UNIVERSITY OF EDINBURGH

SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM

Bruce Worton

Semester 1, 2022–2023

Multiple linear regression models and generalized linear models are widely used in statistical analysis. These models are appropriate for investigating how the distribution of a response variable is influenced by explanatory variables. Linear regression models under the assumption of Normality are shown to be special cases of a general Normal Linear Model, which is conveniently expressed in matrix notation, while generalized linear models provide a broad framework for statistical modelling of discrete or continuous data. Special cases of such models are considered, as well as models with random effects.

Course outline

1. Statistical modelling.
2. The exponential family of distributions and GLMs.
3. Multiple regression.
4. Inference for linear models.
5. Generalized linear models.
6. Using models with random effects.

Textbooks

1. Dobson, A.J. & Barnett, A., *An Introduction to Generalized Linear Models*, 2nd/3rd Edition, Chapman & Hall/CRC.
   Provides an introduction to linear models and generalized linear models.

2. McCullagh, P. & Nelder, J.A., *Generalized Linear Models*, 2nd Edition, Chapman & Hall.
   Provides a more advanced and detailed coverage than Dobson's book.

3. Venables, W.N. & Ripley, B.D., *Modern Applied Statistics with S*, 4th Edition, Springer.
   Comprehensive guide to S/R. Introduces the language, and then covers a wide range of statistical techniques, including linear models and generalized linear models.

4. Wood, S.N. *Generalized Additive Models: An Introduction with* R, Second Edition, CRC.
   Excellent coverage of the material. I would highly recommend!

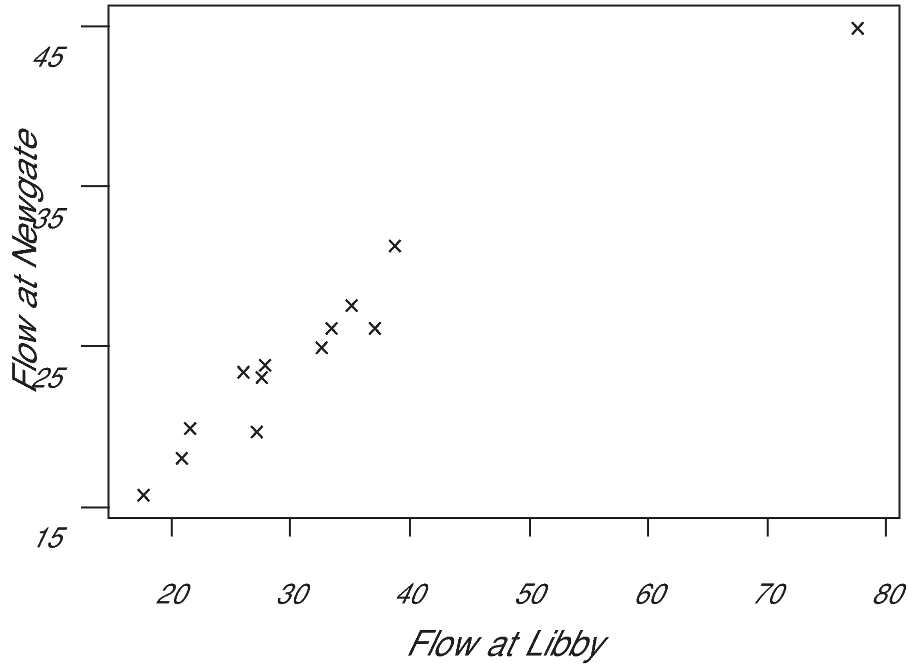
1 Statistical modelling: Relationships between variables

The following examples illustrate various aspects of regression analysis, as well as generalized linear modelling. All the sets of data described are available from Learn.

1.1 January flows on the Kootenai river

While planning a hydro-electric scheme on the Kootenai River in the USA, engineers examined records of the January river flow at Newgate, where a dam was to be built, and at Libby, a place about 50 miles downstream. The records for Libby went back 19 years but those for Newgate existed for only 13, so they wanted to use the earlier records at Libby to estimate the flow at Newgate for the six earlier years. The January flows at the two places are given below (in hundreds of cubic feet per second) and in the file Riverflow.txt: the plot shows the flows for the 13 later years.

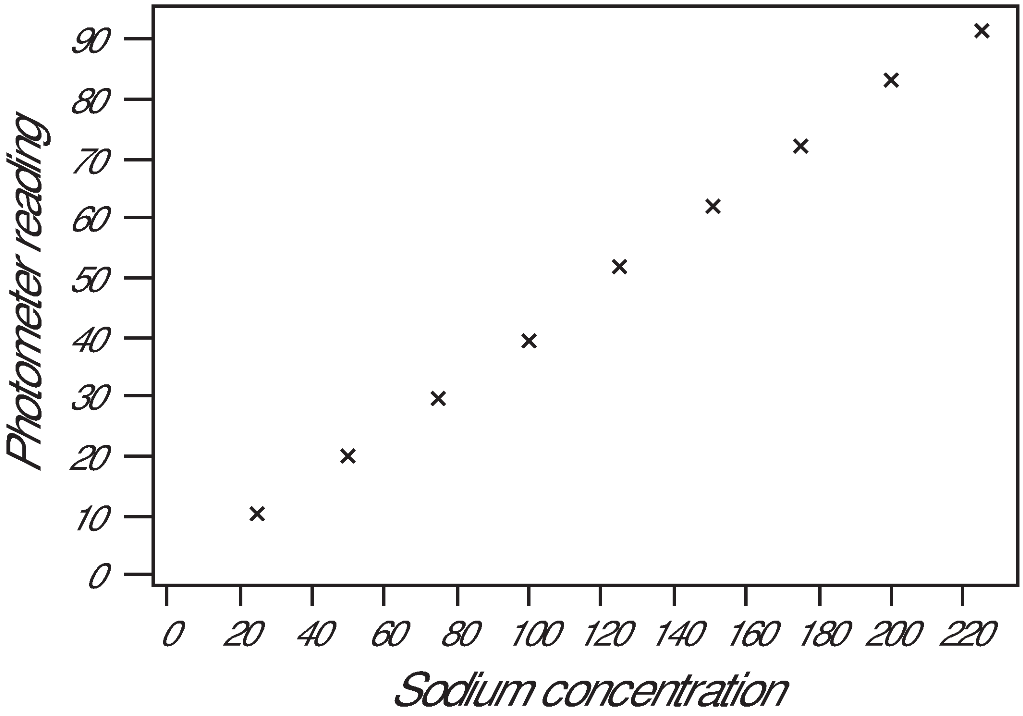
Newgate	–	–	–	–	–	–	19.7	18.0	26.1	44.9	26.1	19.9	15.7	27.6	24.9	23.4	23.1	31.3	23.8
Libby	42.0	24.0	38.0	49.4	24.6	24.2	27.1	20.9	33.4	77.6	37.0	21.6	17.6	35.1	32.6	26.0	27.6	38.7	27.8



1.2 Calibration of a flame photometer

A flame photometer is an instrument that can be used to measure the concentration of sodium in chemical samples. It needs to be calibrated before analysing unknown concentrations, and so standard samples are tested first to provide a 'calibration curve' for converting further measurements into estimates of concentration. The results of such a calibration trial are given below and in Flame.txt.

