# 3 Multiple regression

## 3.1 Simple linear regression

In simple linear regression, we assume that responses $Y_1, Y_2, \ldots, Y_n$ are uncorrelated with common variance $\sigma^2$ and expectations of the form $\beta_0 + \beta_1 x_i$ given the values $x_1, x_2, \ldots, x_n$ of an explanatory variable. We can rewrite the $n$ expectations in vector notation by defining the $n$-vectors $\mathbf{x}$ and $\mathbf{Y}$ with components $x_1, x_2, \ldots, x_n$ and $Y_1, Y_2, \ldots, Y_n$ respectively. If $\mathbf{1}_n$ denotes an $n$-vector of 1's, we have

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{x}) = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \beta_1 = \mathbf{1}_n \beta_0 + \mathbf{x} \beta_1 \,. \tag{3.1.1}$$

A more concise notation uses an $n \times 2$ matrix $\mathbf{X}$ whose columns are $\mathbf{1}_n$ and $\mathbf{x}$, and a 2-vector $\beta$ whose elements are the unknown parameters $\beta_0$ and $\beta_1$. The model then becomes

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{x}) = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \mathbf{X}\beta \,. \tag{3.1.2}$$

The assumptions about the variances and covariances of the $Y_i$ (given the $x_i$) can also be expressed in matrix notation as

$$\mathrm{var}(\mathbf{Y}\,|\,\mathbf{x}) = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_n \,. \tag{3.1.3}$$

The expression $\mathbf{X}\beta$ for the expectation of $\mathbf{Y}$ (given a known matrix $\mathbf{X}$) can be used for a wide range of statistical models if $\mathbf{X}$ and $\beta$ are suitably defined. For example, the alternative formulation of simple linear regression (Question 2, Problem Sheet 1) has $\mathrm{E}(Y_i\,|\,x_i) = \gamma + \beta_1 (x_i - \bar{x})$ $(i = 1, \ldots, n)$ or

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{x}) = \begin{pmatrix} \gamma + \beta_1 (x_1 - \bar{x}) \\ \gamma + \beta_1 (x_2 - \bar{x}) \\ \vdots \\ \gamma + \beta_1 (x_n - \bar{x}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \gamma + \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} \beta_1 \,. \tag{3.1.4}$$

This can be put in the form $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$ by taking

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma \\ \beta_1 \end{pmatrix} . \tag{3.1.5}$$

Similarly, simple linear regression through the origin has $\mathrm{E}(Y_i\,|\,x_i) = \beta x_i$ or

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \beta \,. \tag{3.1.6}$$

## 3.2 Some other linear models

The following are some of the other linear statistical models which can be expressed using the matrix formulation $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$. Again the expectation of $\mathbf{Y}$ is conditional on the values of any explanatory variables which are contained in $\mathbf{X}$.

(a) **Regression on two or more explanatory variables**. With explanatory variables $x_1$ and $x_2$, the $n$ expectations given by $\mathrm{E}(Y_i\,|\,x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ $(i = 1, \ldots, n)$ may be combined into a single equation as

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{X}\beta. \tag{3.2.1}$$

This model can be extended to a regression equation with $q$ explanatory variables:

$$\mathrm{E}(Y_i\,|\,x_{i1}, \ldots, x_{iq}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} \quad (i = 1, \ldots, n) \tag{3.2.2}$$

For the extended model, $\mathbf{X}$ and $\beta$ are given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nq} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}. \tag{3.2.3}$$

(b) **Random sample from a distribution**. If random variables $Y_1, \ldots, Y_n$ have a common distribution with expectation $\mu$, then the vector of expectations is $\mathrm{E}(\mathbf{Y}) = \mu\mathbf{1}_n$: this has the form $\mathbf{X}\beta$ with $\mathbf{X} = \mathbf{1}_n$ and $\beta = \mu$.

