Applied Statistics
Lectures 1–8: Linear Models

Neil Laws

Michaelmas Term 2022 (version of 30-09-2022)

In some places these notes will contain more than the lectures and in other places they may contain less. There will be plenty of examples involving data in lectures. These examples are an extremely important part of the course, just as important as the notes – the examples will be available separately, they are not part of this document.

Most of these notes are closely based on Geoff Nicholls' Applied Statistics notes. I have added material in places, in particular these notes include more detail in parts of Section 1 as the course assumes less about normal linear models than it did a few years ago.

If you spot any errors please let me know (`neil.laws@stats.ox.ac.uk`).

Updates:

For MT 2019 these notes were updated so that $\mathbf{x}_i$ is always a *column* vector – previously $\mathbf{x}_i$ was a row vector in the linear models part of the course. If you spot any places where $\mathbf{x}_i$ still needs to be changed to $\mathbf{x}_i^T$ please let me know. You will need to watch out for this if you refer to old course material.

MT 2022:

Nothing yet

Contents

0 Preliminaries

0.1 Course Info

- Lectures 1–8: Linear Models (LMs) = Neil Laws
- Lectures 9–13: Generalised Linear Models (GLMs) = Frank Windmeijer

These lectures are SB1.1 Applied Statistics and are shared by undergraduates and MSc students.

There are also practicals and problems classes – these are separate for UG and MSc.

0.2 R

I suggest that you install R and RStudio and start practising with them as soon as possible.

`https://www.r-project.org/`

`https://www.rstudio.com/`

0.3 Sheet 0

There is a Problem Sheet 0 with questions you can practice on as these lectures begin. Solutions to these questions will be available shortly – I strongly recommend that you try the questions as much as possible (hopefully complete them all) before looking at the solutions.

These questions will not be covered in problems classes. The sheet is designed to revise some material that will be useful in this course.

I recommend that you complete Sheet 0 as soon as possible.

0.4 Books

Recommended reading:

A. C. Davison. *Statistical Models*. CUP, 2003.

J. J. Faraway. *Linear Models with R*, 2nd edition. CRC Press, 2015.

J. J. Faraway. *Extending the Linear Model with R*, 2nd edition. CRC Press, 2016.

Other very helpful texts and more advanced reading:

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical models*. CUP, 2007.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*, 4th edition. Springer, 2002.

F. L. Ramsey and D. W. Schafer. *The Statistical Sleuth*, 3rd edition. Brooks/Cole, 2013.

A. J. Dobson and A. G. Barnett. *An Introduction to Generalized Linear Models*, 3rd edition. CRC Press, 2008 – mainly for GLMs.

J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

T. A. B. Snijders and R. J. Bosker. *Multilevel Analysis*, 2nd edition. Sage, 2012.

1 Normal Linear Models

EXAMPLES 1.1. Introductory examples here.

We are interested in modelling the pattern of dependence between a *response variable* y and some *explanatory variables* $x_1, x_2, \ldots$.

Other possible names for $x_1, x_2, \ldots$ are *regressors* or *predictors* or *features* or *input variables* (or *independent variables*).

Other possible names for y are the *outcome* or *output variable* (or *dependent variable*).

1.1 Introduction

In a *linear model* we model the random variable response $Y = y$ as a linear function of the explanatory variables $x_1, \ldots, x_p$, plus a normal error:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

We assume that

- we have n observations $y_1, y_2, \ldots, y_n$ from this model
- and associated with observation y_i is a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ of the values of the p explanatory variables for observation i, for $i = 1, 2, \ldots, n$.

So we assume that

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

In (1.1),

- y_i is a random variable, the ith observed value of the response variable
- $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ are known constants (not random variables), the values of the explanatory variables for observation i (note $\mathbf{x}_i$ is a *column* vector)
- $\beta_1, \beta_2, \ldots, \beta_p$ are fixed but unknown parameters, also called *regression coefficients*, describing the dependence of y_i on $\mathbf{x}_i$
- $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are iid random variables, or "random errors"
- and σ^2 is another fixed but unknown parameter, the variance of the ϵ_i.

Let

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and define the $n \times p$ matrix X by

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

Note that $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a *row vector*, the ith row of X.

Let X_j be the jth column of X given by

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

So X_j records the values taken by explanatory variable j.

We can write the matrix X in terms of its rows, or in terms of its columns:

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = (X_1, X_2, \ldots, X_p).$$

The matrix X is called the *design matrix*. If we are able to choose X then we are designing our experiment. In some cases X is chosen for us when we have observational data in which we are not able to influence the values in X.

From (1.1) we have

$$\begin{aligned} E(y_i) &= x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p \\ &= \sum_{j=1}^{p} x_{ij}\beta_j \\ &= \mathbf{x}_i^T \beta \\ &= [X\beta]_i. \end{aligned}$$

So another way to think of (1.1) is that the y_i are independent and normal, with $E(y_i)$ as above, and with $\text{var}(y_i) = \sigma^2$.

The most economical way to write the linear model (1.1) is in matrix notation:

$$y = X\beta + \epsilon \qquad \text{where} \quad \epsilon \sim N(0_n, \sigma^2 I_n). \tag{1.2}$$

In (1.2) the 0_n means an $n \times 1$ vector of 0s and I_n is the $n \times n$ identity matrix. (In future we sometimes prefer to write just 0 rather than 0_n, for simplicity.)

So (1.2) is saying that the $n \times 1$ vector y has a multivariate normal distribution with an $n \times 1$ mean vector of $X\beta$, and an $n \times n$ covariance matrix of $\sigma^2 I_n$.

Unless otherwise stated we will assume that $p < n$ (this means that matrix X has more rows than columns) and that the columns of X are linearly independent, so that matrix X has rank p.

Example 1.2. Almost always we will have an intercept, denoted by β_0, and say m further explanatory variables, so

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \epsilon_i.$$

In this case the number of regression parameters is $p = m + 1$.

If $m = 2$ then $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$ so

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

We write this concisely as $y = X\beta + \epsilon$, where X is the $n \times 3$ matrix above.

- In the example above, and other similar examples, it is usual to start numbering the regression coefficients from zero and denote them as $\beta_0, \beta_1, \ldots$, leading to the model

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + \epsilon_i.$$

- But in the general case we will start numbering the regression coefficients from 1 and denote the p regression coefficients as $\beta_1, \ldots, \beta_p$. So our general model will be

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i.$$

  An intercept term, of β_1 in this numbering scheme, then corresponds to the first explanatory variable being $x_{i1} = 1$ for all i.

1.2 Estimators

We have $y_i \sim N(\mathbf{x}_i^T\beta, \sigma^2)$ independently for $i = 1, \ldots, n$. So the joint density of $(y_1, \ldots, y_n)$, i.e. the likelihood function, is

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T\beta)^2 \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \mathrm{SS}(\beta) \right] \end{aligned}$$

where

$$\mathrm{SS}(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2.$$

The likelihood L is a function of the $p + 1$ variables $\beta_1, \ldots, \beta_p, \sigma^2$.

The log-likelihood is, after dropping a constant of $-\frac{n}{2}\log 2\pi$,

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{SS}(\beta).$$

We see that maximising ℓ (or L) over β is equivalent to *minimising* $\mathrm{SS}(\beta)$ over β: so the vector $\widehat{\beta}$ that minimises $\mathrm{SS}(\beta)$ is both the *maximum likelihood estimator* (MLE) of β and also the *least squares estimator* of β.

If we want to emphasise that the (log-)likelihood depends on y, then we write $L(\beta, \sigma^2; y)$ or $\ell(\beta, \sigma^2; y)$.

MLE of β via differentiation

We have

$$\frac{\partial SS(\beta)}{\partial \beta_j} = -2\sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i^T\beta).$$

We obtain the MLE $\widehat{\beta}$ by putting the above partial derivatives equal to zero, i.e. by solving

$$\sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i^T\beta) = 0, \quad j = 1, \ldots, p.$$

Now $\sum_{i=1}^{n} x_{ij}(y_i - \mathbf{x}_i^T\beta)$ is the jth component of $X^T(y - X\beta)$, where X^T is the transpose of X. So the equations we want to solve are

$$[X^T(y - X\beta)]_j = 0, \quad j = 1, \ldots, p$$

or

$$X^T(y - X\beta) = 0$$

where the 0 on the right is the $p \times 1$ vector of 0s. So

$$X^T X\beta = X^T y. \tag{1.3}$$

Equations (1.3) are called the *normal equations*. The solution $\beta = \widehat{\beta}$ of (1.3) is $\widehat{\beta} = (X^T X)^{-1} X^T y$.

The matrix $X^T X$ is invertible because of the assumption that $\text{rank}(X) = p$.

MLE of β via geometric approach

We would like to minimise

$$\text{SS}(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2 = (y - X\beta)^T(y - X\beta).$$

Above, and from now on, T denotes transpose.

Let $\text{col}(X)$ denote the p-dimensional subspace spanned by the columns of X,

$$\text{col}(X) = \{z \in \mathbb{R}^n : z = X\beta \text{ for some } \beta \in \mathbb{R}^p\}.$$

Now $\text{SS}(\beta)$ is the squared length of the vector $y - X\beta$, and as β varies the value of $X\beta$ varies over $\text{col}(X)$. So $\beta = \widehat{\beta}$ minimises $\text{SS}(\beta)$ when $X\widehat{\beta}$ is the point in $\text{col}(X)$ that is closest to y. The point $\widehat{y} = X\widehat{\beta}$ is therefore the orthogonal projection of y onto $\text{col}(X)$.

So $y - X\widehat{\beta}$ is orthogonal to all vectors in $\text{col}(X)$, in particular it is orthogonal to each column of X, so

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

Hence $\widehat{\beta} = (X^T X)^{-1} X^T y$.

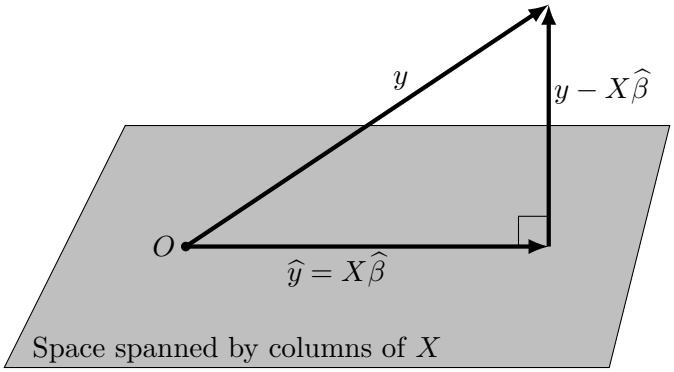


Figure 1.1: The data vector y is in $\mathbb{R}^n$. The shaded subspace is $\text{col}(X)$, and as β varies the value of $X\beta$ varies within this subspace. The value of β for which the length of $y - X\beta$ is minimised is $\widehat{\beta}$, i.e. $\widehat{y} = X\widehat{\beta}$ is the orthogonal projection of y onto $\text{col}(X)$.

MLE of σ^2

The *residual sum of squares* RSS is defined by

$$\text{RSS} = \text{SS}(\widehat{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \widehat{\beta})^2 = (y - X\widehat{\beta})^T (y - X\widehat{\beta}).$$

That is, the RSS is the minimum value of $\text{SS}(\beta)$.

Substituting $\beta = \widehat{\beta}$ into the log-likelihood we get

$$\ell(\widehat{\beta}, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \text{RSS}.$$

We find the MLE $\widehat{\sigma}^2$ of σ^2 by differentiating this with respect to σ^2 (or with respect to σ) and setting the result equal to zero. Solving this equation gives the maximising value of σ^2: it gives the MLE of

$$\widehat{\sigma}^2 = \frac{1}{n} \text{RSS}.$$

We will need this MLE when we carry out likelihood ratio tests.

However this is a biased estimator of σ^2 and we will shortly derive an unbiased estimator of σ^2.

The value of the log-likelihood at the joint MLE $(\widehat{\beta}, \widehat{\sigma}^2)$ is

$$\begin{aligned} \ell(\widehat{\beta}, \widehat{\sigma}^2) &= -\frac{n}{2} \log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2} \text{RSS} \\ &= -\frac{n}{2} \log\left(\frac{\text{RSS}}{n}\right) - \frac{n}{2}. \end{aligned} \tag{1.4}$$

We will need this expression when we get to likelihood ratio tests.

EXAMPLE 1.3. Advertising data – model fitting, interpretation here.