Sodium concentration	25	50	75	100	125	151	175	200	225
Photometer reading	10	20	29.5	39.5	52	62	72	83.5	91.5



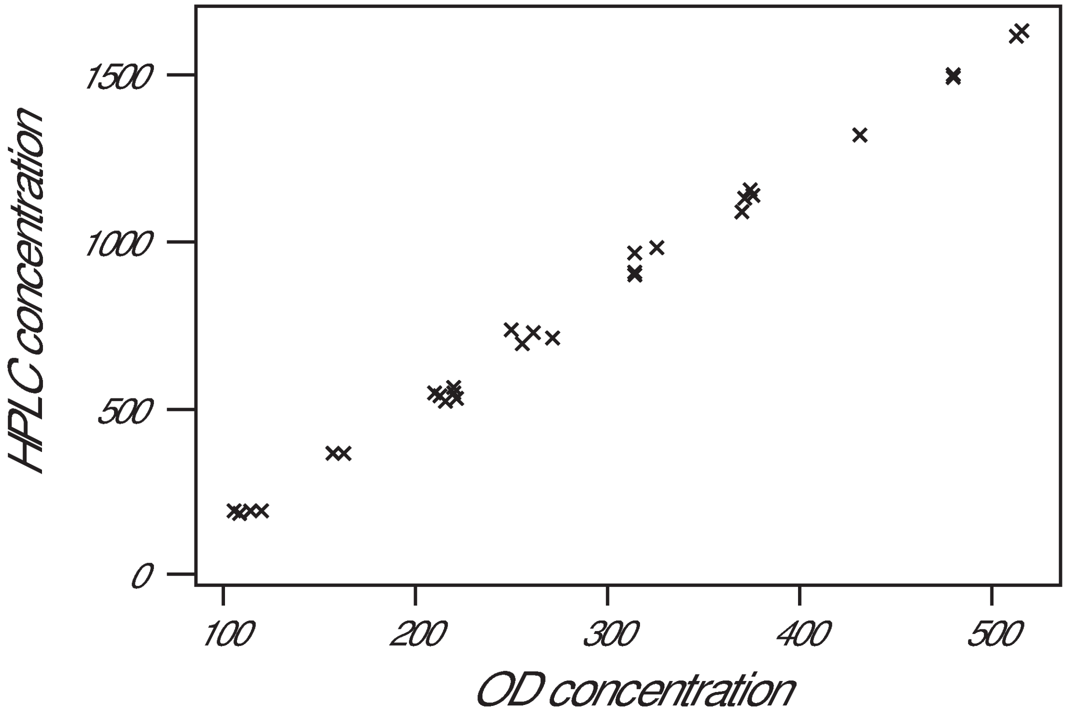
1.3 Comparison of two methods for measuring the bitterness of beer

The bitterness of a beer is a function of the concentration of *iso-alpha acids*; the standard method of measuring this concentration is the optical density (OD) method. The research department of a brewery conducted an experiment to compare this method with a new and more convenient one based on high-pressure liquid chromatography (HPLC). Fifteen samples of beer were prepared with different degrees of bitterness intended to cover the range of interest, and each was split into two sub-samples. The concentration of iso-alpha acids in each sub-sample was measured using both methods: the results are given in Beer.txt and shown below. [For the sake of commercial confidentiality, linear transformations have been applied to the measurements.]

Sample number	1	1	2	2	3	3	4	4	5	5
HPLC reading	193.1	190.4	538.5	544.5	707.6	723.3	911.6	903.7	1153.5	1144.5
OD reading	115	121	214	210	272	262	314	314	374	376

Sample number	6	6	7	7	8	8	9	9	10	10
HPLC reading	1321.4	1319.0	1495.4	1500.4	1621.2	1634.2	181.9	191.4	366.1	360.8
OD reading	432	432	480	480	512	516	109	106	158	163

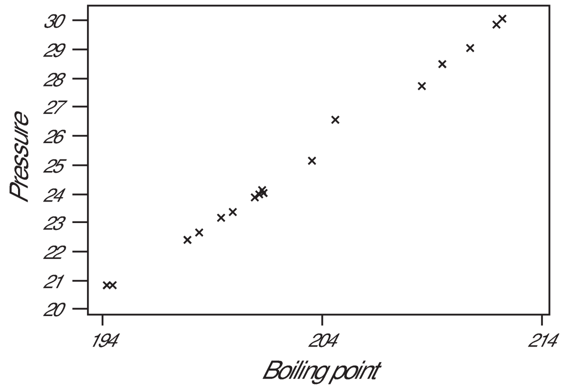
Sample number	11	11	12	12	13	13	14	14	15	15
HPLC reading	531.2	522.3	548.0	560.3	732.1	693.9	963.2	986.9	1135.0	1089.8
OD reading	222	216	220	220	250	256	314	326	372	370

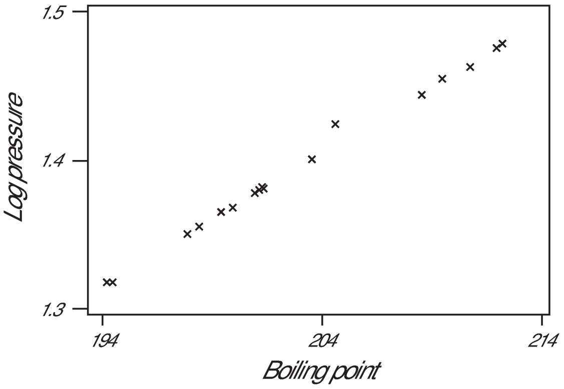


1.4 Forbes' data: Atmospheric pressure and the boiling point of water

In 1857 James Forbes, Professor of Natural Philosophy at the University of Edinburgh, published a paper entitled 'Further experiments and remarks on the measurement of heights by the boiling point of water' (*Transactions of the Royal Society of Edinburgh*, 21, 135–143). It was known that altitude could be estimated from atmospheric pressure, measured with a barometer, but barometers in those days were fragile and clumsy; he wanted to show that atmospheric pressure could be more conveniently measured by determining the boiling point of water and then using the relationship between boiling point and barometric pressure. The following values (also given in Forbes.txt) are the barometric pressures (in inches of mercury and adjusted for ambient air temperature) and the corresponding boiling points (in °F) recorded at 17 locations in Edinburgh and in the Alps. In the graphs which follow, the pressures and their logarithms (to base 10) are plotted against the boiling points: Forbes' theory suggested that the *logarithm* of pressure should be linearly related to boiling point.

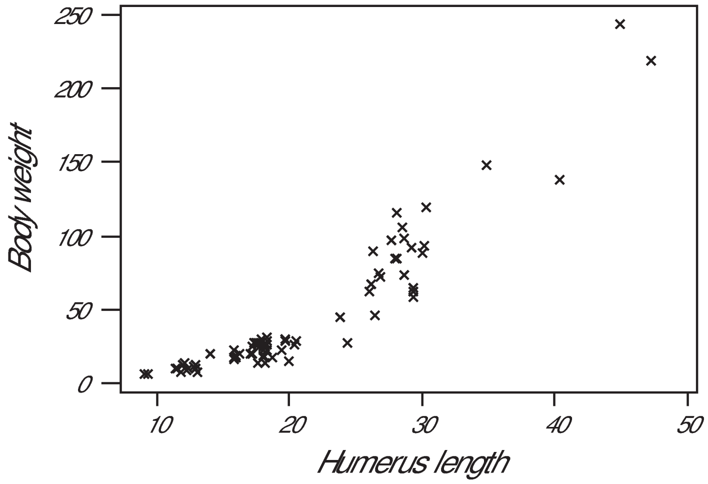
Pressure	20.79	20.79	22.40	22.67	23.15	23.35	23.89	23.99	24.02
Boiling point	194.50	194.25	197.90	198.43	199.45	199.95	200.93	201.15	201.35
Pressure	24.10	25.14	26.57	27.76	28.49	29.04	29.88	30.06	
Boiling point	201.30	203.55	204.60	208.57	209.47	210.72	211.95	212.18	

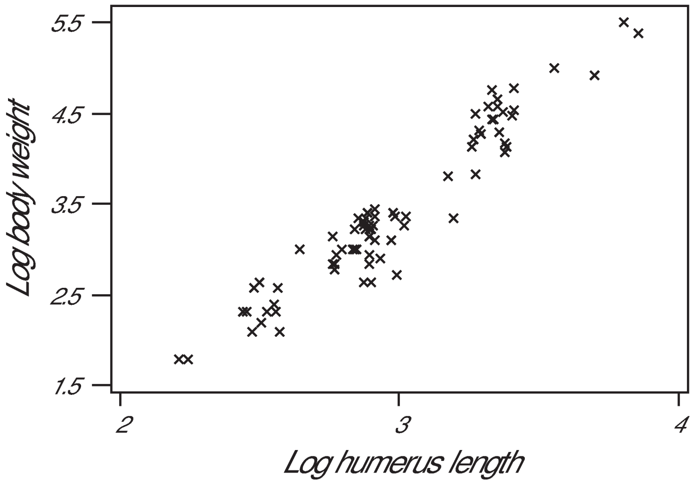




1.5 Estimating the weights of birds from the lengths of wing bones

A zoologist studying the diets of owls wanted to estimate the weight of birds caught and eaten by owls using measurements on birds' bones found in owl pellets. He collected and measured fresh specimens of 78 birds belonging to species he thought were typical of the prey of owls, ranging in size from goldcrests at about 5 grams to a magpie at 244 grams: the data are in BirdWt.txt. The first plot shows the fresh body weight (g) of each bird and length of its humerus bone (mm). The second plot is of the (natural) logarithms of these two measurements.