(c) **Random samples from two distributions**. Suppose that random variables $Y_1, \ldots, Y_m$ have a common distribution with expectation $\mu_1$, and $Y_{m+1}, \ldots, Y_n$ have a common distribution with expectation $\mu_2$. Then the random vector with elements $Y_1, \ldots, Y_n$ has expectation

$$\mathrm{E}(\mathbf{Y}) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \mathbf{1}_m \\ \mu_2 \mathbf{1}_{n-m} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_m & \mathbf{0}_m \\ \mathbf{0}_{n-m} & \mathbf{1}_{n-m} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}. \tag{3.2.4}$$

(d) **Two simple linear regressions**. Suppose that responses $Y_1, \ldots, Y_m$ satisfy the simple linear regression $\mathrm{E}(Y_i\,|\,x_i) = \alpha_1 + \beta_1 x_i$, while $Y_{m+1}, \ldots, Y_n$ satisfy $\mathrm{E}(Y_i\,|\,x_i) = \alpha_2 + \beta_2 x_i$ (using $\alpha$ and $\beta$ rather than $\beta_0$ and $\beta_1$ for the intercepts and slopes to avoid double subscripts). Thus the first $m$ responses follow one regression equation while the remaining $n-m$ follow another (as might be assumed in Example 1.9). Writing $\mathbf{x}_1 = (x_1 \ldots x_m)^T$ and $\mathbf{x}_2 = (x_{m+1} \ldots x_n)^T$, the combined model may be expressed as

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \begin{pmatrix} \alpha_1 \mathbf{1}_m + \beta_1 \mathbf{x}_1 \\ \alpha_2 \mathbf{1}_{n-m} + \beta_2 \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_m & \mathbf{x}_1 & \mathbf{0}_m & \mathbf{0}_m \\ \mathbf{0}_{n-m} & \mathbf{0}_{n-m} & \mathbf{1}_{n-m} & \mathbf{x}_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}, \tag{3.2.5}$$

## 3.3 The 'Normal Linear Model'

Let $\mathbf{Y}$ be a random $n$-vector of responses, $\mathbf{X}$ an $n \times p$ matrix (with $n > p$) whose elements are known values $x_{ij}$, and $\beta$ a $p$-vector of unknown parameters. For the *Normal Linear Model,* we assume

$$\begin{aligned} \mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) &= \mathbf{X}\beta, & (3.3.1)\\ \mathrm{var}(\mathbf{Y}\,|\,\mathbf{X}) &= \sigma^2\,\mathbf{I}_n\,. & (3.3.2) \end{aligned}$$

The $n$ elements $\mathrm{E}(Y_i\,|\,\mathbf{X})$ of $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X})$, are then given by

$$\mathrm{E}(Y_i\,|\,\mathbf{X}) = \sum_{j=1}^{p} x_{ij}\,\beta_j\,, \qquad (3.3.3)$$

which is a *linear* function of the coefficients $\beta_j$. Thus the Normal Linear Model is linear in the $\beta$'s. Even a quadratic regression model with $\mathrm{E}(Y_i\,|\,x_i) = \beta_0 + \beta_1\,x_i + \beta_2\,x_i^2$ is linear in this sense, as is a model $\mathrm{E}(Y_i\,|\,x_i) = \beta_0 + \beta_1\,\ln x_i$.

If the model contains an 'intercept' or 'constant term', such as $\beta_0$ in (3.1.1) and (3.2.1) or $\gamma$ in (3.1.4), this corresponds to a column of 1's in $\mathbf{X}$. If $\mathbf{X}$ has full rank $p$ then the $p \times p$ matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, and the least squares estimates are unique. It is sometimes convenient to consider a model which is not of full rank, and to define the least squares estimates using generalized inverses. For making inferences about $\beta$, we also assume that $\mathbf{Y}$ has a multivariate Normal distribution (given $\mathbf{X}$). The distribution of $\mathbf{Y}$ is then $N_n(\mathbf{X}\beta, \sigma^2\,\mathbf{I}_n)$.

## 3.4 Least squares estimation

For least squares estimation we find the values of the $\beta$'s which minimize the sum of squares

$$Q = \sum_{i=1}^{n} \{y_i - \mathrm{E}(Y_i\,|\,\mathbf{X})\}^2 \qquad (3.4.1)$$

for the observed responses $y_1, \ldots, y_n$. In terms of vectors and matrices, the function to be minimized is

$$\begin{aligned} Q &= \{\mathbf{y} - \mathrm{E}(\mathbf{Y}\,|\,\mathbf{X})\}^T\,\{\mathbf{y} - \mathrm{E}(\mathbf{Y}\,|\,\mathbf{X})\}\\ &= (\mathbf{y} - \mathbf{X}\beta)^T\,(\mathbf{y} - \mathbf{X}\beta) \qquad (3.4.2)\\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta. \end{aligned}$$