1.3 Properties of estimators

For our linear model (1.1), or equivalently (1.2), we want to be able to find confidence intervals for parameters and test hypotheses. To do this we need to find the distributions of $\widehat{\beta}$, $\widehat{\sigma}^2$ and related quantities.

We begin with reminders about mean vectors, covariance matrices and the multivariate normal distribution.

Revision: mean vector, covariance matrix

See also Sheet 0.

For any vector of random variables $y = (y_1, \ldots, y_n)^T$, the mean vector $E(y)$ is defined by

$$E(y) = \begin{pmatrix} E(y_1) \\ \vdots \\ E(y_n) \end{pmatrix}$$

and the covariance matrix $\mathrm{var}(y)$ is defined by

$$\mathrm{var}(y) = \begin{pmatrix} \mathrm{var}(y_1) & \mathrm{cov}(y_1, y_2) & \ldots & \mathrm{cov}(y_1, y_n) \\ \mathrm{cov}(y_2, y_1) & \mathrm{var}(y_2) & \ldots & \mathrm{cov}(y_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(y_n, y_1) & \mathrm{cov}(y_n, y_2) & \ldots & \mathrm{var}(y_n) \end{pmatrix}.$$

Let $\mu = (\mu_1, \ldots, \mu_n)^T$ be defined by $\mu = E(y)$.

Then $\mathrm{cov}(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)]$ and so $\mathrm{var}(y)$ is a symmetric matrix.

Also $\mathrm{cov}(y_i, y_i) = \mathrm{var}(y_i)$, and if y_i, y_j are independent then $\mathrm{cov}(y_i, y_j) = 0$.

Observe that

$$\begin{aligned} &(y - \mu)(y - \mu)^T \\ &= \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix} \begin{pmatrix} y_1 - \mu_1, & \ldots, & y_n - \mu_n \end{pmatrix} \\ &= \begin{pmatrix} (y_1 - \mu_1)^2 & (y_1 - \mu_1)(y_2 - \mu_2) & \ldots & (y_1 - \mu_1)(y_n - \mu_n) \\ (y_2 - \mu_2)(y_1 - \mu_1) & (y_2 - \mu_2)^2 & \ldots & (y_2 - \mu_2)(y_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (y_n - \mu_n)(y_1 - \mu_1) & (y_n - \mu_n)(y_2 - \mu_2) & \ldots & (y_n - \mu_n)^2 \end{pmatrix}. \end{aligned}$$

Taking expectations on both sides of the above equation, where "taking expectations" of a matrix means taking the expectation of each element of the matrix, we obtain the useful relation

$$\mathrm{var}(y) = E[(y - \mu)(y - \mu)^T]. \tag{1.5}$$

Suppose $a = (a_1, \ldots, a_n)^T$ is a vector of constants. Then

$$E(a^T y) = E\left(\sum_{i=1}^n a_i y_i \right) = \sum_{i=1}^n E(a_i y_i) = \sum_{i=1}^n a_i \mu_i = a^T \mu.$$

Now let A be a constant matrix with n columns. In a similar way to finding $E(a^T y)$ above, we can find $E(Ay)$ and $\text{var}(Ay)$.

(i) To find $E(Ay)$:

$$E(Ay) = AE(y) = A\mu.$$

(ii) To find $\text{var}(Ay)$:

$$\begin{aligned}
\text{var}(Ay) &= E[(Ay - A\mu)(Ay - A\mu)^T] \qquad \text{using (1.5)} \\
&= E[A(y-\mu)\{A(y-\mu)\}^T] \\
&= E[A(y-\mu)(y-\mu)^T A^T] \qquad \text{since } (AB)^T = B^T A^T \\
&= AE[(y-\mu)(y-\mu)^T]A^T \\
&= A\,\text{var}(y)A^T.
\end{aligned}$$

Revision: multivariate normal distribution

The vector $y = (y_1, \ldots, y_k)^T$ has a multivariate normal distribution if y can be written as

$$y = Az + \mu$$

where the k-vector μ and the $k \times k$ matrix A are constant, and where $z = (z_1, \ldots, z_k)^T$ with $z_1, \ldots, z_k \overset{\text{iid}}{\sim} N(0,1)$.

Here $E(z) = 0$ and $\text{var}(z) = I_k$. So y has mean vector $E(y) = \mu$ and covariance matrix $\text{var}(y) = \Sigma$, where $\Sigma = AA^T$, because:

$$\begin{aligned}
E(y) &= AE(z) + \mu = \mu \\
\text{var}(y) &= A\,\text{var}(z)A^T = AA^T.
\end{aligned}$$

We say that y is (multivariate) normal with mean μ and covariance Σ, and we write $y \sim N(\mu, \Sigma)$ or $y \sim N_k(\mu, \Sigma)$.

- Let B be a matrix, and b a vector, of constants. If $y \sim N(\mu, \Sigma)$ then $By + b \sim N(B\mu + b, B\Sigma B^T)$, i.e. linear combinations of y are also normal.

Let $\det \Sigma$ denote the determinant of Σ. Then $\det \Sigma = \det(AA^T) = (\det A)^2 \geqslant 0$.

- If $\det \Sigma = 0$ then the normal distribution of y is called singular and no probability density function exists.
- If $\det \Sigma > 0$ then the density of y is

$$f(y) = \frac{1}{(2\pi)^{k/2}(\det \Sigma)^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right\}, \quad y \in \mathbb{R}^k$$

  though we may not need this density in the course.

We will meet vectors which have singular normal distributions: e.g. the vector $\widehat{y}$ of fitted values and the vector e of residuals.

- If y is multivariate normal, then components of y are independent if and only if they are uncorrelated (i.e. iff the covariance matrix is diagonal).

- If $y \sim N(\mu, \Sigma)$ is partitioned as $y = (y^{(1)}, \ldots, y^{(m)})$ where the corresponding partitioned covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & \ldots & 0 \\ 0 & \Sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \Sigma_m \end{pmatrix}$$

  then the vectors $y^{(1)}, \ldots, y^{(m)}$ are independent.

Sometimes it is useful to know that there is a square root $\Sigma^{1/2}$ of Σ. And if $\det \Sigma > 0$ then there is a square root $\Sigma^{-1/2}$ of Σ^{-1}.

Distribution of $\widehat{\beta}$

Our general model is $y = X\beta + \epsilon$ where $\epsilon \sim N(0_n, \sigma^2 I_n)$. Hence

$$E(y) = X\beta + E(\epsilon) = X\beta \tag{1.6}$$
$$\text{var}(y) = \text{var}(\epsilon) = \sigma^2 I_n. \tag{1.7}$$

We have

$$\begin{aligned} \widehat{\beta} &= (X^T X)^{-1} X^T y \\ &= Ay \end{aligned}$$

where $A = (X^T X)^{-1} X^T$. We can now find (a) the expectation, (b) the covariance matrix, and then (c) the distribution of $\widehat{\beta}$.

(a)

$$\begin{aligned} E(\widehat{\beta}) &= (X^T X)^{-1} X^T E(y) \\ &= (X^T X)^{-1} X^T X\beta \qquad \text{using (1.6)} \\ &= \beta. \end{aligned}$$

Hence $\widehat{\beta}$ is unbiased for β.

(b)

$$\text{var}(\widehat{\beta}) = \text{var}(Ay) = A \,\text{var}(y) A^T.$$

Now $X^T X$ is symmetric, hence $(X^T X)^{-1}$ is symmetric also, hence $A^T = X(X^T X)^{-1}$. So, using $\text{var}(y) = \sigma^2 I_n$ from (1.7),

$$\begin{aligned} \text{var}(\widehat{\beta}) &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}. \end{aligned}$$

(c) We know that y is normal $y \sim N(X\beta, \sigma^2 I)$, therefore the linear combination $\widehat{\beta} = Ay$ is also normal.

Combining (a)–(c), we have that $\widehat{\beta}$ is normal with mean β and covariance $\sigma^2 (X^T X)^{-1}$, i.e.

$$\widehat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}). \tag{1.8}$$

This is a multivariate normal distribution in p-dimensions, the mean vector is $p \times 1$ and the covariance matrix is $p \times p$.

Fitted values, residuals

The vector of *fitted values* $\widehat{y}$, or the *predicted response values*, or the *estimated response values*, is defined by $\widehat{y} = X\widehat{\beta}$. So

$$\widehat{y} = X\widehat{\beta} = X(X^TX)^{-1}X^Ty = Hy$$

where the *hat matrix* H is given by $H = X(X^TX)^{-1}X^T$.

So the ith fitted value is $\widehat{y}_i = \mathbf{x}_i^T\widehat{\beta}$.

The vector of *residuals* e is defined by $e = y - \widehat{y}$.

So the ith residual is $e_i = y_i - \widehat{y}_i$.

So the residual sum of squares is $\text{RSS} = e^Te = \sum_{i=1}^n e_i^2$.

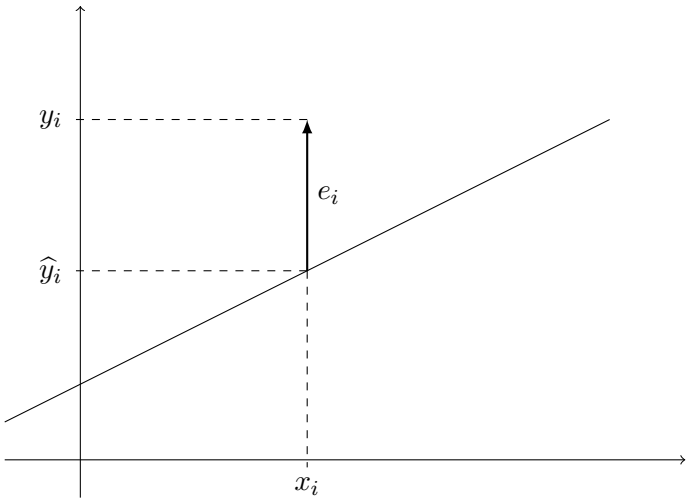


Figure 1.2: For a simple linear regression $y = \beta_0 + \beta_1 x + \epsilon$: the observed value y_i, the fitted value $\widehat{y}_i$, and the residual e_i, at $x = x_i$. The solid line is the fitted regression line. The residual e_i is positive if y_i is above the fitted line, and negative if y_i is below the fitted line.

Theorem 1.4.

(i) *e and $\widehat{y}$ are independent.*

(ii) *$\widehat{\beta}$ and RSS are independent.*

(iii) *$\text{RSS} \sim \sigma^2\chi^2_{n-p}$.*

Proof. Pick an orthonormal basis of column vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ for $\text{col}(X)$, and extend it to an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{R}^n$. Note that

$$\mathbf{v}_i^T\mathbf{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{1.9}$$

Since $y \in \mathbb{R}^n$ we can write y in terms of the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, say as

$$y = \sum_{i=1}^n z_i\mathbf{v}_i.$$

In this expression we can think of z_i being a weight, it is random variable and is given by $z_i = \mathbf{v}_i^T y$. (Multiply the above equation by $\mathbf{v}_i^T$ and use (1.9) to see this.)

For $i = 1, \ldots, p$, we have $\mathbf{v}_i \in \text{col}(X)$ and hence $H\mathbf{v}_i = \mathbf{v}_i$.

For $i = p+1, \ldots, n$, we have that $\mathbf{v}_i$ is orthogonal to the columns of X, that is

$$X^T \mathbf{v}_i = 0 \tag{1.10}$$

and hence $H\mathbf{v}_i = 0$.

So

$$\widehat{y} = Hy = H\Big(\sum_{i=1}^{n} z_i \mathbf{v}_i\Big) = \sum_{i=1}^{p} z_i \mathbf{v}_i.$$

Hence

$$\begin{aligned}
\text{RSS} &= (y - \widehat{y})^T (y - \widehat{y}) \\
&= \Big(\sum_{i=p+1}^{n} z_i \mathbf{v}_i\Big)^T \Big(\sum_{j=p+1}^{n} z_j \mathbf{v}_j\Big) \\
&= \sum_{i=p+1}^{n} z_i^2 \qquad \text{using (1.9).}
\end{aligned}$$

Let A be the $n \times n$ matrix whose rows are given by $\mathbf{v}_1^T, \ldots, \mathbf{v}_n^T$, i.e.

$$A = \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix}.$$

Note that $AA^T = I_n$, the identity matrix, by (1.9). Now y is multivariate normal and $z = Ay$, hence z is also multivariate normal (since we are taking linear combinations). The covariance matrix of z is

$$\begin{aligned}
\text{var}(z) &= \text{var}(Ay) \\
&= A \,\text{var}(y) A^T \\
&= \sigma^2 I_n \qquad \text{since } \text{var}(y) = \sigma^2 I_n \text{ and } AA^T = I_n.
\end{aligned} \tag{1.11}$$

So z is multivariate normal with a diagonal covariance matrix, hence $z_1, \ldots, z_n$ are independent.