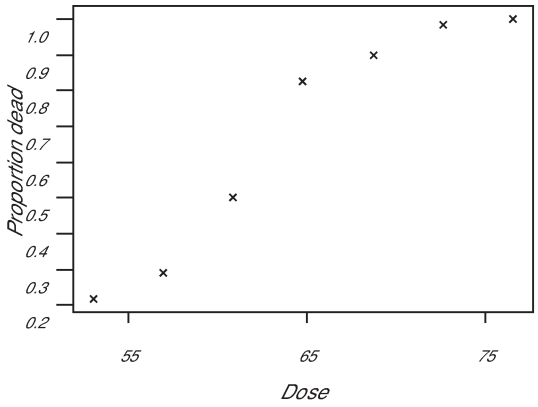


1.6 Numbers of beetles killed by different doses of insecticide

Seven doses of carbon disulphide, an insecticide, were applied to groups of beetles. The doses (in mg / l), the numbers of beetles receiving these doses, and the numbers found to be dead after 5 hours are given below and in Beetles.txt. The first plot, of the proportion dead against dose, shows that its slope increases and then decreases with increasing dose. The second plot, of the logit of this proportion against log dose, is closer to linearity.

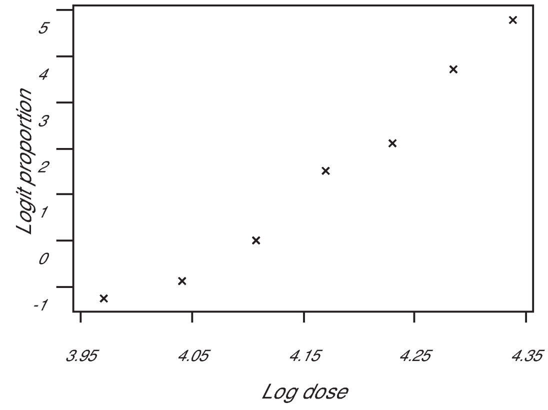
Dose	53.00	56.91	60.84	64.76	68.69	72.61	76.54
Number of beetles	60	62	56	63	59	62	60
Number dead	13	18	28	52	53	61	60

$$\texttt{Proportion dead} = \frac{\texttt{Number dead}}{\texttt{Number of beetles}}$$



`logit = loge((No. dead + 0.5)/(No. in group - No. dead + 0.5))`

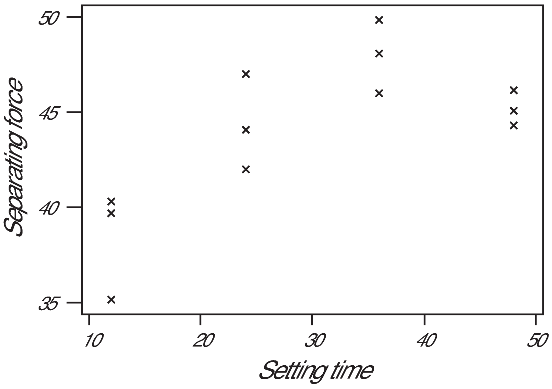
Plot of logit against log dose.



1.7 Effect of setting time on the strength of an adhesive

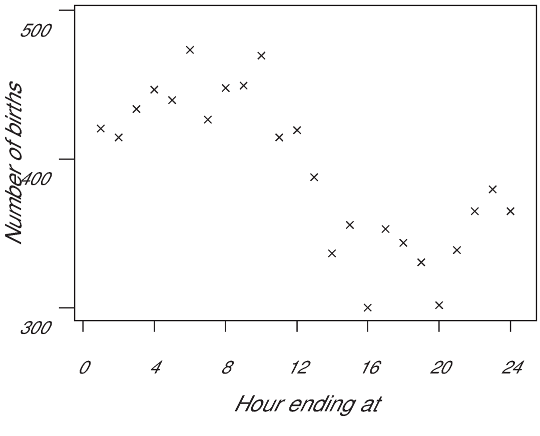
In a study of the setting strength of a woodwork adhesive, the following procedure was carried out. Adhesive was applied to two strips of wood. After 15 minutes the two strips were clamped together at right angles for a fixed time while the adhesive set. Then the force (in kg) required to separate the two strips was measured. Three pairs of pieces were used at each of four setting times with the results shown below (and given in Adhesive.txt).

Setting time	12	12	12	24	24	24	36	36	36	48	48	48
Separating force	35.1	39.7	40.3	42.0	44.1	47.0	46.0	48.1	49.9	44.3	45.1	46.2



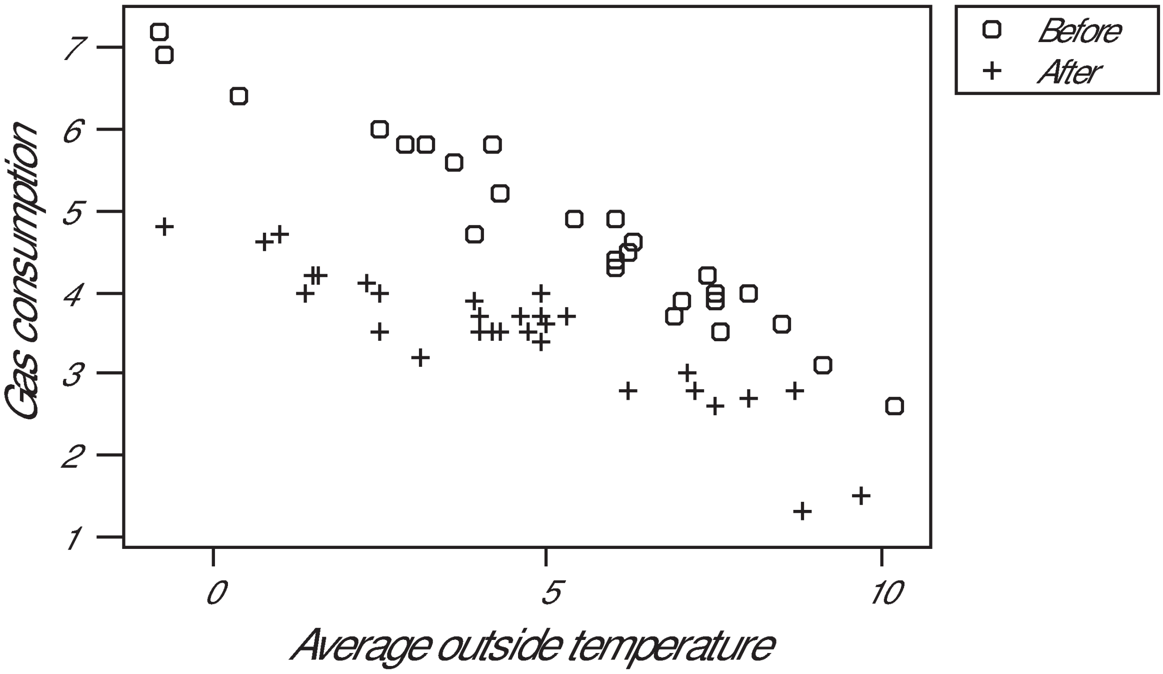
1.8 Numbers of hospital births and time of day

The numbers of normal human births in a hospital were recorded for each hour of the day over several years. The hours and numbers per hour are plotted below and given in the file Births.txt.