Using Section 10 of *Useful Matrix Results*, the vector of partial derivatives of $Q$ with respect to the $\beta$'s is given by

$$\frac{\partial Q}{\partial \beta} = 2\left(\mathbf{X}^T\mathbf{X}\beta - \mathbf{X}^T\mathbf{y}\right). \qquad (3.4.3)$$

Equating this vector to $\mathbf{0}$, the vector $\widehat{\beta}$ of least squares estimates satisfies the $p$ normal equations

$$\mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T\mathbf{y}\,. \qquad (3.4.4)$$

These may also be written

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \mathbf{0}\,. \qquad (3.4.5)$$

Note that the $jk$-th element of $\mathbf{X}^T\mathbf{X}$ and the $j$-th element of $\mathbf{X}^T\mathbf{y}$ are respectively

$$\sum_{i=1}^{n} x_{ij}x_{ik}\,, \quad \sum_{i=1}^{n} x_{ij}y_i\,. \tag{3.4.6}$$

If $\mathbf{X}$ has full rank $p$ then $(\mathbf{X}^T\mathbf{X})^{-1}$ exists and there is a unique *least squares estimate* of $\beta$ given by

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,; \tag{3.4.7}$$

otherwise the estimate is not unique. To show that a solution of (3.4.4) gives a minimum of $Q$, note that

$$\begin{aligned} Q &= \left\{(\mathbf{y}-\mathbf{X}\widehat{\beta}) + \mathbf{X}(\widehat{\beta}-\beta)\right\}^T \left\{(\mathbf{y}-\mathbf{X}\widehat{\beta}) + \mathbf{X}(\widehat{\beta}-\beta)\right\} \\ &= (\mathbf{y}-\mathbf{X}\widehat{\beta})^T(\mathbf{y}-\mathbf{X}\widehat{\beta}) + (\widehat{\beta}-\beta)^T\mathbf{X}^T\mathbf{X}(\widehat{\beta}-\beta) + 2(\widehat{\beta}-\beta)^T\mathbf{X}^T(\mathbf{y}-\mathbf{X}\widehat{\beta})\,. \end{aligned} \tag{3.4.8}$$

The third term is 0 (by (3.4.5)); the first and second terms are non-negative. Hence

$$Q \geq (\mathbf{y}-\mathbf{X}\widehat{\beta})^T(\mathbf{y}-\mathbf{X}\widehat{\beta})\,, \tag{3.4.9}$$

and the minimum is attained when $\beta = \widehat{\beta}$. The right-hand side of (3.4.9) is called the *residual sum of squares* for the Normal Linear Model, and is considered in §3.7.

## 3.5 Expectation and variance matrix of least squares estimator

If $\mathbf{X}$ has full rank $p$ then the least squares estimator of $\beta$ is

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\,. \tag{3.5.1}$$

This is a linear function of $\mathbf{Y}$, which makes finding its expectation and variance matrix straightforward.

$$\begin{aligned} \mathrm{E}(\widehat{\beta}\,|\,\mathbf{X}) &= \mathrm{E}\left\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\,|\,\mathbf{X}\right\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= \beta\,, \end{aligned} \tag{3.5.2}$$

$$\begin{aligned} \mathrm{var}(\widehat{\beta}\,|\,\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ \mathrm{var}(\mathbf{Y}\,|\,\mathbf{X})\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\sigma^2\mathbf{I}_n\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\,. \end{aligned} \tag{3.5.3}$$

Hence if we estimate a linear function $\mathbf{c}^T\beta$ of $\beta$, the least squares estimator $\mathbf{c}^T\widehat{\beta}$ has variance

$$\begin{aligned} \mathrm{var}(\mathbf{c}^T\widehat{\beta}\,|\,\mathbf{X}) &= \mathbf{c}^T\,\mathrm{var}(\widehat{\beta}\,|\,\mathbf{X})\,\mathbf{c} \\ &= \sigma^2\mathbf{c}^T\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\,. \end{aligned} \tag{3.5.4}$$