For $i = p+1, \ldots, n$,

$$\begin{aligned}
E(z_i) &= E(\mathbf{v}_i^T y) \\
&= \mathbf{v}_i^T E(y) \\
&= \mathbf{v}_i^T X\beta \\
&= 0 \qquad \text{using (1.10).}
\end{aligned} \tag{1.12}$$

Hence, from (1.11), (1.12), and the independence of the z_i, we have

$$z_{p+1}, \ldots, z_n \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{1.13}$$

(i) We have that $\widehat{y}$ is a function of $(z_1, \ldots, z_p)$, and $e = y - \widehat{y}$ is a function of $(z_{p+1}, \ldots, z_n)$. Since these two groups of z_i are non-overlapping, and since the z_i are independent, it follows that $\widehat{y}$ and e are independent.

(ii) Now $\widehat{\beta}$ is a function of $\widehat{y}$ (i.e. $\widehat{\beta} = (X^T X)^{-1} X^T \widehat{y}$), and RSS is a function of e (i.e. $\text{RSS} = e^T e$). Using (i) it follows that $\widehat{\beta}$ and RSS are independent.

(iii) From above,

$$\frac{\text{RSS}}{\sigma^2} = \sum_{i=p+1}^{n} \left(\frac{z_i}{\sigma}\right)^2$$

and the sum on the right is a sum of $n - p$ squared independent $N(0, 1)$ random variables (see (1.13)). Hence $\text{RSS} \sim \sigma^2 \chi^2_{n-p}$. □

The estimator $s^2 = \text{RSS}/(n - p)$ is unbiased for σ^2 since

$$E(s^2) = \frac{1}{n-p} E(\text{RSS}) = \frac{\sigma^2}{n-p} E(\chi^2_{n-p}) = \sigma^2$$

since $E(\chi^2_{n-p}) = n - p$.

EXAMPLE 1.5. Advertising data – summaries here.

1.4 Tests

Testing a single parameter

Suppose we want to test the significance of a single parameter β_j. That is, we want to test $H_0 \colon \beta_j = 0$ against $H_1 \colon \beta_j \neq 0$ for one particular value of j, where $1 \leqslant j \leqslant p$.

Under H_0 and using (1.8) we have

$$\frac{\widehat{\beta}_j}{\sigma \sqrt{(X^T X)^{-1}_{jj}}} \sim N(0, 1).$$

Note: $(X^T X)^{-1}_{jj}$ means the (j, j) element of $(X^T X)^{-1}$.

We also have $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$, where this χ^2_{n-p} is independent of the above $N(0, 1)$. So with $s^2 = \text{RSS}/(n - p)$, the quantity

$$\begin{aligned} t &= \frac{\widehat{\beta}_j}{\sigma \sqrt{(X^T X)^{-1}_{jj}}} \cdot \sqrt{\frac{\sigma^2 (n-p)}{\text{RSS}}} \\ &= \frac{\widehat{\beta}_j}{s \sqrt{(X^T X)^{-1}_{jj}}} \end{aligned}$$

is a suitably scaled ratio of an independent $N(0, 1)$ and χ^2. Under H_0 the quantity t has a t-distribution with $n - p$ degrees of freedom.

If t_{obs} is the observed value of t then the p-value for our two-sided test is $2P(t_{n-p} > |t_{\text{obs}}|)$, where t_{n-p} denotes a random variable with a t-distribution with $n - p$ degrees of freedom.

We can write $t = \widehat{\beta}_j / \operatorname{se}(\widehat{\beta}_j)$ where the *standard error* of $\widehat{\beta}_j$ is defined by $\operatorname{se}(\widehat{\beta}_j) = s\sqrt{(X^TX)^{-1}_{jj}}$.

Similarly we can find a $1-\alpha$ confidence interval for β_j as $\big(\widehat{\beta}_j \pm t_{n-p}(\frac{\alpha}{2})\operatorname{se}(\widehat{\beta}_j)\big)$.

Here $t_{n-p}(\frac{\alpha}{2})$ means the upper $\frac{\alpha}{2}$ point of the t_{n-p} distribution (i.e. a t_{n-p} random variable is greater than this value with probability $\frac{\alpha}{2}$).

EXAMPLE 1.6. Advertising data – tests here.

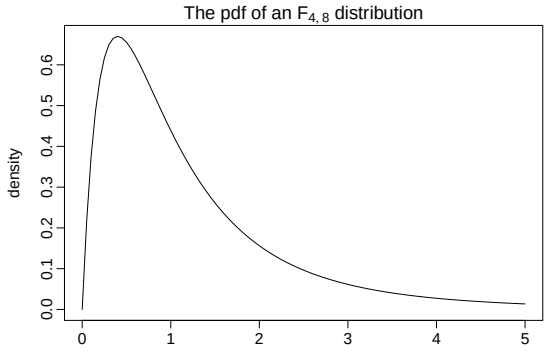
The F distribution

Suppose $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$ are independent. Then the distribution of

$$F = \frac{U_1/d_1}{U_2/d_2}$$

is called an F_{d_1,d_2} *distribution.*

We call d_1 and d_2 the *numerator and denominator degrees of freedom.* The expectation is $E(F_{d_1,d_2}) = \frac{d_2}{d_2-2}$ for $d_2 > 2$, which is about 1 when d_2 is large. The quantiles of F distributions are known and available in R/statistical tables.



Let $F_{d_1,d_2}(\alpha)$ denote the upper α point of this distribution (i.e. the $1-\alpha$ quantile).

Testing a group of parameters

When we test for the significance of a group of parameters we use a test called an *F-test.* If there is just one parameter in the group then the F-test reduces to the t-test described above.

Suppose we have a conjecture that there is no linear relation between the response y and the last k explanatory variables $x_{p-k+1}, x_{p-k+2}, \ldots, x_p$. That is, we want to test

$$H_0\colon \beta_{p-k+1} = \beta_{p-k+2} = \cdots = \beta_p = 0 \qquad (1.14)$$
$$\text{(and } \beta_1, \ldots, \beta_{p-k}, \sigma^2 \text{ are unrestricted)}$$

against the alternative that at least one of the parameters $\beta_{p-k+1}, \ldots, \beta_p$ is non-zero.

Under H_0 we are fitting the model

$$y = \sum_{j=1}^{p-k} \beta_j^{(0)} x_j + \epsilon$$

where $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_{p-k}^{(0)})^T$ is the parameter vector.

Let $\widetilde{X}$ be the matrix made up of the first $p-k$ columns of X. Then under H_0 we have $y = \widetilde{X}\beta^{(0)} + \epsilon$. When we fit this model we get $\widehat{\beta}^{(0)} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T y$. Let $H^{(0)} = \widetilde{X}(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T$ be the hat matrix for this H_0-model. Let $\widehat{y}^{(0)} = H^{(0)}y$ and

$$\mathrm{RSS}^{(0)} = (y - \widehat{y}^{(0)})^T(y - \widehat{y}^{(0)}).$$

Under H_0, the MLE for σ^2 is

$$\widehat{\sigma}_0^2 = \frac{\mathrm{RSS}^{(0)}}{n}.$$

The dimension of parameter space under H_0 is $p-k+1$ (i.e. $\beta_1, \ldots, \beta_{p-k}, \sigma^2$).

Under H_1, with $\beta = \beta^{(1)}$, we are fitting the model

$$y = \sum_{j=1}^{p} \beta_j^{(1)} x_j + \epsilon.$$

This is the usual setup and we have $\widehat{\beta}^{(1)} = (X^TX)^{-1}X^Ty$, $\widehat{y}^{(1)} = Hy$ where $H = X(X^TX)^{-1}X^T$, and

$$\mathrm{RSS}^{(1)} = (y - \widehat{y}^{(1)})^T(y - \widehat{y}^{(1)})$$

and the MLE for σ^2 is

$$\widehat{\sigma}_1^2 = \frac{\mathrm{RSS}^{(1)}}{n}.$$

The dimension of parameter space under the alternative H_1 is $p+1$.

We can now find the likelihood ratio statistic Λ for testing H_0. Using the maximised log-likelihood expression (1.4) twice (once for H_0, once for H_1) we obtain

$$\Lambda(y) = 2\left\{\ell(\widehat{\beta}^{(1)}, \widehat{\sigma}_1^2) - \ell(\widehat{\beta}^{(0)}, \widehat{\sigma}_0^2)\right\} \tag{1.15}$$

$$= n\log\left(\frac{\mathrm{RSS}^{(0)}}{\mathrm{RSS}^{(1)}}\right) \tag{1.16}$$

$$= n\log\left(1 + \frac{\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)}}{\mathrm{RSS}^{(1)}}\right). \tag{1.17}$$

We know Λ has an approximate χ_k^2 distribution for large n and this would give an approximate test of H_0. But we can do better than this.

Usng the above expression for $\Lambda(y)$, the exact likelihood ratio test of H_0 is:

$$\begin{aligned} \text{reject } H_0 &\iff \Lambda(y) > \text{constant} \\ &\iff F(y) > \text{constant} \end{aligned}$$

where

$$F(y) = \frac{(\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)})/k}{\mathrm{RSS}^{(1)}/(n-p)}. \tag{1.18}$$

Under H_0, the test statistic $F(y)$ has an $F_{k,n-p}$ distribution (see Theorem 1.7). So for an exact test of size α we reject H_0 at significance level $\alpha \iff F(y) > F_{k,n-p}(\alpha)$.

The p-value of this test is $P(F_{k,n-p} > F(y))$.

What is the intuition here?

The question which the test answers is: is there evidence that the H_1-model fits the data significantly better than the H_0-model? Tests which look for significant changes in the RSS, such as the F-test above, are called *analysis of variance* (ANOVA).

When we fit the more complex model H_1, there will be a reduction in the residual sum of squares compared to the residual sum of squares we get when we fit the simpler model H_0.

(i) The $\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)}$ in the numerator is the reduction in the residual sum of squares in moving from H_0 to H_1. This reduction is taken relative to the number of parameters we gain in moving from H_0 to H_1, i.e. it is divided by the additional number of degrees of freedom in H_1, which is k.

(ii) The denominator is the residual sum of squares $\mathrm{RSS}^{(1)}$ for the more general model H_1 relative to its number of degrees of freedom, i.e. divided by $n - p$ (see Theorem 1.4(iii)).

If (i) relative to (ii) (i.e. divided by) is large enough, then H_1 is significantly better and we reject H_0. The relevant distribution to compare to is $F_{k,n-p}$.

Theorem 1.7. *Under H_0 (at (1.14)),*

(i) $\dfrac{RSS^{(1)}}{\sigma^2} \sim \chi^2_{n-p}$ *independently of* $\dfrac{RSS^{(0)} - RSS^{(1)}}{\sigma^2} \sim \chi^2_k$

(ii) $F(y) \sim F_{k,n-p}$.

We have already shown part of (i): we saw $\frac{\mathrm{RSS}^{(1)}}{\sigma^2} \sim \chi^2_{n-p}$ in Theorem 1.4.

Proof. The proof is similar to that of Theorem 1.4. We need to show that (i) and (ii) are true under H_0, which means we are assuming H_0 is true in this proof.

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_{p-k}\}$ be an orthonormal basis for $\mathrm{col}(\widetilde{X})$, and extend it to an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ for $\mathrm{col}(X)$, and then extend this to an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{R}^n$.

As in Theorem 1.4 we can write $y = \sum_{i=1}^n z_i \mathbf{v}_i$ where $z_i = \mathbf{v}_i^T y$.