1.9 Domestic gas consumption with and without cavity-wall insulation

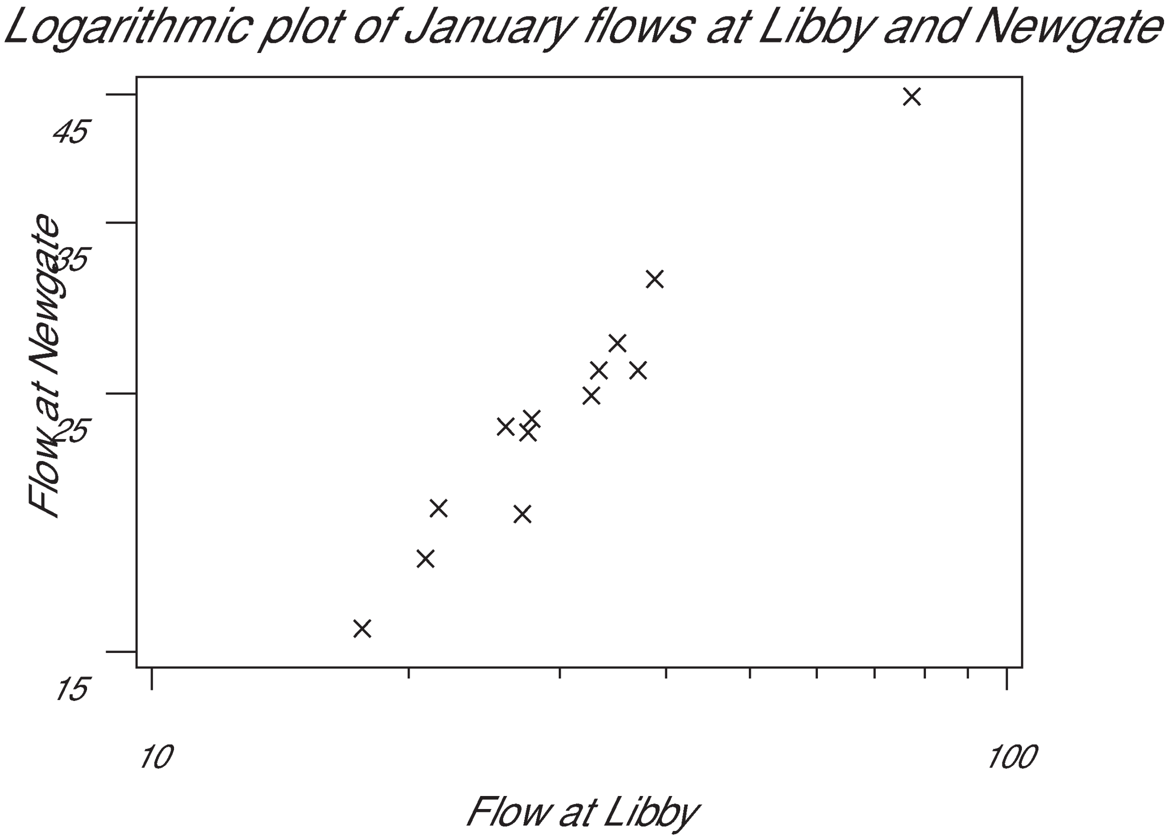
The average outside temperature (in °C) and the weekly gas consumption (in thousands of cubic feet) were recorded at a house in south-east England which had gas-fired central heating for 26 weeks before and 30 weeks after cavity-wall insulation was installed. The house thermostat was set to 20°C throughout the 56 weeks. The 56 pairs of values are given in Insulate.txt along with a column indicating which of them were recorded before and after installation.



1.10 Transformations

As we can see from some of the previous examples, transformations of the data are often necessary to obtain a linear relationship (or, at least, to improve linearity) between a response variable (y) and an explanatory variable (x).

As a further example, compare the following (logarithmic, to base 10) plot with the plot shown in Section 1.1.



2 Exponential family of distributions and GLMs

Definition: A distribution is said to belong to the **exponential family of distributions** if its probability density function (or probability function in the discrete case) can be written in the form

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

where $a(y)$ and $d(y)$ are functions of y but not θ, and $b(\theta)$ and $c(\theta)$ are functions of θ but not y. Many of the common distributions are members of this family. If $a(y) = y$, i.e. a is the identity function, then the exponential family distribution is said to be in **canonical form**, and in this case $b(\theta)$ is called the **natural parameter** of the distribution.

Examples:

- Normal distribution with mean θ and variance σ^2, $Y \sim N(\theta, \sigma^2)$ (σ^2 known):

  $$f(y;\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y\theta}{\sigma^2} - \frac{1}{2}\frac{\theta^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{y^2}{\sigma^2}\right\}.$$

  Thus, we may take, $a(y) = y$, $b(\theta) = \frac{\theta}{\sigma^2}$, $c(\theta) = -\frac{\theta^2}{2\sigma^2}$ and $d(y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{y^2}{\sigma^2}$.

- Poisson distribution with mean θ, $Y \sim \text{Poisson}(\theta)$:

  $$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!} = \exp(y\log\theta - \theta - \log y!).$$

  Thus, we may take, $a(y) = y$, $b(\theta) = \log(\theta)$, $c(\theta) = -\theta$ and $d(y) = -\log y!$.

- Bernoulli distribution with probability parameter θ, i.e. binomial distribution with parameters $m = 1$ and θ, $Y \sim Bi(1,\theta)$:

  $$f(y;\theta) = \theta^y(1-\theta)^{1-y} = \exp\left\{y\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\}.$$

  Thus, we may take, $a(y) = y$, $b(\theta) = \log(\frac{\theta}{1-\theta})$, $c(\theta) = \log(1-\theta)$ and $d(y) = 0$.

In each of the above examples $a(y) = y$, and thus the normal, Poisson and Bernoulli distributions may be expressed in canonical form. The natural parameters are given by $b(\theta)$, i.e.

Distribution	*Natural parameter*
Normal	$b(\theta) = \frac{\theta}{\sigma^2}$
Poisson	$b(\theta) = \log(\theta)$
Bernoulli	$b(\theta) = \log(\frac{\theta}{1-\theta})$

2.1 Mean and variance of $a(Y)$

In this section we show that the mean and variance of $a(Y)$ are given by:

$$\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad \mathrm{var}\{a(Y)\} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}.$$

To obtain these expressions, we require the following results from likelihood theory: *If* $l(\theta)$ *is the log likelihood function for* θ, *then*

(i) $\mathrm{E}(U) = 0$, *and*

(ii) $\mathrm{var}(U) = \mathrm{E}(U^2) = -\mathrm{E}(U')$,

(under very general conditions) where $U = l'(\theta) = \frac{dl(\theta)}{d\theta}$. [See Problem Sheet 1 which gives examples in the cases of exponential and binomial distributions.]

Solving $U = l'(\theta) = \frac{dl(\theta)}{d\theta} = 0$ gives the **maximum likelihood estimator** (MLE) of θ. U is called the **score function**, and var(U) is called **Fisher's information** (the inverse of which is the asymptotic variance of the maximum likelihood estimator).