## 3.6 Fitted values and residuals

If the linear model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$, $\mathrm{var}(\mathbf{Y}\,|\,\mathbf{X}) = \sigma^2\,\mathbf{I}_n$ defined in §3.3 is fitted using least squares, the vector of *fitted values* is $\mathbf{X}\widehat{\beta}$, and the vector of (raw) *residuals* is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\beta}. \tag{3.6.1}$$

Hence the normal equations (3.4.5) for the least squares estimates may be expressed as

$$\mathbf{X}^T\mathbf{e} = \mathbf{0}, \tag{3.6.2}$$

showing that the vector of residuals is orthogonal to each of the $p$ columns of $\mathbf{X}$. If the model includes an 'intercept' (usually written as $\beta_0$) then $\mathbf{X}$ includes the column $\mathbf{1}_n$, so that (3.6.2) implies

$$\sum_{i=1}^{n} e_i = \mathbf{1}_n^T\mathbf{e} = 0, \tag{3.6.3}$$

and the raw residuals sum to zero.

### 3.6.1 Fitted values and residuals as projections

If $\mathbf{X}$ has full rank $p$ then the vectors of fitted values and residuals are given (using (3.4.7)) by

$$\begin{aligned}\mathbf{X}\widehat{\beta} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{P}_\mathbf{X}\mathbf{y}\end{aligned} \tag{3.6.4}$$

and

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{P}_\mathbf{X}\mathbf{y} \\ &= (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{y},\end{aligned} \tag{3.6.5}$$

where $\mathbf{P}_\mathbf{X}$ is defined by

$$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{3.6.6}$$

The matrix $\mathbf{P}_\mathbf{X}$ is $n \times n$, symmetric, idempotent and of rank $p$; $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ is therefore $n \times n$, symmetric and idempotent and has rank $n - p$. Also

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{X} = \mathbf{0}, \tag{3.6.7}$$

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{P}_\mathbf{X} = \mathbf{0}. \tag{3.6.8}$$

The matrices $\mathbf{P}_\mathbf{X}$ and $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ both represent projections in $R^n$: $\mathbf{X}\widehat{\beta}$ (which equals $\mathbf{P}_\mathbf{X}\mathbf{y}$) is the projection of the vector $\mathbf{y}$ of responses onto the column space $\mathcal{C}(\mathbf{X})$ of $\mathbf{X}$, (i.e. the $p$-dimensional subspace spanned by the columns of $\mathbf{X}$) and $\mathbf{e}$ is the projection of $\mathbf{y}$ onto the orthogonal complement of $\mathcal{C}(\mathbf{X})$, i.e. the $(n-p)$-dimensional subspace orthogonal to $\mathcal{C}(\mathbf{X})$.

### 3.6.2 Expectation and variance matrix of the residuals

If the vector $\mathbf{E}$ of residuals is considered as a random vector, its expectation and variance matrix are as follows (using (3.6.7) and the idempotency of $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$).

$$\mathrm{E}(\mathbf{E}\,|\,\mathbf{X}) = (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{X}\beta = \mathbf{0}, \tag{3.6.9}$$

$$\mathrm{var}(\mathbf{E}\,|\,\mathbf{X}) = (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\sigma^2\mathbf{I}_n\,(\mathbf{I}_n - \mathbf{P}_\mathbf{X}) = \sigma^2\,(\mathbf{I}_n - \mathbf{P}_\mathbf{X})^2 = \sigma^2\,(\mathbf{I}_n - \mathbf{P}_\mathbf{X}). \tag{3.6.10}$$

## 3.7 Estimation of $\sigma^2$

The *residual sum of squares* is the minimum value of $Q$, as given by the right hand side of (3.4.9). It is also the sum of the squares of the residuals $e_1, \ldots, e_n$. It may be expressed in several ways (using (3.6.1), (3.6.5), (3.6.6), (3.4.7) and the symmetry and idempotency of $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$) as follows.

$$\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \mathbf{e}^T\mathbf{e} \\
&= (\mathbf{y} - \mathbf{X}\widehat{\beta})^T(\mathbf{y} - \mathbf{X}\widehat{\beta}) \\
&= \mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{y} \\
&= \mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\,\mathbf{y} && (3.7.1) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \widehat{\beta}^T\mathbf{X}^T\mathbf{y} && (3.7.2) \\
&= \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{p} \widehat{\beta}_j \sum_{i=1}^{n} x_{ij}y_i. && (3.7.3)
\end{aligned}$$