Since $H^{(0)}$ projects onto the space spanned by the first $p - k$ columns of X, we have $H^{(0)}\mathbf{v}_i = \mathbf{v}_i$ if $i \leqslant p - k$ and $H^{(0)}\mathbf{v}_i = 0$ if $i > p - k$. Hence

$$\widehat{y}^{(0)} = H^{(0)} y = z_1 \mathbf{v}_1 + \cdots + z_{p-k} \mathbf{v}_{p-k}.$$

As in Theorem 1.4,

$$\widehat{y}^{(1)} = Hy = z_1 \mathbf{v}_1 + \cdots + z_p \mathbf{v}_p.$$

So $y - \widehat{y}^{(1)} = z_{p+1}\mathbf{v}_{p+1} + \cdots + z_n \mathbf{v}_n$ and

$$\mathrm{RSS}^{(1)} = z_{p+1}^2 + \cdots + z_n^2$$

and similarly $\mathrm{RSS}^{(0)} = z_{p-k+1}^2 + \cdots + z_n^2$ and so

$$\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)} = z_{p-k+1}^2 + \cdots + z_p^2.$$

The weights $z_1, \ldots, z_n$ are independent (since they are normal with a diagonal covariance matrix, exactly as in Theorem 1.4). They are distributed as $z_i \sim N(\mathbf{v}_i^T E(y), \sigma^2)$ and $E(y) = \widetilde{X}\beta$ as H_0 is assumed true.

The expectation $E(z_i) = \mathbf{v}_i^T \widetilde{X}\beta = 0$ for $i > p-k$ since $\mathbf{v}_i$ is orthogonal to the columns of $\widetilde{X}$ for $i > p-k$. Hence $z_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $i = p-k+1, \ldots, n$ (i.e. not just for $i = p+1, \ldots, n$ as we had in Theorem 1.4).

So $\text{RSS}^{(1)}$ and $\text{RSS}^{(0)} - \text{RSS}^{(1)}$ are independent since they are functions of disjoint sets of the independent random variables z_i. Since there k terms in the $\text{RSS}^{(0)} - \text{RSS}^{(1)}$ sum of squares above, the χ^2_k property also follows. Part (ii) of the Theorem follows immediately from part (i) and the definition of an $F_{k,n-p}$ distributionn. □

Other tests

We often test for two parameters β_1 and β_2 to be equal, so $H_0 : \beta_1 - \beta_2 = 0$. The MLE of $\beta_1 - \beta_2$ is $\widehat{\beta}_1 - \widehat{\beta}_2$ with variance

$$\begin{aligned}\text{var}(\widehat{\beta}_1 - \widehat{\beta}_2) &= \text{var}(\widehat{\beta}_1) + \text{var}(\widehat{\beta}_2) - 2\,\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) \\ &= \sigma^2 (X^T X)^{-1}_{11} + \sigma^2 (X^T X)^{-1}_{22} - 2\sigma^2 (X^T X)^{-1}_{12}\end{aligned}$$

so the test statistic and its null distribution are

$$\frac{\widehat{\beta}_1 - \widehat{\beta}_2}{s\sqrt{(X^T X)^{-1}_{11} + (X^T X)^{-1}_{22} - 2(X^T X)^{-1}_{12}}} \sim t(n-p).$$

This approach also works for linear combinations of parameters. Suppose v is a known $p \times 1$ vector and we want to test $v^T\beta = 0$. Then the MLE of $v^T\beta$ is $v^T\widehat{\beta}$ and

$$\text{var}(v^T\widehat{\beta}) = v^T \,\text{var}(\widehat{\beta}) v = \sigma^2 v^T (X^T X)^{-1} v$$

so the test statistic and its null distribution are

$$\frac{v^T\widehat{\beta}}{s\sqrt{(v^T (X^T X)^{-1} v)}} \sim t(n-p).$$

The quantities s^2 and $v^T\widehat{\beta}$ are independent since s^2 and $\widehat{\beta}$ are independent.

There are some shortcuts. For example consider a test for $\beta_1 = \beta_2$. The reduced model, with $\beta_1' = \beta_1 = \beta_2$, is

$$y = \beta_1'(x_1 + x_2) + \beta_3 x_3 + \ldots$$

and the full model $y = X\beta + \epsilon$ can be written

$$y = \beta_1'(x_1 + x_2) + \beta_2'(x_1 - x_2) + \beta_3 x_3 + \ldots$$

so the test for $\beta_1 = \beta_2$ can be framed as a test for $\beta_2' = 0$. We can run a t- or F-test to drop β_2', with design matrix $X = [X_1 + X_2, X_1 - X_2, X_3, \ldots, X_p]$ and parameter vector $\beta = (\beta_1', \beta_2', \beta_3, \ldots, \beta_p)$.

1.5 ANOVA

Because ANOVA tests are used frequently, the important numbers in the tests are laid out in a standard way. The ANOVA table sets out the numbers we need to make tests for dropping certain collections of variables. Suppose the variables $x_1, \ldots, x_p$ come in $m = 3$ groups and are ordered so that the groups are

$$\{1\},\ \{2, 3, \ldots, p-k\},\ \{p-k+1, p-k+2, \ldots, p\}.$$

This splits off the variables x_{p-k+1} to x_p. This is the variable grouping relevant for the hypothesis $H_0 \colon \beta_{p-k+1} = \beta_{p-k+2} = \cdots = \beta_p = 0$, which we test with an F-test.

We assume that group $\{1\}$, the variable x_1, corresponds to an intercept.

A typical ANOVA table gives fitting information for a sequence of models starting from a simplest model with just intercept β_1, adding the groups one at a time, up to the model with all p variables. For the $(m = 3)$-group case, testing $H_0 \colon \beta_{p-k+1} = \beta_{p-k+2} = \cdots = \beta_p = 0$, the model sequence is

$$\begin{aligned} y &= \beta_1 + \epsilon \\ y &= \beta_1 + \beta_2 x_2 + \cdots + \beta_{p-k} x_{p-k} + \epsilon \\ y &= \beta_1 + \beta_2 x_2 + \cdots + \beta_{p-k} x_{p-k} + \beta_{p-k+1} x_{p-k+1} + \cdots + \beta_p x_p + \epsilon. \end{aligned}$$

If we order the variables in the right way, we can sometimes do model selection at a glance, as we read down the table from top to bottom. If $X_{1:i} = [X_1, X_2, \ldots, X_i]$, then the design matrices build from X_1 to $X = X_{1:p}$. Let $\mathrm{RSS}_{1:i}$ be the residual sum of squares for the fit with design matrix $X_{1:i}$. The decrease in the residual sum of squares when we add the variables $x_{i+1}, \ldots, x_{i+k}$ to a model that already has the variables $x_1, x_2, \ldots, x_i$ is $\mathrm{RSS}_{1:i} - \mathrm{RSS}_{1:(i+k)}$. The number of residual degrees of freedom in the fit for the model with design matrix $X_{1:i}$ is $n - i$ (assuming the columns of $X_{1:i}$ are linearly independent).

Terms added	Degrees of freedom	Reduction in RSS	Mean square	F-statistic
$X_{2:(p-k)}$	$p-k-1$	$\mathrm{TSS} - \mathrm{RSS}_{1:(p-k)}$	$\dfrac{\mathrm{TSS} - \mathrm{RSS}_{1:(p-k)}}{p-k-1}$	$\dfrac{(\mathrm{TSS} - \mathrm{RSS}_{1:(p-k)})/(p-k-1)}{\mathrm{RSS}_{1:p}/(n-p)}$
$X_{(p-k+1):p}$	k	$\mathrm{RSS}_{1:(p-k)} - \mathrm{RSS}_{1:p}$	$\dfrac{\mathrm{RSS}_{1:(p-k)} - \mathrm{RSS}_{1:p}}{k}$	$\dfrac{(\mathrm{RSS}_{1:(p-k)} - \mathrm{RSS}_{1:p})/k}{\mathrm{RSS}_{1:p}/(n-p)}$
Residual	$n-p$	$\mathrm{RSS}_{1:p}$	$\dfrac{\mathrm{RSS}_{1:p}}{n-p}$	

Table 1.1: ANOVA table for the groups of variables $\{x_2, \ldots, x_{p-k}\}$ and $\{x_{p-k+1}, \ldots, x_p\}$ added incrementally to the intercept group $\{x_1\}$. In some tables a final column giving the p-value is included.

The layout of an ANOVA table for the three groups $\{1\}$, $\{2, \ldots, p-k\}$, $\{p-k+1, \ldots, p\}$ is shown in Table 1.1. $\mathrm{TSS} = (y - \overline{y})^T (y - \overline{y})$ is the residual sum of squares for a model with just intercept, in other words, the total sum of squares adjusted for intercept. $\mathrm{RSS}_{1:p}$ is the residual sum of squares for the full model.

The second F-statistic in the table (in the $X_{(p-k+1):p}$ row) is the F-test statistic for the test to add the variables $\{x_{p-k+1}, \ldots, x_p\}$ to the model with variables $\{x_1, \ldots, x_{p-k}\}$ already included, which is the test we set up at (1.18).

The first F-statistic in the table (in the $X_{2:(p-k)}$ row) is an F-test statistic for the test to add the variables $\{x_2, \ldots, x_{p-k}\}$ to a model with just x_1, the intercept variable. It might seem natural to use the divisor $\mathrm{RSS}_{1:(p-k)}/(n-(p-k))$, for an F with $p-k-1$ numerator and $n-(p-k)$ denominator degrees of freedom. However:

(i) the divisor $\mathrm{RSS}_{1:p}/(n-p)$ is "just as good" as $\mathrm{RSS}_{1:(p-k)}/(n-(p-k))$, since it too is independent of $\mathrm{TSS}-\mathrm{RSS}_{1:(p-k)}$, so we can see $(\mathrm{TSS}-\mathrm{RSS}_{1:(p-k)})(n-p)/\mathrm{RSS}_{1:p)}(p-k-1)$ has an $F(p-k-1, n-p)$ distribution under the null, and

(ii) the divisor $\mathrm{RSS}_{1:p}/(n-p)$ is better as it is an estimate of σ^2 which is not biased if the variables $x_{p-k+1}, \ldots, x_p$ added in the row below turn out to have a significant explanatory effect

(iii) we might possibly add: the divisor $\mathrm{RSS}_{1:p}/(n-p)$ has a higher variance than $\mathrm{RSS}_{1:(p-k)}/(n-(p-k))$ if variables $x_{p-k+1}, \ldots, x_p$ really were not related to the response, so in that case we would do better to drop them from the ANOVA – this is equivalent to using the $\mathrm{RSS}_{1:(p-k)}/(n-(p-k))$ divisor.

Another way to make point (iii) is that the $\mathrm{RSS}_{1:(p-k)}/(n-(p-k))$ divisor is the one given by the likelihood ratio test, whereas $\mathrm{RSS}_{1:p)}/(n-p)$ is just some statistic with a distribution we happen to know under the null.

On balance, item (ii) controls our choice of test statistic, so the table opts for higher variance in return for lower bias.

Table 1.1 is the table we might set out if we were carrying out the F-test for $H_0 : \beta_{p-k+1} = \beta_p - k + 2 = \cdots = \beta_p = 0$ against $H_1 : \beta \in \mathbb{R}^p$, though we could omit the first row.

Testing for no linear dependence

Other relevant test statistics can be computed from the numbers in an ANOVA table like Table 1.1. For example, the test for the hypothesis H_0': $\beta_2 = \cdots = \beta_p = 0$ against the full model H_1: $\beta \in \mathbb{R}^p$ is the test for no linear relation to the variables $x_2, \ldots, x_p$. The test statistic is

$$F' = \frac{(\mathrm{TSS} - \mathrm{RSS}_{1:p})/(p-1)}{\mathrm{RSS}_{1:p}/(n-p)}$$

since $k = p - 1$ here.

The denominator is given at the bottom of the ANOVA table. The quantity $\mathrm{TSS}-\mathrm{RSS}_{1:p}$ is the sum of the terms in the "reduction in RSS" column,

$$\mathrm{TSS} - \mathrm{RSS}_{1:p} = (\mathrm{TSS} - \mathrm{RSS}_{1:(p-k)}) + (\mathrm{RSS}_{1:(p-k)} - \mathrm{RSS}_{1:p})$$

so we can form F' by taking appropriate sums and ratios of table elements.

We can think of F' as replacing the widely used statistic R^2 as a measure of fit quality (or the lack of it). R^2 runs from zero to one but we have no absolute scale for quality of fit (i.e. there is no scale to say how close to one R^2 needs to be for a good fit). Whereas F' runs from 0 to infinity (where large F' is poor fit) and *does* give a direct test for significant linear dependence. Note that R reports both F' (the test statistic for no linear relation) and its p-value as part of its standard summary of a linear model. This is more useful to us than R^2, though R gives this as well.

In the above notation R^2 is given by $R^2 = 1 - \text{RSS}_{1:p}/\text{TSS}$ and R^2 can be thought of as the proportion of variance (in y) that is explained by the model with p parameters.