If observations $y_1, \ldots, y_n$ have been drawn independently from a probability density function $f(y;\theta)$, then the likelihood for θ is

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i;\theta),$$

and the log likelihood is given by

$$l(\theta) = \log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i;\theta).$$

Consider the case when $n = 1$, i.e. the log likelihood for a **single observation** y drawn from a distribution with probability density function $f(y;\theta)$. The log likelihood is then given by

$$l(\theta) = \log L(\theta; y) = \log f(y;\theta) = a(y)b(\theta) + c(\theta) + d(y).$$

The score function is given by

$$U = l'(\theta) = a(y)b'(\theta) + c'(\theta).$$

Differentiating the score function with respect to θ gives

$$U' = l''(\theta) = a(y)b''(\theta) + c''(\theta).$$

Note that in the following the observation y is replaced by a random variable Y (θ is treated as fixed), and thus U and U' are random variables, as in the likelihood theory results given above.

Since $\mathrm{E}(U) = 0$ it follows that

$$0 = \mathrm{E}(U) = \mathrm{E}\{a(Y)\}b'(\theta) + c'(\theta),$$

and thus

$$\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)}.$$

Also

$$\begin{aligned} \mathrm{var}(U) &= \{b'(\theta)\}^2 \mathrm{var}\{a(Y)\} \\ \text{and } -\mathrm{E}(U') &= -b''(\theta)\mathrm{E}\{a(Y)\} - c''(\theta). \end{aligned}$$

But since $\mathrm{var}(U) = -\mathrm{E}(U')$, we obtain

$$\mathrm{var}\{a(Y)\} = \frac{-b''(\theta)\mathrm{E}\{a(Y)\} - c''(\theta)}{\{b'(\theta)\}^2} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}.$$

Examples:

- Normal distribution, $Y \sim N(\theta,\sigma^2)$ (σ^2 known), $a(y)=y$, $b(\theta)=\frac{\theta}{\sigma^2}$, $c(\theta)=-\frac{\theta^2}{2\sigma^2}$:

$$\begin{aligned} \mathrm{E}(Y) &= -\frac{c'(\theta)}{b'(\theta)} = -\frac{-\theta/\sigma^2}{1/\sigma^2} = \theta, \\ \mathrm{var}(Y) &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} = \frac{(0)c'(\theta)-(-1/\sigma^2)(1/\sigma^2)}{\{1/\sigma^2\}^3} = \sigma^2. \end{aligned}$$

- Poisson distribution, $Y \sim \text{Poisson}(\theta)$, $a(y)=y$, $b(\theta)=\log(\theta)$, $c(\theta)=-\theta$:

$$\begin{aligned} \mathrm{E}(Y) &= -\frac{c'(\theta)}{b'(\theta)} = -\frac{-1}{1/\theta} = \theta, \\ \mathrm{var}(Y) &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} = \frac{(-1/\theta^2)(-1)}{\{1/\theta\}^3} = \theta. \end{aligned}$$

Using these general results it is easy to find the mean and variance of $a(Y)$ for **any** member of the exponential family.

2.2 Maximum likelihood estimation

Suppose that $y_1,\ldots,y_n$ is a sample drawn independently from a distribution with probability density function

$$f(y;\theta)=\exp\{a(y)b(\theta)+c(\theta)+d(y)\}.$$

The maximum likelihood estimate (MLE) of θ is determined by θ which maximizes the likelihood function $L(\theta)=\prod_{i=1}^n f(y_i;\theta)$, but, as usual, it is often more convenient to maximize the log likelihood function

$$\begin{aligned} l(\theta) &= \log\left[\prod_{i=1}^n \exp\{a(y_i)b(\theta)+c(\theta)+d(y_i)\}\right] \\ &= b(\theta)\sum_{i=1}^n a(y_i)+nc(\theta)+\text{constant}. \end{aligned}$$

Differentiate the log likelihood to obtain the score function

$$U(\theta)=l'(\theta)=b'(\theta)\sum_{i=1}^n a(y_i)+nc'(\theta).$$

Solving $U(\widehat{\theta})=0$ determines the maximum likelihood estimate for θ (provided that $\widehat{\theta}$ corresponds to a maximum, i.e. $l''(\widehat{\theta})<0$, and $l(\theta)$ is twice differentiable at $\widehat{\theta}$).

Examples:

- $Y \sim \text{Poisson}(\theta)$, $a(y_i)=y_i$, $b'(\theta)=\frac{1}{\theta}$, $c'(\theta)=-1$.
  Thus, the maximum likelihood estimate for θ is the solution of

  $$b'(\theta)\sum_{i=1}^n a(y_i)+nc'(\theta)=\frac{1}{\theta}\sum_{i=1}^n y_i-n=0,$$

  which gives the MLE of θ as $\widehat{\theta}=\frac{\sum_{i=1}^n y_i}{n}=\bar{y}$, the sample mean of the observations.

- Bernoulli, $Y \sim \text{Binomial}(1,\theta)$, $a(y_i) = y_i$, $b'(\theta) = \frac{1}{\theta(1-\theta)}$, $c'(\theta) = -\frac{1}{(1-\theta)}$.

  The maximum likelihood estimate for θ is the solution of
  $$b'(\theta)\sum_{i=1}^{n} a(y_i) + nc'(\theta) = \left\{\frac{1}{\theta(1-\theta)}\right\}\sum_{i=1}^{n} y_i - \frac{n}{(1-\theta)} = 0,$$
  which gives the MLE of θ as $\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$, the sample proportion.

Example:

- Pareto distribution, $Y \sim f(y;\theta) = \theta y^{-\theta-1} \quad (y > 1)$,
  $$f(y;\theta) = \exp\{-\theta\log y + \log\theta - \log y\} \quad (y > 1).$$
  Thus the Pareto is a member of the exponential family, but is **not** in canonical form since $a(\cdot)$ is not the identity function. To transform to canonical form, we use $z = \log y$ thus
  $$\begin{aligned} f(z;\theta) &= f(y;\theta) \times \left|\frac{dy}{dz}\right| \\ &= \exp\{-\theta z + \log\theta - z\} \times \left|\frac{dy}{dz}\right| \\ &= \exp\{-\theta z + \log\theta + d(z)\} \quad (z > 0). \end{aligned}$$
  Applying the general result from Section 2.1 for the mean of a distribution which is a member of the exponential family we obtain
  $$\mu = \mathrm{E}(Z) = -\frac{c'(\theta)}{b'(\theta)} = -\frac{1/\theta}{-1} = \frac{1}{\theta}.$$
  One way to view this transformation of the response variable (to produce canonical form) is that it corresponds to making a **log transformation** of the data, $z_i = \log y_i$.

2.3 Definition of a generalized linear model

Definition: A generalized linear model has the following three components:

- **Model matrix:**
  $$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$
  of known constants, with associated parameters $\beta = (\beta_1, \ldots, \beta_p)^T$.

- **Link function:** A link function $g(\cdot)$ which links together the mean
  $$\mu_i = \mathrm{E}(Y_i),$$
  and the **linear component** $\mathbf{x}_i^T\beta$,
  $$g(\mu_i) = \mathbf{x}_i^T\beta.$$

- **Exponential family:** Each response Y_i has a distribution that is from a member of the exponential family with pdf
  $$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\}.$$

3 Multiple regression

3.1 Simple linear regression

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

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

A more concise notation uses an $n \times 2$ matrix $\mathbf{X}$ whose columns are $\mathbf{1}_n$ and $\mathbf{x}$, and a 2-vector β whose elements are the unknown parameters β_0 and β_1. The model then becomes

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

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

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

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

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

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

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

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

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

3.2 Some other linear models

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

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

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

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

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

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

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

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

(c) **Random samples from two distributions**. Suppose that random variables $Y_1, \ldots, Y_m$ have a common distribution with expectation μ_1, and $Y_{m+1}, \ldots, Y_n$ have a common distribution with expectation μ_2. Then the random vector with elements $Y_1, \ldots, Y_n$ has expectation