The residual sum of squares can be shown to have expectation $(n-p)\sigma^2$, so we estimate $\sigma^2$ using the *residual mean square*,

$$\widehat{\sigma}^2 = \frac{\text{residual sum of squares}}{n-p} = \frac{\mathbf{y}^T\mathbf{y} - \widehat{\beta}^T\mathbf{X}^T\mathbf{y}}{n-p}. \qquad (3.7.4)$$

The *model sum of squares* is the portion of the total sum of squares accounted for by fitting the model. It is therefore expressible as

$$\mathbf{y}^T\mathbf{P}_\mathbf{X}\mathbf{y} = \widehat{\beta}^T\mathbf{X}^T\mathbf{y} = \sum_j \widehat{\beta}_j \sum_i x_{ij}y_i. \qquad (3.7.5)$$

The last of these expressions is convenient for hand calculation. Rearranging (3.7.2), the total sum of squares, $\sum_i y_i^2$ or $\mathbf{y}^T\mathbf{y}$, may be decomposed into the *model sum of squares* $\mathbf{y}^T\mathbf{P}_\mathbf{X}\mathbf{y}$ and the *residual sum of squares* $\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{y}$. These two sums of squares may be interpreted as the squared lengths of the projection $\mathbf{X}\widehat{\beta} = \mathbf{P}_\mathbf{X}\mathbf{y}$ of $\mathbf{y}$ onto the column space $\mathcal{C}(\mathbf{X})$ of $\mathbf{X}$ and of the projection $\mathbf{e} = (\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{y}$ onto the subspace orthogonal to $\mathcal{C}(\mathbf{X})$; these two squared lengths sum to $\mathbf{y}^T\mathbf{y}$ by Pythagoras' Theorem.

## 3.8 Distributions of the sums of squares under Normality

Now consider the two sums of squares and the least squares estimator $\widehat{\beta}$ as random variables, and suppose that the vector $\mathbf{Y}$ of responses is Normally distributed, so that its distribution is $N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$. Because the matrices $\mathbf{P}_\mathbf{X}$ and $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ are symmetric and idempotent with ranks $p$ and $n-p$, it follows that

(a) the model sum of squares $\mathbf{Y}^T\mathbf{P}_\mathbf{X}\mathbf{Y}$ has the distribution

$$\sigma^2\chi^2(p, \sigma^{-2}\beta^T\mathbf{X}^T\mathbf{P}_\mathbf{X}\mathbf{X}\beta) \quad \text{or} \quad \sigma^2\chi^2(p, \sigma^{-2}\beta^T\mathbf{X}^T\mathbf{X}\beta);$$

(b) the residual sum of squares $\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}$ has the distribution

$$\sigma^2\chi^2(n-p, \sigma^{-2}\beta^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{X}\beta) \quad \text{or} \quad \sigma^2\chi^2(n-p);$$

(c) the two sums of squares are independent because $\mathbf{P_X}(\mathbf{I}_n - \mathbf{P_X}) = \mathbf{0}$;

(d) since $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}_n - \mathbf{P_X}) = \mathbf{0}$, the residual sum of squares is independent of $\widehat{\beta}$.

## 3.9 An alternative formulation for models with an intercept

Consider a linear regression model of the form

$$\mathrm{E}(Y_i \,|\, \mathbf{X}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq} \quad (i = 1, \ldots, n). \tag{3.9.1}$$

Here the 'intercept' or 'constant term' $\beta_0$ corresponds to a column of 1's in $\mathbf{X}$, as in (3.1.1), (3.1.4), (3.2.1) and (3.2.3). For this sort of model, we usually decompose the sum of squares $\sum_i (y_i - \overline{y})^2$ about the mean response $\overline{y}$ rather than the (raw) total $\sum_i y_i^2$ considered in Section 3.8. It is convenient to use the equivalent model

$$\mathrm{E}(Y_i \,|\, \mathbf{X}) = \gamma + \beta_1(x_{i1} - \overline{x}_1) + \beta_2(x_{i2} - \overline{x}_2) + \ldots + \beta_q(x_{iq} - \overline{x}_q) \quad (i = 1, \ldots, n) \tag{3.9.2}$$