ANOVA tables can be more general than above, we may have m groups of variables, say the groups are

$$\{i_1 = 1\},\ \{2, 3, \ldots, i_2\},\ \{i_2 + 1, i_2 + 2, \ldots, i_3\},\ \ldots,\ \{i_{m-1} + 1, i_{m-1} + 2, \ldots, i_m\}.$$

So i_k gives the index of the largest entry in the kth group. We typically insist that $i_1 = 1$ is a group of one, the intercept, and necessarily $i_m = p$. The simple case above is $m = 3$ with $i_2 = p - k$ so there are k variables in the last group. The quantity $\text{TSS} - \text{RSS}_{1:p}$, the total reduction in RSS in moving from the model with just an intercept to the model with all p explanatory variables, can be written

$$\begin{aligned}\text{TSS} - \text{RSS}_{1:p} = (\text{TSS} - \text{RSS}_{1:i_2}) + (\text{RSS}_{1:i_{i_2}} - \text{RSS}_{1:i_3}) + \ldots \\ + (\text{RSS}_{1:i_{m-2}} - \text{RSS}_{1:i_{m-1}}) + (\text{RSS}_{1:i_{m-1}} - \text{RSS}_{1:p}).\end{aligned}$$

where the bracketed terms on the RHS would be the entries in the "reduction in RSS" column of the ANOVA table.

EXAMPLE 1.8. Trees example here.

1.6 Prediction

Suppose we have fitted the model $y = X\beta + \epsilon$ and we are interested in predicting y at a new set of predictor variables $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})$.

The predicted value is $\widehat{y}_0 = \mathbf{x}_0^T \widehat{\beta}$. How do we assess the uncertainty in this prediction?

Predicting y at $\mathbf{x}_0$ could mean two things:

- prediction of the mean response – this means predicting the *expectation* of y at $\mathbf{x}_0$
- prediction of a future value – this means predicting the *value* of a future observation at $\mathbf{x}_0$.

(i) Confidence interval for the mean response: we have $\text{var}(\mathbf{x}_0^T \widehat{\beta}) = \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0 \sigma^2$ and so a $1 - \alpha$ confidence interval for the mean response is

$$\widehat{y}_0 \pm t_{n-p}(\tfrac{\alpha}{2}) s \sqrt{\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}.$$

(ii) Prediction interval for the value of a future observation: a future observation is predicted to be $\mathbf{x}_0^T \widehat{\beta} + \epsilon_0$. We do not know the future error ϵ_0, we assume $\epsilon_0 \sim N(0, \sigma^2)$ independent of $(\epsilon_1, \ldots, \epsilon_n)$. Then

$$\text{var}(\mathbf{x}_0^T \widehat{\beta} + \epsilon_0) = \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0 \sigma^2 + \sigma^2.$$

So a $1 - \alpha$ interval is

$$\widehat{y}_0 \pm t_{n-p}(\tfrac{\alpha}{2}) s \sqrt{1 + \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}.$$

We call this second interval a prediction interval (rather than a confidence interval) because it is an interval for a random variable (rather than a parameter).

The confidence interval above is clearly narrower than the prediction interval, often much narrower.

1.7 Categorical variables

So far our explanatory variables have been continuous variables. Explanatory variables that are qualitative, e.g. hair colour or eye colour, are *categorical variables* or *factors*. The values taken by a categorical variable are called its *levels*, and the levels may be ordered or unordered. We will discuss unordered categorical variables.

E.g. Eye colour $\in$ {Brown, Blue, Hazel, Green} would be a factor with 4 levels.

A categorical explanatory variable $x^{(\text{cat})}$ with c levels, say $x^{(\text{cat})} \in \{1, 2, \ldots, c\}$, is equivalent to c binary indicator variables, say $g_1, g_2, \ldots, g_c$:

- let $x_i^{(\text{cat})}$ be the value of the categorical variable for observation i
- and then define $g_{ir} = 1_{\{x_i^{(\text{cat})} = r\}}$ for $r = 1, 2, \ldots, c$.

So the ith response y_i has one explanatory variable g_{ir} for each level of the original categorical variable.

Suppose we want to allow the response y_i to have a mean which depends on the level of $x_i^{(\text{cat})}$, and suppose there are m other explanatory variables $x_{i1}, \ldots, x_{im}$ which include an intercept $x_{i1} \equiv 1$. The model

$$y_i = \alpha + \alpha_2 g_{i2} + \cdots + \alpha_c g_{ic} + \sum_{j=2}^{m} \beta_j x_{ij} + \epsilon_i \tag{1.19}$$

allows the mean of y_i to vary with the level of $x_i^{(\text{cat})}$. The parameter vector is $\beta = (\alpha, \alpha_2, \ldots, \alpha_c, \beta_2, \ldots, \beta_m)$.

What happened to the term $\alpha_1 g_{i1}$? If we include it then our model is over-parameterised. So we omit it.

In (1.19) if the level for response i is $x_i^{(\text{cat})} = 1$, then $g_{i1} = 1$ and $g_{i2} = \cdots = g_{ic} = 0$, so

$$E(y_i) = \alpha + \sum_{j=2}^{m} \beta_j x_{ij}.$$

Whereas if the level is $x_i^{(\text{cat})} = r$ where $r \geqslant 2$, then $g_{ir} = 1$ and the others are zero, so

$$E(y_i) = \alpha + \alpha_r + \sum_{j=2}^{m} \beta_j x_{ij}.$$

So we see that α_r is the offset in the intercept of the level-r samples relative to the intercept α of the level-1 samples. We are using level 1 as the *baseline* level.

If G_r is the binary column vector $G_r = (g_{1r}, \ldots, g_{nr})^T$ for the level-r indicator, and $X_1, \ldots, X_m$ are column vectors for other variables, then the design matrix for the model above is $X = (X_1, G_2, \ldots, G_c, X_2, \ldots, X_m)$. The dimension of β is $p = m + c - 1$. Including the categorical variable has added $c - 1$ additional parameters to the model (not c extra).

We left out G_1 when we formed the design matrix because X has a first column of ones, corresponding to the intercept. But then $X_1 = \sum_{r=1}^{c} G_r$ since each observation

must have its categorical variable in one of the levels $1, \ldots, c$, and so the columns of $(X_1, G_1, G_2, \ldots, G_c, X_2, \ldots, X_m)$ are not linearly independent and the model with all c columns $G_1, \ldots, G_c$ is over-parameterised.

The variables $G_1, \ldots, G_c$ are sometimes called *dummy variables* for the levels and the matrix $(G_2, \ldots, G_c)$ is called a *contrast* matrix.

EXAMPLE 1.9. Gas consumption example here.

1.8 Variable interactions

We can form new explanatory variables from old ones by taking functions of explanatory variables. We may want to do this because variables interact.

Interactions of the form $\beta_i x_i + \beta_j x_j + \beta_I x_i x_j$ are particularly common and mean something like "variable j has more impact on the response when variable i is large" (and vice versa). The parameter β_I is an interaction parameter.

For we would then have

$$E(y) = \cdots + \beta_i x_i + (\beta_j + \beta_I x_i) x_j + \ldots$$

and so, with x_j fixed, the impact of the term involving x_j is larger when x_i is large. And there is a corresponding interpretation with i and j swapped.

If x_i is a binary dummy variable for some level r of a categorical variable $x^{(\text{cat})}$, i.e. if $x_i = 1_{\{x^{(\text{cat})}=r\}}$, then

$$\begin{aligned} E(y) &= \cdots + \beta_i x_i + (\beta_j + \beta_I x_i) x_j + \ldots \\ &= \begin{cases} \cdots + \beta_j x_j + \ldots & \text{if } x^{(\text{cat})} \neq r, \text{ i.e. when } x_i = 0 \\ \cdots + \beta_i + (\beta_j + \beta_I) x_j & \text{if } x^{(\text{cat})} = r, \text{ i.e. when } x_i = 1. \end{cases} \end{aligned}$$

So the slope with respect to increasing x_j is β_j for observations with $x^{(\text{cat})} \neq r$ (where $x_i = 0$), and the slope is $\beta_j + \beta_I$ for observations with $x^{(\text{cat})} = r$ (where $x_i = 1$).

This change in slope is in addition to the additive offset effect of the categorical variable (an offset of 0 when $x_i = 0$, and of β_i when $x_i = 1$).

The *hierarchy principle* is that if an interaction is included in the model, then so are the corresponding lower order terms. That is, if the $x_i x_j$ interaction term is included then so are the individual main effect terms for x_i and x_j. We usually follow this principle.

For example, suppose $y = \alpha + \beta x_1 x_2 + \epsilon$ where x_1 is in degrees Celsius. If we switch from x_1 in Celsius to x_1' in Fahrenheit, where $x_1 = m x_1' + d$ with $m = 5/9$, $d = -160/9$, then we get $y = \alpha + d\beta x_2 + m\beta x_1' x_2 + \epsilon$. Now we have new kind of term: $d\beta x_2$. We may dislike the idea that the kinds of terms in our model (rather than just the parameter values) are dependent on the location of the zero of our measurement scale, and instead at least begin our modelling with $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_I x_1 x_2 + \epsilon$. Faraway (1st edition, 2005, p122; 2nd edition, 2015, p150) has more on this.

If x_1 and x_2 are continuous variables, then introducing an $x_1 x_2$ term adds just one parameter (β_I above).

If x_1 is a categorical variable with c levels and x_2 is continuous, then the interaction model would have: an intercept, plus $c - 1$ parameters for the main effects of x_1 (i.e. different

intercepts for the different levels of x_1), plus 1 parameter for the main effect of x_2, plus $c-1$ further parameters for interactions of x_1 and x_2 (i.e. different gradients for each level of x_1).

If x_1 is categorical with c levels and x_2 is categorical with d levels, then the interaction model would have: an intercept, plus $c-1$ parameters for main effects of x_1, plus $d-1$ parameters for main effects of x_2, plus $(c-1)(d-1)$ parameters for interactions of x_1 and x_2.

EXAMPLE 1.10. See gas consumption example again.

1.9 Blocks, Treatments and Designs

Chapters 14–16 of Faraway (1st edition, 2005; or chapters 15–17 of 2nd edition, 2015) covers this material at about the right level for us. See the discussion in Davison (2003) for more detail.

One important kind of categorical variable arises when the data have been gathered from b *blocks*, say y_{ij} for $i = 1, \ldots, b$ and $j = 1, \ldots, n_i$, where the group of subjects in a block (n_i subjects in block i) are expected to have similar response to the explanatory variables, but where there may be differences from one block to another.

Imagine collecting observations of the body mass index (BMI) of 80 five-year-old children from different schools. In one design we measure the BMI of eight randomly selected five-year-olds in each of 10 schools. For each child we record the BMI and the school. In another design, we might do exactly the same, but fail to record the school. The first data set has a block structure, with 10 blocks. The block index is often explanatory. If, for example, there is a correlation between parent income-level, school and incidence of obesity, then 'school' will be explanatory for 'BMI'. In such cases we code the block index i as a categorical explanatory variable in the design matrix.

Besides taking subjects from distinct groups, and distinguishing group responses, we may also give subjects different treatments, and distinguish responses to different treatments. If there is a block structure to the population, with subjects in different blocks having different treatment responses, and we ignore it, then this will tend to inflate the estimated error variance s^2, and real differences in the treatment response may not be detected.

"Treatment factors are those for which we wish to determine if there is an effect. Blocking factors are those for which we believe there is an effect. We wish to prevent a presumed blocking effect from interfering with our measurement of the treatment effect."
(Heiberger and Holland, *Statistical Analysis and Data Display*, Springer, 2004.)

EXAMPLE 1.11. Pigs diet example here.

When we set up a design, we may have some choice in the assignment of treatments to subjects. The point of the trial is to see if the response depends on the treatment. Suppose the 'treatment' levels are 'Give drug' and 'Give placebo' and the response is some index of health. If we are allowed to choose which subject gets which treatment, we could distort the trial outcome by, for example, choosing to give the drug to subjects who for some reason are more likely to get better anyway. Wonderdrug! In order to avoid all traps of this kind, we typically assign the treatments to subjects completely at random. A design with everyone in one block, and treatments assigned to subjects at random, is called *completely randomised*. If the subjects are in blocks, with m_a subjects in block a, and we

apply treatment $t = 1, \ldots, T$ to $m_{a,t}$ different subjects in block $a = 1, \ldots, b$, then, for each treatment, we choose $m_{a,t}$ subjects independently at random and without replacement from the m_a subjects in the a'th block. If $m_{a,t} = m$, so that each treatment is applied to the same number of subjects in each block, then the design is called a *randomised complete block design*. The treatments are distributed in a balanced way through the blocks.

