_i v_i^2 + \sum_i w_i^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2$.

Since the covariance matrix of $\mathbf{f}$ is $\mathbf{I}_p\sigma^2$, it also follows from (1.5) in Section 1.5.1 that the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathbf{V}_{\hat{\beta}} = \mathbf{R}^{-1}\mathbf{I}_p\mathbf{R}^{-\mathrm{T}}\sigma^2 = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\sigma^2. \tag{7.4}$$

Furthermore, since $\hat{\boldsymbol{\beta}}$ is just a linear transformation of the normal random vector $\mathbf{f}$, it must have a multivariate normal distribution:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{V}_{\hat{\beta}}).$$

This result is not usually directly useful for making inferences about $\boldsymbol{\beta}$, because σ^2 is generally unknown and must be estimated, thereby introducing an extra component of variability that should be accounted for.

7.1.3 $(\hat{\beta}_i-\beta_i)/\hat{\sigma}_{\hat{\beta}_i} \sim t_{n-p}$

This section derives a result that *is* generally useful for testing hypotheses about individual β_i, as well as for finding confidence intervals for β_i. Since the $n-p$ elements of $\mathbf{r}$ are i.i.d. $N(0,\sigma^2)$ random variables,

$$\frac{1}{\sigma^2}\|\mathbf{r}\|^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n-p} r_i^2 \sim \chi^2_{n-p}$$

(see Section A.1.2). The mean of a χ^2_{n-p} r.v. is $n-p$, so this result is sufficient (but not necessary) to imply that

$$\hat{\sigma}^2 = \|\mathbf{r}\|^2/(n-p) \tag{7.5}$$

is an unbiased estimator of σ^2. The independence of the elements of $\mathbf{r}$ and $\mathbf{f}$ also implies that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.[4]

Now consider a single-parameter estimator, $\hat{\beta}_i$, with standard deviation, $\sigma_{\hat{\beta}_i}$, given by the square root of element i,i of $\mathbf{V}_{\hat{\beta}}$. An unbiased estimator of $\mathbf{V}_{\hat{\beta}}$ is $\hat{\mathbf{V}}_{\hat{\beta}} = \mathbf{V}_{\hat{\beta}}\hat{\sigma}^2/\sigma^2 = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\hat{\sigma}^2$, so an estimator, $\hat{\sigma}_{\hat{\beta}_i}$, is given by the square root of element i,i of $\hat{\mathbf{V}}_{\hat{\beta}}$, and it is clear that $\hat{\sigma}_{\hat{\beta}_i} = \sigma_{\hat{\beta}_i}\hat{\sigma}/\sigma$. Hence, using Section A.1.3,

$$\frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{\hat{\beta}_i-\beta_i}{\sigma_{\hat{\beta}_i}\hat{\sigma}/\sigma} = \frac{(\hat{\beta}_i-\beta_i)/\sigma_{\hat{\beta}_i}}{\sqrt{\frac{1}{\sigma^2}\|\mathbf{r}\|^2/(n-p)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p} \tag{7.6}$$

(where the independence of $\hat{\beta}_i$ and $\hat{\sigma}^2$ has been used). This result enables

[4] Recall that $\|\mathbf{r}\|^2 = \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2$.

confidence intervals for β_i to be found and is the basis for hypothesis tests about individual β_is (for example, $\mathrm{H}_0 : \beta_i = 0$).

7.1.4 F-ratio results

It is also of interest to obtain distributional results for testing, for example, the simultaneous equality to zero of several model parameters. Such tests are particularly useful for making inferences about factor variables and their interactions, because each factor (or interaction) is typically represented by several elements of $\boldsymbol{\beta}$. Suppose that we want to test

$$\mathrm{H}_0 : \boldsymbol{\mu} = \mathbf{X}_0\boldsymbol{\beta}_0 \text{ against } \mathrm{H}_1 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}_0$ is 'nested' within $\mathbf{X}$ (meaning that $\mathbf{X}\boldsymbol{\beta}$ can exactly match any $\mathbf{X}_0\boldsymbol{\beta}_0$, but the reverse is not true). Without loss of generality we can assume that things are actually arranged so that $\mathbf{X} = [\mathbf{X}_0 : \mathbf{X}_1]$: it is always possible to re-parameterise the model so that this is the case. Suppose that $\mathbf{X}_0$ and $\mathbf{X}_1$ have $p - q$ and q columns, respectively, and let $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ be the corresponding subvectors of $\boldsymbol{\beta}$. The null hypothesis can hence be rewritten as $\mathrm{H}_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

Now consider (7.2), the original QR decomposition of $\mathbf{X}$, in partitioned form:

$$\mathbf{X} = \mathbf{Q}\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix} \Rightarrow \mathbf{Q}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}$$
$$\Rightarrow \mathbf{Q}^{\mathrm{T}}[\mathbf{X}_0 : \mathbf{X}_1] = \begin{bmatrix}\tilde{\mathbf{R}}_0 : \mathbf{R}_1\\ \mathbf{0}\end{bmatrix} \Rightarrow \mathbf{Q}^{\mathrm{T}}\mathbf{X}_0 = \begin{bmatrix}\tilde{\mathbf{R}}_0\\ \mathbf{0}\end{bmatrix},$$