in which explanatory variables are measured from their means (generalizing the model in Question 2, Problem Sheet 1). In matrix terms this becomes

$$\mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) = \gamma \mathbf{1}_n + \dot{\mathbf{X}}\dot{\beta}, \tag{3.9.3}$$

where the $n \times q$ matrix $\dot{\mathbf{X}}$ has $ij$-th element $x_{ij} - \overline{x}_j$ and

$$\left.\begin{array}{l} \dot{\beta} = \left(\beta_1 \ldots \beta_q\right)^T, \\ \gamma = \beta_0 + \beta_1\overline{x}_1 + \ldots + \beta_q\overline{x}_q. \end{array}\right\} \tag{3.9.4}$$

The values of the $q$ explanatory variables are said to be *centred* when the means are subtracted. The sum over the $j$-th column of $\dot{\mathbf{X}}$ is $\sum_i (x_{ij} - \overline{x}_j) = 0$ $(j = 1, \ldots, q)$ i.e. it satisfies

$$\dot{\mathbf{X}}^T \mathbf{1}_n = \mathbf{0}_q. \tag{3.9.5}$$

### 3.9.1 Least squares estimation

The least squares estimates $\widehat{\gamma}$ and $\widehat{\dot{\beta}}$ of $\gamma$ and $\dot{\beta}$ in (3.9.4) satisfy the $q+1$ linear equations

$$\mathbf{1}_n^T \left(\mathbf{y} - \widehat{\gamma}\mathbf{1}_n - \dot{\mathbf{X}}\widehat{\dot{\beta}}\right) = 0, \tag{3.9.6}$$

$$\dot{\mathbf{X}}^T \left(\mathbf{y} - \widehat{\gamma}\mathbf{1}_n - \dot{\mathbf{X}}\widehat{\dot{\beta}}\right) = \mathbf{0}, \tag{3.9.7}$$

or (using (3.9.5))

$$\widehat{\gamma} = \overline{y}, \tag{3.9.8}$$

$$\dot{\mathbf{X}}^T\dot{\mathbf{X}}\widehat{\dot{\beta}} = \dot{\mathbf{X}}^T\mathbf{y}, \tag{3.9.9}$$

so that

$$\widehat{\dot{\beta}} = (\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}^T\mathbf{y} \tag{3.9.10}$$

if $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ is non-singular. In contrast to (3.4.6), the matrix $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ has $jk$-th element

$$\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(x_{ik}-\overline{x}_k) = \sum_{i=1}^{n} x_{ij}x_{ik} - n^{-1}\sum_{i=1}^{n} x_{ij}\sum_{i=1}^{n} x_{ik}\,, \tag{3.9.11}$$

a sum of squares or products *about the mean*; the vector $\dot{\mathbf{X}}^T\mathbf{y}$ has $j$-th element

$$\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)y_i = \sum_{i=1}^{n} x_{ij}y_i - n^{-1}\sum_{i=1}^{n} x_{ij}\sum_{i=1}^{n} y_i\,. \tag{3.9.12}$$

The estimators $\widehat{\dot{\beta}}$ and $\widehat{\gamma}=\bar{Y}$ are unbiased; their variances and covariances are given by

$$\mathrm{var}(\widehat{\dot{\beta}}\,|\,\mathbf{X}) = \sigma^2\,(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\,,\quad \mathrm{var}\,(\widehat{\gamma}|\,\mathbf{X}) = n^{-1}\sigma^2\,,\quad \mathrm{cov}(\widehat{\dot{\beta}},\widehat{\gamma}|\,\mathbf{X}) = \mathbf{0}\,. \tag{3.9.13}$$

Since $\widehat{\gamma}$ is uncorrelated with $\widehat{\dot{\beta}}$, the variance of a linear function of the estimators is given by

$$\mathrm{var}\left(c_0\widehat{\gamma}+\mathbf{c}^T\widehat{\dot{\beta}}\,|\,\mathbf{X}\right) = \sigma^2\left\{n^{-1}c_0^2+\mathbf{c}^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\mathbf{c}\right\}\,. \tag{3.9.14}$$