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

(d) **Two simple linear regressions**. Suppose that responses $Y_1, \ldots, Y_m$ satisfy the simple linear regression $\mathrm{E}(Y_i\,|\,x_i) = \alpha_1 + \beta_1 x_i$, while $Y_{m+1}, \ldots, Y_n$ satisfy $\mathrm{E}(Y_i\,|\,x_i) = \alpha_2 + \beta_2 x_i$ (using α and β rather than β_0 and β_1 for the intercepts and slopes to avoid double subscripts). Thus the first m responses follow one regression equation while the remaining $n - m$ follow another (as might be assumed in Example 1.9). Writing $\mathbf{x}_1 = (x_1 \ldots x_m)^T$ and $\mathbf{x}_2 = (x_{m+1} \ldots x_n)^T$, the combined model may be expressed as

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

3.3 The ‘Normal Linear Model’

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

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

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

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

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

If the model contains an ‘intercept’ or ‘constant term’, such as β_0 in (3.1.1) and (3.2.1) or γ in (3.1.4), this corresponds to a column of 1’s in $\mathbf{X}$. If $\mathbf{X}$ has full rank p then the $p \times p$ matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, and the least squares estimates are unique. It is sometimes convenient to consider a model which is not of full rank, and to define the least squares estimates using generalized inverses. For making inferences about β, we also assume that $\mathbf{Y}$ has a multivariate Normal distribution (given $\mathbf{X}$). The distribution of $\mathbf{Y}$ is then $N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$.

3.4 Least squares estimation

For least squares estimation we find the values of the β’s which minimize the sum of squares

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

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

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

Using Section 10 of *Useful Matrix Results*, the vector of partial derivatives of Q with respect to the β’s is given by

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

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

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

These may also be written

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

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

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

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

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

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

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

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

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

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

3.5 Expectation and variance matrix of least squares estimator

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

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

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

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

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

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

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

3.6 Fitted values and residuals

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

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

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

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

showing that the vector of residuals is orthogonal to each of the p columns of $\mathbf{X}$. If the model includes an 'intercept' (usually written as β_0) then $\mathbf{X}$ includes the column $\mathbf{1}_n$, so that (3.6.2) implies

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

and the raw residuals sum to zero.

3.6.1 Fitted values and residuals as projections

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

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

and

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

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

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

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

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

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

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

3.6.2 Expectation and variance matrix of the residuals

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

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

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

3.7 Estimation of σ^2

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

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

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

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

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

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

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

3.8 Distributions of the sums of squares under Normality

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

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

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

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

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

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

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

3.9 An alternative formulation for models with an intercept

Consider a linear regression model of the form

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

Here the 'intercept' or 'constant term' β_0 corresponds to a column of 1's in $\mathbf{X}$, as in (3.1.1), (3.1.4), (3.2.1) and (3.2.3). For this sort of model, we usually decompose the sum of squares $\sum_i (y_i - \overline{y})^2$ about the mean response $\overline{y}$ rather than the (raw) total $\sum_i y_i^2$ considered in Section 3.8. It is convenient to use the equivalent model

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

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

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

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

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

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

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

3.9.1 Least squares estimation

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

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

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

or (using (3.9.5))

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

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

so that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.9.2 Sums of squares

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

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

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

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

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

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

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

(c) the two sums of squares are independent.

We again estimate σ^2 using the *residual mean square*, now defined as

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

4 Inferences about parameters in Normal Linear Models

4.1 Introduction

In the general *Normal Linear Model* of §3.3, an n-vector of responses $\mathbf{Y}$ follows the distribution $N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, given the value of an $n \times p$ matrix $\mathbf{X}$. Here $\mathbf{X}$ is assumed to have full rank, so that $\mathbf{X}^T\mathbf{X}$ is invertible, although this condition can be relaxed. We now consider making inferences about one or more of the coefficients in the p-vector β.

The Normality of $\mathbf{Y}$ given $\mathbf{X}$ implies that the least-squares estimator $\widehat{\beta}$ is also Normal: from (3.5.2) and (3.5.3), its distribution is $N_p(\beta, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1})$. Also (from §3.8) the residual sum of squares (*RSS*) is independent of $\widehat{\beta}$ with the distribution $\sigma^2\chi^2(n-p)$. Hence, if $\widehat{\sigma}^2$ denotes the residual *mean* square (given by $RSS/(n-p)$), the corresponding estimator is unbiased for σ^2.

When the R regression function `lm` is used, an intercept is included in the model by default, i.e. $\mathbf{X}$ is assumed to include a column of 1's. To omit the intercept use `-1` in the formula, e.g. `lm(y∼-1+x)`.

4.2 Single linear function of β

For inferences about a *single* linear function of β, such as β_1 or $\beta_1 - \beta_2$, we can use tests and confidence intervals based on Student's t. If $\mathbf{c}$ is a constant p-vector then the linear function

$$\mathbf{c}^T\beta = c_1\beta_1 + \ldots + c_p\beta_p \tag{4.2.1}$$

is estimated by $\mathbf{c}^T\widehat{\beta}$. From (3.5.4), the corresponding least squares estimator has estimated standard deviation

$$\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}, \tag{4.2.2}$$

which is sometimes called its *estimated standard error*. Confidence intervals and tests of hypotheses for $\mathbf{c}^T\beta$ can thus be based on the random variable

$$\frac{\mathbf{c}^T\widehat{\beta} - \mathbf{c}^T\beta}{\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}, \tag{4.2.3}$$

which has a t_{n-p} distribution. For example, if $t_{0.025}$ denotes the upper 2.5% point of this distribution, a 95% confidence interval for $\mathbf{c}^T\beta$ is given by

$$\mathbf{c}^T\widehat{\beta} \pm t_{0.025}\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}. \tag{4.2.4}$$

4.2.1 Individual coefficients

Inference about an individual element of β, β_r say, corresponds to taking $\mathbf{c}^T$ equal to $(0\ldots0\,1\,0\ldots0)$ in (4.2.1) with $c_r = 1$, so that (4.2.2) gives the estimated standard error of $\widehat{\beta}_r$ as

$$\widehat{\sigma}\sqrt{r\text{-th diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}}. \tag{4.2.5}$$

The output from R's regression analysis `summary(lm(y∼x))` includes the estimated standard error for each regression coefficient (called '`Std. Error`') with a corresponding t-statistic and significance probability for a two-sided test of the hypothesis that the true coefficient is zero. It is sometimes convenient to parametrize a linear model so that a linear function of interest (such as a difference between two regression slopes) becomes an individual parameter: see §4.3.3 for an example.

4.2.2 Future responses

If $\mathbf{x}_*$ denotes a future possible value of the vector of explanatory variables, then the corresponding expected response is $\mathrm{E}(Y\,|\,\mathbf{x}_*) = \mathbf{x}_*^T\beta$. From (4.2.2), the estimator $\mathbf{x}_*^T\widehat{\beta}$ of this expectation has estimated standard error

$$\widehat{\sigma}\sqrt{\mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}\,. \tag{4.2.6}$$

While, the estimated standard error for an *individual* future response is [following the same argument as in Question 2(b) of Problem Sheet 1],

$$\widehat{\sigma}\sqrt{1+\mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}\,. \tag{4.2.7}$$

In R the function `predict` can be used to calculate standard errors, as well as confidence intervals and prediction intervals,
e.g. for confidence intervals `predict(lm(y∼x),se.fit = TRUE, interval = "confidence")`.