The piglets data is balanced, as each of the four litters contains one piglet on each of the three diets, so $b = 4, T = 3$ and $m_{a,t} = 1$ for all $a = 1, \ldots, 4$ and $t = 1, 2, 3$. A *randomised complete block design* is balanced in such a way that the block and treatment parameter estimates are independent, and so the treatments can be analyzed separately from blocks. Informally, in a completely balanced design the treatments are all tested under the same conditions, so when we compare treatments, it doesn't matter what those conditions were.

Sometimes the number of subjects m_a in a block is smaller than T the number of treatments. In that case the design will be *incomplete*, since some blocks will have no instances of some treatments. An incomplete block design may still be balanced.

2 Model Checking and Model Selection

2.1 Model checking

We are fitting a normal linear model $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$. Here X is an $n \times p$ design matrix and X has one column $X_i = (x_{1i}, x_{2i}, \ldots, x_{ni})^T$ for each of the $i = 1, 2, \ldots, p$ explanatory variables.

Model violations take many forms. A misspecified model is structurally inappropriate for essentially all responses. For example an important explanatory variable (or more than one) may be missing from the current model. Or perhaps the response y may not be a linear function of the linear predictors $X\beta$ and we may need to transform the response. We may choose to work with a misspecified model if we have reason to believe that the biases in fitted values or parameter estimates are not large enough to invalidate the inference, for parameters of interest.

Some possible problems with the model:

- the errors ϵ do not have constant variance
- the errors ϵ are not normal
- the errors ϵ are correlated with each other
- the response y is not a linear function of $x_1, \ldots, x_p$.

On the other hand the problem may lie with the data. The normal linear framework may be good for most data points, but a few of the responses may have quite different causative factors. Such data points are called *outliers*. We try to identify them and then typically remove them from further analysis.

We have a range of validatory checks for model misspecification and outlier detection. The most straightforward check for linearity is to plot the response against each of the explanatory variables in turn. However variation in the response caused by variation in other variables may obscure the linear response to any single variable. We also have checks on the independence and constant variance of ϵ, the errors: we saw that the fitted values $\widehat{y} = Hy$ and the vector of residuals $e = y - \widehat{y}$ are independent under the model, so a plot of residuals against fitted values should show no correlation.

Misfit

One weakness of the residuals v. fitted-values plot (which we have already used) as a diagnostic tool is that the residuals may have unequal variance under the model. A large residual e_i could be a sign that the ith data point is an outlier, but it might have a large variance as a consequence of the experimental design.

The vector of residuals $e = y - \widehat{y} = (I - H)y$ is normal (as linear combinations are normal), with mean

$$\begin{aligned} E(e) &= (I - H)E(y) \\ &= (I - H)X\beta \\ &= 0 \qquad \text{since } HX = X \end{aligned}$$

and covariance matrix

$$\begin{aligned}\operatorname{var}(e) &= (I-H)\operatorname{var}(y)(I-H)^T \\ &= \sigma^2(I-H)(I-H)^T \\ &= \sigma^2(I-H) \qquad \text{since } H^2 = H = H^T.\end{aligned}$$

Hence $\operatorname{var}(e_i) = \sigma^2(1-h_{ii})$ where h_{ii} is the (i,i)-entry of H, showing that the residual variances of the e_i can be unequal.

Similarly the variances of the $\widehat{y}_i$ are unequal: $\operatorname{var}(\widehat{y}) = \sigma^2 H$, so $\operatorname{var}(\widehat{y}_i) = \sigma^2 h_{ii}$.

As $\operatorname{var}(e_i) \geqslant 0$ and $\operatorname{var}(\widehat{y}_i) \geqslant 0$, we have $0 \leqslant h_{ii} \leqslant 1$.

Standardised residuals

We can adjust the e_i to account for unequal variances. The *standardised residuals* $r = (r_1, \ldots, r_n)$ are defined by

$$r_i = \frac{e_k}{s\sqrt{1-h_{ii}}}$$

where s^2 is the usual unbiased estimate of σ^2. The r_i have approximately unit variance and can be compared with standard normal variables, an approximation which will be good when large $n-p$ is large. Because s and e are correlated, we cannot easily compute the distribution of the standardised residuals.

A value of $|r_i| > 2$ is possible misfit. That is, assuming the model is true, we expect about 95% of the standardised residuals to be in the interval $(-2, 2)$. Note: we do not expect *all* of the r_i to be in $(-2, 2)$.

Exercise 2.1. Show that the standardised residuals are (like the residuals e) independent of $\widehat{y}$.

We can make normal qqplots of standardised residuals, and we can plot them (the r_i) against fitted values $\widehat{y}_i$. We often mistakenly fit a model of constant variance to data in which the variance of the response increases with the underlying mean. This model misspecification is shown by a trend of increasing variance in r as $\widehat{y}$ increases.

Studentised residuals

A response y_i which generates a relatively large residual e_i need not be an outlier, since $\operatorname{var}(e_i) = \sigma^2(1-h_{ii})$, and $1-h_{ii}$ may be relatively large. We might expect the standardised residuals r to be a good basis for outlier detection, since these should have variance about one under the normal linear model. Standardised residuals exceeding two (standard deviations) are large. The problem here is that $s^2 = e^T e/(n-p)$, the denominator in the expression for r_i, is computed from the residuals e themselves. A response y_i, which is truly outlying, may have a large residual e_i, but this will inflate our estimate s^2 for σ, and we may end up with a moderate standardised residual r_i. A bad response with an h_{ii} value close to one has a low variance. It 'pulls' the fitted surface towards itself, so that e_i is small, and again r_i is not obviously large. What to do?

We can treat this problem using an idea related to 'cross-validation'. We remove response y_i and row i, given by $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, from the data and compare the fitted values $\widehat{y}$ we got from all of the data with the fitted values $\widehat{y}_{-i}$ we get when $\mathbf{x}_i^T, y_i$ are removed.

Denote by y_{-i} and X_{-i} the remaining response and design data with y_i and $\mathbf{x}_i$ removed. Let $\widehat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i}$ give the new parameter estimates. The ith *studentised residual* (or *deletion residual*) is defined by

$$r'_i = \frac{y_i - \mathbf{x}_i^T \widehat{\beta}_{-i}}{\text{std.err}(y_i - \mathbf{x}_i^T \widehat{\beta}_{-i})}. \tag{2.1}$$

Note: the estimated standard error in (2.1) is calculated as follows (see Sheet 2 for further details). The standard error is the square root of $\text{var}(y_i - \mathbf{x}_i^T \widehat{\beta}_{-i})$, and this square root is a multiple of σ. In the estimated standard error we replace the σ by an estimate of σ: since we are considering the analysis without observation i, the appropriate estimate is s_{-i} where this denotes the value of s calculated for the data with $\mathbf{x}_i, y_i$ removed.

It may be shown (see Sheet 2) that $r'_i \sim t(n-p-1)$, and that r' and $\widehat{y}$ are independent. So the studentised residuals have equal variance, known distribution, and can be compared to a standard normal (using for example a qqplot). We can plot r' against $\widehat{y}$. Any visible correlation is a sign of model misspecification, and data points with $|r'_i| > 2$ show misfit and are possible outliers.

The studentised residuals r'_i are related to the standardised residuals r_i by

$$r'_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}. \tag{2.2}$$

This formula is useful computationally, since it shows that we can compute the studentised residuals without making n linear regressions for the n deletions, instead we can just using the results of the primary regression.

We state the result (2.2) without proof: deriving it involves multiple pages of algebra and I don't think those calculations particularly help with understanding. For some details of the calculations, see e.g. Sections 2.2 and 3.1 of Atkinson, *Plots, Transformations, and Regression*, Oxford, 1985.

Leverage

The diagonal entries h_{ii} of the hat matrix H are called the *leverage* components.

Recall that $0 \leqslant h_{ii} \leqslant 1$. Since $\text{var}(e_i) = \sigma^2(1 - h_{ii})$, a point with leverage h_{ii} close to one has low variance: since $E(e_i) = 0$, the fitted surface $\mathbf{x}^T \widehat{\beta}$ must be pulled close to the ith response y_i. If $(\mathbf{x}_i, y_i)$ is an outlier with high leverage, then predictions $\mathbf{x}^T \widehat{\beta}$ for $\mathbf{x}$ near $\mathbf{x}_i$ will be poor.

How big is big, when it comes to leverage? The 'average' leverage for a $n \times p$ design is $\overline{h} = p/n$, as we will see shortly, so points with leverage values above $2p/n$ get special attention. Since they are having more impact on the final fit than other points, it is important that they are not outliers.

The sum of leverages is given by the trace of the hat matrix H:

$$\begin{aligned} \sum_{i=1}^{n} h_{ii} &= \text{trace}(H) \\ &= \text{trace}(X(X^T X)^{-1} X^T) \\ &= \text{trace}(X^T X (X^T X)^{-1}) \end{aligned}$$

using the cyclic permutation property of trace, $\text{trace}(ABC) = \text{trace}(CAB)$. It follows that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} h_{ii} &= \frac{1}{n}\,\text{trace}(I_p)\\ &= \frac{p}{n}.\end{aligned}$$

Influence

A point with high leverage need not do much damage if it lies close to the fitted surface through other points – the surface would have gone close to it anyway. Such a point is said to have high leverage but low *influence*. Highly influential points shift the fitted surface far from where it would have lain if the influential point were not included.

The *Cook's distance* for a point $(\mathbf{x}_i, y_i)$ is a measure of influence given by the sum of the squares of the shift in fitted values when point i is removed. The Cook's distance for the ith data point is defined to be

$$C_i = \frac{(\widehat{y} - \widehat{y}_{-i})^T(\widehat{y} - \widehat{y}_{-i})}{ps^2}$$

where $\widehat{y}_{-i} = X\widehat{\beta}_{-i}$. (That is, $\widehat{y}_{-i}$ is an n-component vector of fitted values, where the fit is based on all data points except point i, so $(\widehat{y}_{-i})_k = \mathbf{x}_k^T\widehat{\beta}_{-i}$ for $k = 1, \ldots, n$).

A large value of C_i indicates an observation which has high influence.

It may be shown (see e.g. Atkinson (1985)), that

$$C_i = \frac{r_i^2 h_{ii}}{p(1 - h_{ii})}. \tag{2.3}$$

Our intuition is that high influence occurs where there is misfit and large leverage. The factor $h_{ii}/(1 - h_{ii})$ rises with increasing leverage. The factor r_i^2 (standardised residual squared) is related to misfit. A rough rule of thumb is that $|r_i| > 2$ and $h_{ii} > 2p/n$ are separate causes for concern, so (plugging 2 and $2p/n$ into (2.3)) points with Cook's distance exceeding

$$C_k \gtrsim \frac{8}{n - 2p}$$

have high influence.

EXAMPLES 2.2. Model checking examples here.

EXAMPLE 2.3. `swiss` example here.

2.2 Model selection

In some settings we aim to filter out from some original, possibly large, set of explanatory variables the ones that are explanatory for the response. Where variables are correlated, this may not be possible – which of a group of near linearly dependent variables do we drop? Physical considerations are very important. Does the model make sense? We have to go some way down the road to understanding the application domain in order to make sense of the variables and their interactions, and make sensible selections of subsets of variables.

Model choice v. Exploratory data analysis

The easiest case is where we go into the analysis with some preconceived hypothesis. The trees example, Example 1.8, worked that way. We modelled the tree as something like a cylinder, and that gave us a clear hypothesis about the relation between volume, girth and height. We fit the model and test the hypothesis.

More commonly we look at the parameters in the fitted model, observe that some are not significant, fit a range of reduced models, and formulate some hypothesis about the relations between variables. This leads to a test for the significance of some subset of explanatory variables which is informed by the data. We have in effect made many tests, but report just the final one. This introduces a hazard for multiple testing, which can sometimes be corrected. If we don't correct, and we hardly every do, then paraphrasing Davison (2003, Section 8.7, in the subsection 'Inference after model selection'): "...the only covariates for which subsequent inference using the standard confidence intervals is reliable are those for which the evidence for inclusion is overwhelming". It sometimes happens that we frame a hypothesis by looking at the data, but then realise that the hypothesis has some natural physical meaning. When these notes were first written, the cylinder model in the trees example (Example 1.8) was not the first model tried for that data, but emerged in the analysis. However, the final model is so natural, that we could easily imagine a scientist going into the analysis with precisely this hypothesis (we just did).