The model defined in (3.9.2) and (3.9.3) is a special case of the model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$ of (3.3.1) in which $p=q+1$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \dot{\mathbf{X}} \end{pmatrix}\,. \tag{3.9.15}$$

Using (3.9.2) simplifies computation because (from (3.9.5)) $\mathbf{X}^T\mathbf{X}$ becomes

$$\begin{pmatrix} \mathbf{1}_n & \dot{\mathbf{X}} \end{pmatrix}^T \begin{pmatrix} \mathbf{1}_n & \dot{\mathbf{X}} \end{pmatrix} = \begin{pmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \dot{\mathbf{X}}^T\dot{\mathbf{X}} \end{pmatrix}\,, \tag{3.9.16}$$

so that only $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ has to be inverted to estimate the parameters and the variances and covariances in (3.9.13). The replacement of the $q+1$ normal equations by the $q$ equations in (3.9.9) and

$$\widehat{\beta}_0+\overline{x}_1\,\widehat{\beta}_1+\ldots+\overline{x}_q\,\widehat{\beta}_q = \overline{y} \tag{3.9.17}$$

can be interpreted as the first step in a Gaussian elimination: the first row of $\mathbf{X}^T\mathbf{X}$, which equals $(n\ \sum_i x_{i1}\ \ldots\ \sum_i x_{iq})$ or $n(1\ \ \overline{x}_1\ \ldots\ \overline{x}_q)$ is multiplied by $\overline{x}_1$ and subtracted from the second row, multiplied by $\overline{x}_2$ and subtracted from the third row, and so on. Thus (3.9.13) is consistent with expression (3.5.3) for $\mathrm{var}(\widehat{\beta}\,|\,\mathbf{X})$: $(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}$ is the bottom right $q\times q$ sub-matrix of $(\mathbf{X}^T\mathbf{X})^{-1}$.

### 3.9.2 Sums of squares

The decomposition of $\sum_i y_i^2$ corresponding to the model (3.9.3) is

$$\begin{aligned}\mathbf{y}^T\mathbf{y} &\equiv \mathbf{y}^T\left\{n^{-1}\mathbf{1}_n\mathbf{1}_n^T\right\}\mathbf{y}+\mathbf{y}^T\left\{\dot{\mathbf{X}}(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}^T\right\}\mathbf{y}+\mathbf{y}^T\left\{\mathbf{H}_n-\dot{\mathbf{X}}(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}^T\right\}\mathbf{y}\\ &\equiv n^{-1}(\sum_i y_i)^2 \qquad + \qquad \widehat{\dot{\beta}}^T\dot{\mathbf{X}}^T\mathbf{y} \qquad + \qquad \text{residual SS}\,,\end{aligned} \tag{3.9.18}$$

where $\mathbf{H}_n$ denotes $\mathbf{I}_n-n^{-1}\mathbf{1}_n\mathbf{1}_n^T$. Subtracting $n^{-1}(\sum_i y_i)^2$ from both sides of (3.9.18) gives the following decomposition of the total sum of squares *about the mean* into the *regression sum of* *squares* and the *residual sum of squares*. This decomposition is generally used for a regression model which includes an intercept:

$$\sum_i (y_i - \overline{y})^2 \equiv \mathbf{y}^T \mathbf{H}_n \mathbf{y} \equiv \widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \mathbf{y} \quad + \quad \text{residual SS}\ . \tag{3.9.19}$$

If the response vector $\mathbf{Y}$ is Normally distributed, with distribution $N_n(\gamma \mathbf{1}_n + \dot{\mathbf{X}}\dot{\beta}, \sigma^2 \mathbf{I}_n)$, the joint distribution of the sums of squares is as follows:

(a) the regression sum of squares has distribution $\sigma^2 \chi^2 \left(q, \sigma^{-2} \dot{\beta}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \dot{\beta}\right)$;

(b) the residual sum of squares has distribution $\sigma^2 \chi^2 (n-q-1)$;

(c) the two sums of squares are independent.

We again estimate $\sigma^2$ using the *residual mean square*, now defined as

$$\widehat{\sigma}^2 = \frac{\text{residual sum of squares}}{\text{residual d.f.}} = \frac{\sum_i (y_i - \overline{y})^2 - \widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \mathbf{y}}{n-q-1}\ . \tag{3.9.20}$$