4.3 Single linear function of $\dot{\beta}$ when the model has an intercept

If the model includes an intercept and q explanatory variables, then the alternative formulation given in §3.9 is usually assumed, under which the explanatory variables are measured from their mean values. Thus we have

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \gamma\mathbf{1}_n + \dot{\mathbf{X}}\dot{\beta}\,, \tag{4.3.1}$$

where $\dot{\mathbf{X}}$ has elements $x_{ij}-\overline{x}_j$. If $\dot{\mathbf{X}}$ has full rank then the residual sum of squares, the mean response $\widehat{\gamma}=\overline{Y}$ and the estimator $\widehat{\dot{\beta}}$ of $\dot{\beta}$ are independent of each other with respective distributions $\sigma^2\chi^2(n-q-1)$, $N(\gamma, n^{-1}\sigma^2)$ and $N_q(\dot{\beta}, \sigma^2(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1})$. If $\mathbf{c}$ is now a constant q-vector and c_0 a scalar, then a linear function $c_0\gamma+\mathbf{c}^T\dot{\beta}$ has the estimator $c_0\widehat{\gamma}+\mathbf{c}^T\widehat{\dot{\beta}}$ with estimated standard error

$$\widehat{\sigma}\sqrt{n^{-1}c_0^2+\mathbf{c}^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\mathbf{c}}\,.$$

4.3.1 Individual coefficients

The estimated standard error of $\widehat{\beta}_r$ is expressible as

$$\widehat{\sigma}\sqrt{r\text{-th diagonal element of }(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}}\,. \tag{4.3.2}$$

4.3.2 Future responses

The expected response corresponding to a future possible value $\mathbf{x}_*$ of the q-vector of explanatory variables is $\mathrm{E}(Y\,|\,\mathbf{x}_*) = \gamma + (\mathbf{x}_*-\bar{\mathbf{x}})^T\dot{\beta}$. The estimated standard errors for the estimator $\widehat{\gamma}+(\mathbf{x}_*-\bar{\mathbf{x}})^T\widehat{\dot{\beta}}$ and for an *individual* future response are respectively

$$\widehat{\sigma}\sqrt{n^{-1}+(\mathbf{x}_*-\bar{\mathbf{x}})^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}(\mathbf{x}_*-\bar{\mathbf{x}})}\,, \tag{4.3.3}$$

$$\widehat{\sigma}\sqrt{1+n^{-1}+(\mathbf{x}_*-\bar{\mathbf{x}})^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}(\mathbf{x}_*-\bar{\mathbf{x}})}\,. \tag{4.3.4}$$

4.3.3 Two simple linear regressions

The model described in §3.2(d) for comparing two simple linear regressions is

$$\mathrm{E}(Y_i\,|\,x_i) = \alpha_1 + \beta_1 x_i \; (i=1,\ldots,m), \quad \mathrm{E}(Y_i\,|\,x_i) = \alpha_2 + \beta_2 x_i \; (i=m+1,\ldots,n) \tag{4.3.5}$$

with $\mathrm{var}(Y_i\,|\,x_i) = \sigma^2$. This model does not include a common intercept, since there is no parameter included in all the expectations. It could be fitted in R by omitting the default intercept and using vectors $(1\ldots1\,0\ldots0)^T$, $(x_1\ldots x_m\,0\ldots0)^T$, $(0\ldots0\,1\ldots1)^T$ and $(0\ldots0\,x_{m+1}\ldots x_n)^T$. To compare the slopes of the two fitted lines, we would divide $\widehat{\beta}_1 - \widehat{\beta}_2$ by its estimated standard error and compare this t-statistic with $t(n-4)$.

It is more convenient to redefine the model, replacing (4.3.5) by, say,

$$\mathrm{E}(Y_i\,|\,x_i) = \alpha + \beta x_i \; (i=1,\ldots,m), \quad \mathrm{E}(Y_i\,|\,x_i) = \alpha + \gamma + \beta x_i + \delta x_i \; (i=m+1,\ldots,n), \tag{4.3.6}$$

so that α, β replace α_1, β_1, and parameters γ and δ are respectively the difference between the intercepts $(\alpha_2-\alpha_1)$ and the difference between the slopes $(\beta_2-\beta_1)$. This model may be fitted by *including* the common intercept (corresponding to α) and using vectors $(x_1\ldots x_n)^T$, $(0\ldots0\,1\ldots1)^T$ and $(0\ldots0\,x_{m+1}\ldots x_n)^T$ (corresponding to β, γ and δ respectively). The t-statistic for the final vector tests the difference in the slopes. The difference in the intercepts could also be tested, but this is sensible only if the expected responses at $x=0$ have particular significance.

If the final vector in the above model is omitted, we assume a model with a common slope β but different intercepts α and $\alpha+\gamma$, i.e. the two regression lines are assumed parallel. Under this model, the hypothesis of a common intercept (and hence a common regression) is tested using the t-statistic for γ, i.e. for the vector $(0\ldots0\,1\ldots1)^T$.

Example — Cavity-wall insulation

4.4 Tests of the hypotheses $\beta = \mathbf{0}$ and $\dot{\beta} = \mathbf{0}$

Suppose that we fit the linear model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$, $\mathrm{var}(\mathbf{Y}\,|\,\mathbf{X}) = \sigma^2\mathbf{I}_n$ assuming that the response vector $\mathbf{Y}$ is Normally distributed, and then want to test the hypothesis that β equals $\mathbf{0}$, i.e. that $\beta_1 = \ldots = \beta_p = 0$ or $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{0}$. Separate Student-$t$ tests on the estimated coefficients do not provide an appropriate method, since they test only one coefficient at a time: instead we compare the *model sum of squares* with the *residual sum of squares* for the model.

From §3.8, the model SS is

$$\widehat{\beta}^T\mathbf{X}^T\mathbf{Y} = \mathbf{Y}^T\mathbf{P}_\mathbf{X}\mathbf{Y}, \tag{4.4.1}$$

and has the distribution $\sigma^2\chi^2(p, \sigma^{-2}\beta^T\mathbf{X}^T\mathbf{X}\beta)$; it is independent of the residual SS,

$$(n-p)\widehat{\sigma}^2 = \textstyle\sum_i Y_i^2 - \widehat{\beta}^T\mathbf{X}^T\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}, \tag{4.4.2}$$

which has the distribution $\sigma^2\chi^2(n-p)$. If β *is* equal to $\mathbf{0}$ then the model SS has the distribution $\sigma^2\chi^2(p)$, and the random variable

$$F = \frac{\text{model MS}}{\text{residual MS}} = \frac{\text{model SS}/p}{\widehat{\sigma}^2} \tag{4.4.3}$$

has the distribution $F(p, n-p)$. [If β does *not* equal $\mathbf{0}$ then the model SS has a *non-central* $\sigma^2\chi^2$ distribution with expectation *larger* than $p\sigma^2$.] If the hypothesis is false we expect *large* values for the model SS and hence for F. So we can test the hypothesis that β is $\mathbf{0}$ by comparing the value of F in (4.4.3) with the upper percentage points of $F(p, n-p)$.

More usually, we consider the alternative formulation defined in §3.9, and want to test the model $\mathrm{E}(Y_i \mid \mathbf{X}) = \gamma$ (with no dependence on the explanatory variables) against the model

$$\mathrm{E}(Y_i \mid \mathbf{X}) = \gamma + \beta_1 (x_{i1} - \bar{x}_1) + \ldots + \beta_q (x_{iq} - \bar{x}_q) \quad (i = 1, \ldots, n), \tag{4.4.4}$$

which includes q explanatory variables. This is equivalent to testing the hypothesis that $\dot{\beta}$ is $\mathbf{0}$. From §3.9.2, the regression SS, $\widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \mathbf{Y}$, has distribution $\sigma^2 \chi^2(q, \sigma^{-2} \dot{\beta}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \dot{\beta})$, and is independent of the residual SS, which is distributed as $\sigma^2 \chi^2(n-q-1)$. Thus the expectation of the regression SS is larger than $q\sigma^2$ unless $\dot{\beta}$ is $\mathbf{0}$, so we test the hypothesis $\dot{\beta} = \mathbf{0}$ by comparing the value of

$$F = \frac{\text{regression MS}}{\text{residual MS}} \tag{4.4.5}$$

with the distribution $F(q, n-q-1)$. A *large* value provides evidence against the hypothesis.