where $\tilde{\mathbf{R}}_0$ is the first $p - q$ columns of $\mathbf{R}$. Since $\mathbf{R}$ is upper triangular, the last q rows of $\tilde{\mathbf{R}}_0$ are 0, so let $\mathbf{R}_0$ denote the first $p - q$ rows of $\tilde{\mathbf{R}}_0$ (i.e. the first $p - q$ rows and columns of $\mathbf{R}$). Rotating $\mathbf{y} - \mathbf{X}_0\boldsymbol{\beta}_0$ using $\mathbf{Q}^{\mathrm{T}}$ implies that

$$\|\mathbf{y} - \mathbf{X}_0\boldsymbol{\beta}_0\|^2 = \left\|\mathbf{Q}^{\mathrm{T}}\mathbf{y} - \begin{bmatrix}\mathbf{R}_0\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_0\right\|^2 = \|\mathbf{f}_0 - \mathbf{R}_0\boldsymbol{\beta}_0\|^2 + \|\mathbf{f}_1\|^2 + \|\mathbf{r}\|^2,$$

where $\mathbf{Q}^{\mathrm{T}}\mathbf{y}$ has been partitioned into $\mathbf{f}$ and $\mathbf{r}$, exactly as before, but $\mathbf{f}$ has then been further partitioned into $p - q$ vector $\mathbf{f}_0$ and q vector $\mathbf{f}_1$ so that $\mathbf{f} = \begin{bmatrix}\mathbf{f}_0\\ \mathbf{f}_1\end{bmatrix}$. Since the residual sum of squares for this null model is now $\|\mathbf{f}_1\|^2 + \|\mathbf{r}\|^2$, $\|\mathbf{f}_1\|^2$ is the increase in the residual sum of squares that results from dropping $\mathbf{X}_1$ from the model (i.e. from setting $\boldsymbol{\beta}_1 = \mathbf{0}$).

That is, $\|\mathbf{f}_1\|^2$ is the difference in residual sum of squares between the 'full model' and the 'null model'.

Now, we know that $\mathbf{f} \sim N(\mathbf{R}\boldsymbol{\beta}, \mathbf{I}_p\sigma^2)$, but in addition we know that $\boldsymbol{\beta}_1 = \mathbf{0}$ under H_0 (i.e. the last q elements of $\boldsymbol{\beta}$ are zero). Hence

$$E\begin{bmatrix}\mathbf{f}_0\\ \mathbf{f}_1\end{bmatrix} = \mathbf{R}\boldsymbol{\beta} = (\tilde{\mathbf{R}}_0 : \mathbf{R}_1)\begin{bmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_1\end{bmatrix} = (\tilde{\mathbf{R}}_0 : \mathbf{R}_1)\begin{bmatrix}\boldsymbol{\beta}_0\\ \mathbf{0}\end{bmatrix}$$
$$= \tilde{\mathbf{R}}_0\boldsymbol{\beta}_0 = \begin{bmatrix}\mathbf{R}_0\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_0 = \begin{bmatrix}\mathbf{R}_0\boldsymbol{\beta}_0\\ \mathbf{0}\end{bmatrix}.$$

So, if H_0 is true, $E(\mathbf{f}_1) = \mathbf{0}$ and $\mathbf{f}_1 \sim N(\mathbf{0}, \mathbf{I}_q\sigma^2)$. Consequently

$$\frac{1}{\sigma^2}\|\mathbf{f}_1\|^2 \sim \chi^2_q.$$

We also know that $\mathbf{f}_1$ and $\mathbf{r}$ are independent. So, forming an *F-ratio statistic*, assuming H_0 and using Section A.1.4, we have

$$F = \frac{\|\mathbf{f}_1\|^2/q}{\hat{\sigma}^2} = \frac{\frac{1}{\sigma^2}\|\mathbf{f}_1\|^2/q}{\frac{1}{\sigma^2}\|\mathbf{r}\|^2/(n-p)} \sim \frac{\chi^2_q/q}{\chi^2_{n-p}/(n-p)} \sim F_{q,n-p}, \quad (7.7)$$

and this result can be used to find the p-value for the hypothesis test. Remember that the term $\|\mathbf{f}_1\|^2$ is the difference in residual sum of squares between the two models being compared, and q is the difference in their degrees of freedom. So we could also write F as

$$F = \frac{(\|\mathbf{y} - \mathbf{X}_0\hat{\boldsymbol{\beta}}_0\|^2 - \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2)/\{\dim(\boldsymbol{\beta}) - \dim(\boldsymbol{\beta}_0)\}}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\{n - \dim(\boldsymbol{\beta})\}}.$$

7.1.5 The influence matrix

One useful matrix is the *influence matrix* (or *hat matrix*) of a linear model. This is the matrix that yields the fitted value vector, $\hat{\boldsymbol{\mu}}$, when post-multiplied by the data vector, $\mathbf{y}$. Recalling the definition of $\mathbf{Q}_f$, as being the first p columns of $\mathbf{Q}$, $\mathbf{f} = \mathbf{Q}_f^{\mathrm{T}}\mathbf{y}$, and so

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}_f^{\mathrm{T}}\mathbf{y}.$$

Furthermore $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X} = \mathbf{Q}_f\mathbf{R}$ so

$$\hat{\boldsymbol{\mu}} = \mathbf{Q}_f\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}_f^{\mathrm{T}}\mathbf{y} = \mathbf{Q}_f\mathbf{Q}_f^{\mathrm{T}}\mathbf{y}.$$

So the matrix $\mathbf{A} \equiv \mathbf{Q}_f\mathbf{Q}_f^{\mathrm{T}}$ is the influence (hat) matrix such that $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$.

The influence matrix has two interesting properties. First, the trace of the