Sometimes we use *automatic* variable selection. This means, essentially, any model selection procedure not guided by physical considerations for variable meaning. We might search over all subsets of variables for the 'best' reduced model. One choice is to look for the largest fully significant model. No variable, or set of variables, can be dropped, and if we add variables we have a model with some non-significant variables. There may be many such models. When the model space is very large we may search using stepwise methods. In *Backwards elimination* we start with the full set of variables, and successively drop the least significant, until we have a fully significant set. There is a *Forwards selection* scheme, which adds the next most significant variable. We sometimes sort the variables in an ANOVA table so that the p-values on the RHS of the table show in effect the steps of forward selection. One advantage of backwards selection is that the initial estimator for s^2 is from the fit for all the variables, so it is not biased (upwards) by significant variables which are not included. If forwards selection started with an estimate of s^2 based on just the one or two variables in the initial model, then it might be a large overestimate, since variation in the response y due to variations in significant explanatory variables would inflate s^2, and levels of significance would suffer accordingly. This is the reason we use the same $\mathrm{RSS}_{1:p}/(n-p)$ divisor in every row of the F-coloumn of an ANOVA, rather than $\mathrm{RSS}_{1:i_k}/(n-(i_k-i_{k-1}))$.

Automatic methods are exposed to the hazard for multiple testing mentioned above. If we make alot of tests on a large number of non-significant variables we may find marginally significant variables where there are none. The approach is justified as part of exploratory data analysis. The hope is that the method will turn up some physically natural set of explanatory variables.

AIC

One strategy is to search over all (or at least a wide range of) models and optimise some measure of the relative worth of models. The measure must somehow penalise models

which are too complex or too simple, since both are poor for prediction. The AIC is just such a model choice criterion.

Consider what happens when new data y' come along. We have parameter MLEs $\widehat{\beta}(y)$ and $\widehat{\sigma}^2(y)$ computed using the old data y and some particular design X. If the model is a good model, then the log-likelihood $\ell(\widehat{\beta}(y), \widehat{\sigma}^2(y); y')$ for $\widehat{\beta}$ and $\widehat{\sigma}^2$ in the new data should be large. This should "usually" hold, i.e. the expectation over y and y' of $\ell(\widehat{\beta}(y), \widehat{\sigma}^2(y); y')$ should be large. So if

$$C(y, y') = -2\ell(\widehat{\beta}(y), \widehat{\sigma}^2(y); y')$$

then we like models that make $E[C(y, y')]$ small. Here the expectation $E[C]$ is over both the old y and the new y' data. This leads to the AIC criterion; AIC is a consistent estimator of $E[C]$ (asymptotically, and after unimportant constants are ignored).

AIC = the *Akaike information criterion*, or "an information criterion".

AIC is defined by

$$\text{AIC} = -2\ell(\widehat{\beta}, \widehat{\sigma}^2) + 2p. \tag{2.4}$$

That is, minus twice the maximised log-likelihood plus $2p$, where p is the dimension of $\beta = (\beta_1, \ldots, \beta_p)$.

To use AIC we choose the model which *minimises* the AIC.

The AIC definition of "minus twice the maximised log-likelihood plus twice the number of parameters" generalises to other settings. In our case of linear models, equation (1.4) gives $-2\ell(\widehat{\beta}, \widehat{\sigma}^2) = n \log\left(\frac{\text{RSS}}{n}\right) + n$, so we can use

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + 2p$$

because we can drop the additive constant of n as our data y are fixed, hence n is fixed when we compare different models, i.e. if n is added to all AIC values, it would still be the same model which minimised AIC. The "number of parameters" of our model is really $p + 1$, i.e. we have parameters $\beta_1, \ldots, \beta_p, \sigma^2$, but again it doesn't matter if we ignore the "+1" as this would add a constant (of 2) to all AIC values, hence we could ignore the "+1" when defining AIC at (2.4). We must include variance and covariance/correlation parameters when counting parameters and calculating AIC in Section 3.

The first term in AIC involves RSS and hence is made smaller by adding more explanatory variables. However, adding more explanatory variables increases second *penalty term* of $2p$, hence this penalty term discourages us from adding too many explanatory variables. So there is a trade-off, AIC provides a balance between fit and simplicity.

There are other criteria, e.g. the *Bayes information criterion* (BIC) given by

$$\text{BIC} = -2\ell(\widehat{\beta}, \widehat{\sigma}^2) + p \log n$$

Clearly BIC penalises larger models more, so it will tend to prefer smaller models than AIC.

Faraway (2015): "There is some debate regarding the relative merits of these two criteria although AIC is generally considered better when prediction is the aim".

Where does the expression for AIC come from? See Geoff Nicholls MT 2014 notes, or Davison (2003, Section 8.7.3).

Final remarks on AIC:

- AIC balances model complexity against model fit.
- AIC compares on basis of prediction success, it tends to keep variables that other criteria drop.
- The models we compare *need not be nested*.

EXAMPLE 2.4. AIC example here.

2.3 Two model revision strategies

The following two subsections treat a couple of discrete topics concerning model revision. Suppose that, in the course of our diagnostic analysis we find that the errors are non-normal, or correlated. It may be possible to make a linear transformation of the model to get i.i.d. normal mean zero errors ϵ. We transform the data, fit, and then invert the transformation to get results for the original model. This is weighted regression, which we look at first.

We may find that the response y is not linearly related to the linear predictor $X\beta$. Can we find a transform which restores linearity? It would be convenient to take a family of transformations and then choose the member of that family which best restores linearity. This is the Box-Cox approach, which we look at (briefly) second.

Weighted regression

Suppose our data follow a normal linear model except that the variance varies from observation to observation. So if σ_i^2 is the variance for the ith observation, then

$$y_i = \mathbf{x}_i^T\beta + \epsilon_i, \quad i = 1, \ldots, n$$

as before, but now the independent errors ϵ_i are distributed $\epsilon_i \sim N(0, \sigma_i^2)$. There are three cases to consider: σ_i^2 unequal and unknown; σ_i^2 unequal and known; and $\sigma_i^2 = \sigma^2/w_i$, with w_i known but σ^2 unknown.

The first case cannot be treated without some assumption on the joint distribution of the σ_i since we have a variance parameter for each data point, which we cannot estimate. The second case can be treated in the same framework as the third (see the Problem Sheet 1 question where σ^2 is known).

What is the motivation for considering the third case? If y_i is actually the outcome of n_i independent measurements say $y_{ij} \sim N(\mathbf{x}_i^T\beta, \sigma^2)$ for $j = 1, \ldots, n_i$, then we may pool the data so that $y_i = \frac{1}{n_i}\sum_j y_{ij}$. Then $\text{var}(y_i) = \sigma^2/n_i$. This may also work for the more general case where y_{ij} are i.i.d. for $j = 1, \ldots, n_i$, with mean $\mathbf{x}_i^T\beta$ and variance $\text{var}(y_{ij}) = \sigma^2$, but with the y_{ij} not normal. By the CLT, the approximation $y_i \sim N(\mathbf{x}_i^T\beta, \sigma^2/n_i)$ may be good (if the n_i values are reasonably large).

We can map the weighted variance problem onto our original problem. Let $W = \text{diag}(w_1, \ldots, w_n)$, i.e. W is an $n \times n$ diagonal matrix with the w's on the diagonal. We define 'data' $y' = W^{1/2}y$ and a 'design' $X' = W^{1/2}X$. Then

$$y' = X'\beta + \epsilon'$$

where $\epsilon' = W^{1/2}\epsilon \sim N(0, \sigma^2 I_n)$ and we are back to our standard problem. The weighted least squares estimators (i.e. the MLEs) for β and σ^2 take the usual form when written in terms of X' and y'. Writing those estimators in terms of X and y gives $\widehat{\beta} = (X^TWX)^{-1}X^TWy$ and $s^2 = (y - X\widehat{\beta})^TW(y - X\widehat{\beta})/(n - p)$.

EXAMPLE 2.5. Weighted regression example here.

The Box-Cox family of transformations

Suppose a normal linear model applies not to y, but to some power of y, say to y^λ. We can use the Box-Cox method to find the best value of λ.

So suppose that a normal linear model applies not to y, but to $y^{(\lambda)}$ where

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0. \end{cases}$$

Note the above is consistent because $\lim_{\lambda \to 0} \left(\frac{y^\lambda - 1}{\lambda} \right) = \log y$. This consistency is the reason for using $y^{(\lambda)}$ rather than y^λ to obtain the best λ. Once the best λ has been found, we can use y^λ as the transformed response (rather than $(y^\lambda - 1)/\lambda$).

As λ varies in the range $(-2, 2)$ we get the inverse transformation ($\lambda = -1$), log ($\lambda = 0$), square and cube roots ($\lambda = \frac{1}{2}, \frac{1}{3}$), the original scale ($\lambda = 1$), as well as the squared case ($\lambda = 2$). Where possible we might hope for an interpretable value of λ. Faraway (2015): "If explaining the model is important, you should round λ to the nearest interpretable value."

The above assumes that $y_i > 0$ for all i. If not, the transformation must be applied to $y_i + \alpha$, with α chosen large enough so that all $y_i + \alpha$ are positive.

The Box Cox method treats λ as another parameter and finds the MLE and a confidence interval for λ.

Having computed $\widehat{\lambda}$ and a confidence interval for λ we usually fix on the nearest readily interpreted λ-value in the interval. We then recompute the fit (for $\widehat{\beta}$ and s^2 etc) conditioned on this estimate, i.e. having chosen $\lambda = \lambda^*$ we then take y^{λ^*} as our transformed response.

See Davison (2003), Section 8.6.2 for some details of the calculations.

EXAMPLE 2.6. Box-Cox example here.

3 Normal Linear Mixed Models

3.1 Hierarchical models

Hierarchical models or *mixed-effects models* (or *multilevel models*) generalise normal linear models and GLMs.

Here we consider the extension of normal linear models.

Suppose our data come in J groups, where group j contains n_j observations. So the total number of observations n is given by $n = \sum_{j=1}^{J} n_j$. Let y_{ij} be the response for the ith observation in group j, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$.

Suppose we have one covariate x, i.e. one explanatory variable. Let x_{ij} be the value of this covariate for the ith observation in group j.

Then a simple normal linear model is

$$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}$$
$$\text{where } \epsilon_{ij} \sim N(0, \sigma_y^2) \text{ are independent.}$$

Our first generalization allows the intercept α to vary randomly between groups. We assume that

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij} \tag{3.1}$$
$$\text{where } \epsilon_{ij} \sim N(0, \sigma_y^2) \tag{3.2}$$
$$\text{and } \alpha_j \sim N(\alpha, \sigma_\alpha^2). \tag{3.3}$$

Note that in (3.1)–(3.3) the α_j as well as the ϵ_{ij} are random variables. We assume all of the α_j and ϵ_{ij} are independent, for $i = 1, \ldots, n_j$ and $j = 1 \ldots J$. The unknown parameters are $\alpha, \beta, \sigma_y^2, \sigma_\alpha^2$.

Example 3.1. Gelman and Hill (2007) give radon measurements for US homes. Following that text we restrict to the $n = 919$ measurements from Minnesota. The measurement we work with as y is actually the log of the radon level.

The measurements come in groups, where each county in Minnesota is a group. There are $J = 85$ counties/groups. Denote by y_{ij} the response in the ith home in the jth county.

Measurements y_{ij} can be taken in the basement or on the ground floor. Covariate $x_{ij} \in \{0, 1\}$ records where the measurement was taken ($x = 0$ for basement, $x = 1$ for ground floor).

What is the baseline (log) radon level in each county?

The background radon levels vary from county to county, depending on local geology. It is natural to suppose that this background determines a baseline level α_j and each house varies around that local mean. Baseline levels are not unrelated – they are assumed to vary around the Minnesota-mean level α.

This leads us to the hierarchical model given above, which has a random effect on the intercept.

[There is another covariate u_j giving the (log of the) soil uranium yield in county j. We will ignore this for the moment.]

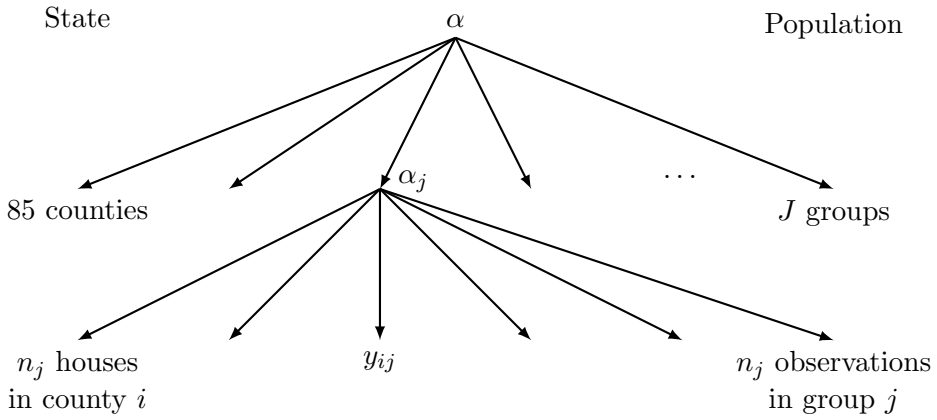


Figure 3.1: From the population/state, whose mean level is α, we draw α_j which is the mean level for group/county j. Then for group/county j we draw n_j observations, which we denote by y_{ij} for $i = 1, \ldots, n_j$.

Alternative notation

There are other ways to write the above model. For example, we can stack the measurements in a big column vector y with n entries:

$$(y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2}, \quad \ldots, \quad y_{n_J J})^T.$$

Say $y = (y_k)$ where k runs from 1 to n. So the ith observation from group j corresponds to observation $k = i + \sum_{r=1}^{j-1} n_r$.

Here we also want a categorical group variable $j_k \in \{1, 2, ..., J\}$, where j_k is the index of the group in which observation k falls.

In this indexing

$$\begin{aligned} &y_k = \alpha_{j_k} + \beta x_k + \epsilon_k, \quad k = 1, \ldots, n \\ &\text{where } \epsilon_k \sim N(0, \sigma_y^2), \quad k = 1, \ldots, n \\ &\text{and } \alpha_j \sim N(\alpha, \sigma_\alpha^2), \quad j = 1, \ldots, J. \end{aligned}$$

Apart from the last line (i.e. the normality assumption for the α_j), this is just the same as treating the group variable as a categorical variable with J levels and using indicator variables because

$$\alpha_{j_k} = \sum_{r=1}^{J} \alpha_r 1_{\{j_k = r\}}.$$

The novelty here is that the effects α_j are treated as random variables: they are called *random effects*.

R works with zero-mean offsets so we can rewrite a bit further. For each j we can write the random effect $\alpha_j \sim N(\alpha, \sigma_\alpha^2)$ as

$$\begin{aligned} &\alpha_j = \alpha + b_j \\ &\text{where } b_j \sim N(0, \sigma_\alpha^2). \end{aligned} \tag{3.4}$$

That is, on the RHS of (3.4) the α is a fixed parameter and the b_j is a normally distributed random effect with mean zero.

So we have

$$y_k = \alpha + b_{j_k} + \beta x_k + \epsilon_k, \quad k = 1, \ldots, n \tag{3.5}$$

$$\epsilon_k \sim N(0, \sigma_y^2), \quad k = 1, \ldots, n \tag{3.6}$$

$$b_j \sim N(0, \sigma_\alpha^2), \quad j = 1, \ldots, J. \tag{3.7}$$

Here α and β are fixed unknown parameters, and σ_y^2 and σ_α^2 are also fixed unknown parameters. The quantities $b_1, \ldots, b_J$ are random effects. The model is a *mixed-effects model* because it involves both fixed effects (α, β) and random effects $(b_1, \ldots, b_J)$. It is also called a *hierarchical model* (HM).

We can actually write the model without any random effects, by integrating them out. In the k-indexing the model above has covariances

$$\operatorname{cov}(y_k, y_{k'}) = \begin{cases} \sigma_y^2 + \sigma_\alpha^2 & \text{if } k = k' \\ \sigma_\alpha^2 & \text{if } k \neq k' \text{ and } j_k = j_{k'} \text{ (i.e. } k \neq k' \text{ are in same group)} \\ 0 & j_k \neq j_{k'} \text{ (i.e. } k \text{ and } k' \text{ are in different groups).} \end{cases} \tag{3.8}$$

Let Σ be the $n \times n$ matrix with entries $\Sigma_{k,k'} = \operatorname{cov}(y_k, y_{k'})$ as above. Let

$$\eta_k = \alpha + \beta x_k, \quad k = 1, \ldots, n. \tag{3.9}$$

Since everything is normal, we have the normal linear model

$$y \sim N(\eta, \Sigma). \tag{3.10}$$

We are not assuming we know σ_y^2 or σ_α^2. This form makes it clearer that although we introduced n random effects, the predictive model for y still has just $p = 2$ explanatory variables and two unknown variances. From (3.8)–(3.10) the likelihood can be found, hence it can be maximised and estimates obtained. We don't attempt the calculations, we leave those to R.

If we ignore the group structure then we ignore the correlation in the errors – observe the covariance matrix at (3.8) is *not* diagonal meaning that there is correlation in the errors. Correlation reduces the effective sample size, yielding falsely small estimates of variance if we ignore it. This leads to falsely inflated significance levels (falsely small p-values).

3.2 HMs interpolate pooled and unpooled models

In order to make transparent the consequence of having random effects, consider two extreme, simple, models:

(i) one with a single common mean (the "pooled" model), and

(ii) one with a separate mean for each group (the "un-pooled" model).

Let $\epsilon_k \sim N(0, \sigma_y^2)$ throughout, for $k = 1, \ldots, n$.

(i) Pooled model: $y_k = \alpha + \epsilon_k$. The MLE is $\widehat{\alpha} = \overline{y}$.

(ii) Un-pooled model: $y_k = \alpha_{j_k} + \epsilon_k$. The MLEs are $\widehat{\alpha}_j = \overline{y}_j$ (where $\overline{y}_j$ means the mean of the observations in group j, so $\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$).

The corresponding HM is:

$$\begin{aligned} &y_{ij} = \alpha_j + \epsilon_{ij} \\ &\text{where } \epsilon_{ij} \sim N(0, \sigma_y^2) \\ &\text{and } \alpha_j \sim N(\alpha, \sigma_\alpha^2). \end{aligned}$$

This is the same model as discussed in Section 3.1 except here there is no explanatory variable x. An equivalent way to write this $y_{ij} \sim N(\alpha_j, \sigma_y^2)$ where $\alpha_j \sim N(\alpha, \sigma_\alpha^2)$. So we have a normal likelihood for the y_{ij}, and a normal prior on the α_j. Hence we can find the joint posterior for $(\alpha_1, \ldots, \alpha_J)$.

The maximum of the posterior is at $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_J)$ where

$$\widehat{\alpha}_j = \frac{\frac{n_j \overline{y}_j}{\sigma_y^2} + \frac{\alpha}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}. \tag{3.11}$$

Once the values of $\alpha, \sigma_y^2, \sigma_\alpha^2$ have been estimated, equation (3.11) gives a value for $\widehat{\alpha}_j$. Since the α_j are random variables (not fixed parameters), the $\widehat{\alpha}_j$ values are usually called "predictions" rather than estimates.

Exercise 3.2. Show (by completing the square) that the joint posterior for $(\alpha_1, \ldots, \alpha_J)$ is normal and maximised at (3.11).

Assuming the estimate of α is close to $\overline{y}$, equation (3.11) gives

$$\widehat{\alpha}_j \approx \frac{\frac{n_j \overline{y}_j}{\sigma_y^2} + \frac{\overline{y}}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \tag{3.12}$$

$$= \overline{y}_j \; - \; \frac{\left(\overline{y}_j - \overline{y}\right)}{\left(1 + \frac{n_j \sigma_\alpha^2}{\sigma_y^2}\right)}. \tag{3.13}$$

The RHS of (3.12) is a weighted average of the pooled estimate $\overline{y}$ and the unpooled estimate $\overline{y}_j$. So the HM model is between the two extremes of the pooled and unpooled models:

- when $\sigma_\alpha^2 = 0$ we have the pooled model and $\widehat{\alpha}_j = \overline{y}$
- for general σ_α^2 we have a general $\widehat{\alpha}_j$
- as $\sigma_\alpha^2 \to \infty$ we approach the unpooled model and $\widehat{\alpha}_j \to \overline{y}_j$.

This is related to the idea of *shrinkage*. By assuming a common underlying distribution for the α_j, we shrink them towards the pooled mean, see (3.13). How much we shrink depends on the ratio $n_j \sigma_\alpha^2 / \sigma_y^2$.

3.3 REML, likelihood ratio tests and model selection

The `lme()` function in `library(nlme)` fits using REML (Restricted Maximum Likelihood) by default.

REML is a two stage process which estimates the variances σ_y^2 and σ_α^2 using a first regression and then computes a 'profile' MLE for all other parameters conditioned on these estimates. It can be shown to have lower bias than the MLE which tends to underestimate (recall that the MLE $\widehat{\sigma}^2$ underestimates σ^2 in normal linear models, Section 1).

The details of this algorithm are not part of the course, though we do use REML estimators. See Pinheiro and Bates (2000, Section 2.2.5) for more info.

Although REML gives good estimates, the REML value isn't a maximum likelihood value. We will tell `lme()` to do the fit by seeking the MLEs if we want to do a likelihood ratio test comparing two nested models. (In some cases – models with the same fixed-effects structure – it is possible to compare REML values (Pinheiro and Bates, 2000, Chapter 2).)

3.4 Random effects, blocks and treatments

In an experiemental design with blocks and treatments, the block variables often come from a larger population. We have seen block variables index e.g. litters of pigs – we could regard the litters we observe as coming from a larger population of litters. In this context it is natural to shrink the block effects using a mixed-effects model.

Example: Piglet diet data.

	Diet		
Litter	A	B	C
I	89	68	62
II	78	59	61
III	114	85	83
IV	79	61	82

Let X be the matrix of treatment variables (diets) and Z be the matrix of block variables (litters). We can write

$$y = \alpha + X\beta + Z\gamma + \epsilon.$$

Here β will be the fixed effects and we could model γ_j, for $j = 1, \ldots, J$, as random effects (with J the number of blocks).

If we use random effects, and so `lme()`, we are adding $\gamma_j \sim N(0, \sigma_\gamma^2)$ to the model. See R-example.

3.5 Random effects on slopes

If some variables are drawn from a population then the same principles which apply to the intercepts may apply to the slopes.

We are sticking with the case where we just have one explanatory variable x, so a basic regression would be $y_k = \alpha + \beta x_k + \epsilon_k$. In Section 3.1 we considered a random effect on the intercept: we replaced α by a random effect α_j (or equivalently by $\alpha + b_j$, see (3.4)). We can also consider a random effect on the slope β by replacing it by a random effect β_j. The model with random effects on both intercept and slope is:

$$y_k = \alpha_{j_k} + \beta_{j_k} x_k + \epsilon_k, \quad k = 1, \ldots, n \tag{3.14}$$

where

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \overset{\text{iid}}{\sim} N\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right], \quad j = 1, \ldots, J. \tag{3.15}$$

In (3.14) we have observation errors $\epsilon_k \overset{\text{iid}}{\sim} N(0, \sigma_y^2)$ as usual. The group intercepts and slopes (α_j, β_j) are modelled in (3.15) as coming from an underlying bivariate normal

population (the same bivariate normal for all j). The pairs (α_j, β_j) are independent as j varies. But for a particular value of j the quantities α_j and β_j are not independent: ρ is a parameter allowing correlation of the random effects in slope and intercept.

As in Section 3.1 (see (3.4), (3.7)) we can work with zero mean random effects: let

$$\alpha_j = \alpha + b_{1j}$$
$$\beta_j = \beta + b_{2j}$$

where $(b_{1j}, b_{2j})^T$ will be bivariate normal with zero mean below.

We can write the model (3.14), (3.15) as

$$y_{ij} = \alpha + b_{1j} + (\beta + b_{2j})x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, n_j \text{ and } j = 1, \ldots, J$$

where

$$\begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix} \overset{\text{iid}}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right], \quad j = 1, \ldots, J$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_y^2)$.

3.6 Conclusions

Hierarchical models and more general models with random effects are very widely used to accommodate group structure in explanatory variables.

If the number of groups is large then σ_α is well estimated. If there are just a few groups it is poorly estimated and the unpooled or pooled models could be used.

Gelman and Hill (2007): "There is little risk from applying a multilevel model, assuming we are willing to put in the effort to set up the model and interpret the resulting inferences."

This try-it-and-see approach makes sense – we should used mixed effects models fairly routinely for data sets of any reasonable size.