4.5 Testing a linear hypothesis using the ‘Extra Sum of Squares’

We now generalize from the hypotheses $\beta = \mathbf{0}$ and $\dot{\beta} = \mathbf{0}$ considered in §4.4 to a linear hypothesis H_0 specifying a set of c linear constraints on β or $\dot{\beta}$. In general, these have the form

$$\mathbf{C}\beta = \mathbf{d} \qquad \text{or} \qquad \mathbf{C}\dot{\beta} = \mathbf{d} \tag{4.5.1}$$

where $\mathbf{C}$ is a specified $c \times p$ (or $c \times q$) matrix of rank c and $\mathbf{d}$ is a specified c-vector. As in the comparison of two simple linear regressions in §4.3.3, we can redefine our linear model so that the c constraints specify the values of c individual parameters rather than of linear combinations.

To suggest the form of the statistic to be used for testing (4.5.1), first note that the null hypothesis in the test of $\dot{\beta} = \mathbf{0}$ is that the responses have a common expectation γ. The least squares estimate of γ under this hypothesis is $\bar{y}$, so the total SS about the mean, $\sum_i (y_i - \bar{y})^2$, is the residual SS under the *null* hypothesis. The regression SS is therefore the *increase* in the residual SS due to imposing the constraint that $\dot{\beta}$ is $\mathbf{0}$: this is called the *extra sum of squares* for this hypothesis, and the magnitude of this extra SS is compared with the residual SS using the F-statistic of (4.4.5).

We can extend the idea of calculating an F-statistic to compare the extra SS for a linear hypothesis with the residual SS to include the more general form of hypothesis H_0 given in (4.5.1). We calculate the statistic

$$\begin{aligned} F &= \frac{(\text{extra SS for } \mathrm{H}_0)/c}{\text{residual MS}} &(4.5.2)\\ &= \frac{(\text{residual SS under } \mathrm{H}_0 - \text{residual SS under full model})/c}{\text{residual MS}}, &(4.5.3)\end{aligned}$$

and compare its value with the upper percentage points of $F(c, n-p)$ or $F(c, n-q-1)$; *large* values of F provide evidence against H_0. The distributions given for F are based on the results that

(a) the residual SS under the full model has the distribution $\sigma^2\chi^2(n-p)$ or $\sigma^2\chi^2(n-q-1)$;

(b) the extra SS is independent of the residual SS;

(c) the extra SS has the distribution $\sigma^2\chi^2(c)$ under H_0 and is non-central $\sigma^2\chi^2(c)$ otherwise.

Assertion (a) comes from §3.8 and §3.9.2. Assertions (b) and (c) are not proved in detail here, but an indication of the method of proof is as follows:

- The random c-vector $\mathbf{C}\widehat{\beta}-\mathbf{d}$ has the distribution $N_c(\mathbf{C}\beta-\mathbf{d}, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)$ and is independent of the residual SS under the linear model of §3.3.
- The expectation of $\mathbf{C}\widehat{\beta}-\mathbf{d}$ is $\mathbf{0}$ under H_0, so the quadratic form

$$(\mathbf{C}\widehat{\beta}-\mathbf{d})^T\left\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right\}^{-1}(\mathbf{C}\widehat{\beta}-\mathbf{d}) \tag{4.5.4}$$

  has the distribution $\sigma^2\chi^2(c)$ independently of the residual SS under H_0.

- The realised value of the quadratic form in (4.5.4) is equal to the extra SS in (4.5.2).
- Under the alternative formulation for models with an intercept, $\dot{\mathbf{X}}$ and $\widehat{\dot{\beta}}$ replace $\mathbf{X}$ and $\widehat{\beta}$.

4.5.1 Calculating an extra sum of squares as a quadratic form

To illustrate the use of the extra SS, we test the linear hypothesis $\beta_1=\beta_2=1$ for a plum-leaf data example. [This hypothesis means that leaf area is approximately proportional to lamina length times breadth.] The residual MS is $\widehat{\sigma}^2=0.000176$ and $\left(\widehat{\beta}_1\ \widehat{\beta}_2\right)=(1.132 \quad 0.836)$, while $\mathbf{C}=\mathbf{I}_2$ and $\mathbf{d}=\mathbf{1}_2$ in (4.5.1). Thus

$$\mathbf{C}\widehat{\dot{\beta}}-\mathbf{d}=(1.132-1 \quad 0.836-1)^T=(0.132 \quad -0.164)^T,$$

and (4.5.4) takes the value

$$\begin{aligned}(\mathbf{C}\widehat{\dot{\beta}}-\mathbf{d})^T\left\{\mathbf{C}(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\mathbf{C}^T\right\}^{-1}(\mathbf{C}\widehat{\dot{\beta}}-\mathbf{d}) &= \begin{pmatrix} 0.132 & -0.164 \end{pmatrix}\begin{pmatrix} 0.6615 & 0.6926 \\ 0.6926 & 0.8268 \end{pmatrix}\begin{pmatrix} 0.132 \\ -0.164 \end{pmatrix} \\ &= 0.0039.\end{aligned}$$

Hence the F-statistic for the hypothesis is

$$f=\frac{0.0039/2}{0.000176}=11.06.$$

The 0.5% point of $F(2,12)$ is 8.51, so the data appear *not* to be consistent with this hypothesis.

4.5.2 Calculating an extra sum of squares as a difference of sums of squares

Rather than calculate the extra SS using (4.5.4), it is usually more convenient to find the extra SS for the hypothesis by fitting the full model and the model constrained by the hypothesis separately and finding the difference between the two residual SS.

For example, suppose a total of n responses Y_{jk} are independent with distributions $N(\mu_j,\sigma^2)$ $(j=1,\ldots,g; k=1,2,\ldots,n_j)$: this structure is sometimes called a *one-way classification*. We might want to test the hypothesis that the expected values $\mu_1,\ldots,\mu_g$ are equal. This amounts to imposing $g-1$ linear constraints on the μ_j, which might be written as $\mu_j-\mu_g=0\ (j=1,\ldots,g-1)$, so that $\beta=\left(\mu_1\ldots\mu_g\right)^T$, $\mathbf{d}=\mathbf{0}$ and $\mathbf{C}$ is $(g-1)\times g$ in (4.5.1).

Rather than find the extra SS from (4.5.4), we note that the residual SS under the full model and under the hypothesis of a common expectation μ, say, are respectively the within-groups SS, $\sum_j \sum_k (y_{jk} - \overline{y}_j)^2$ (with $n-g$ degrees of freedom), and the total SS about the mean, $\sum_j \sum_k (y_{jk} - \overline{y})^2$, so that the extra SS for the hypothesis is their difference, the between-groups SS, $\sum_j n_j (\overline{y}_j - \overline{y})^2$ (with $g-1$ df). The hypothesis is therefore tested by comparing

$$F = \frac{\text{between-groups MS}}{\text{within-groups MS}} = \frac{\text{between-groups SS} \,/\, (g-1)}{\text{within-groups SS} \,/\, (n-g)} \tag{4.5.5}$$

with the distribution $F(g-1, n-g)$.

To apply this analysis in R use `anova(y~as.factor(x))`, where `x` contains the levels of the factor and `y` is the response.

4.5.3 Calculating an extra sum of squares from sequential sums of squares

In R we might use `anova(lm(y~x1))` and then `anova(lm(y~x1),lm(y~x1+x2),lm(y~x1+x2+x3))` to give us 'sequential sums of squares' (*or* this may be read directly from the ANOVA table `anova(lm(y~x1+x2+x3))`). For regression on variables $x_1, \ldots, x_q$, say, these comprise

- the regression SS for fitting x_1 alone,
- the extra SS for fitting x_2 *as well as* x_1, equal to

$$\text{regression SS for fitting } x_1, x_2 \quad - \quad \text{regression SS for fitting } x_1\,,$$

- the extra SS for fitting x_3 in addition to x_1, x_2,

and so on. These can be used to find the extra SS for a hypothesis which sets some of the regression coefficients $\beta_1, \ldots, \beta_q$ to zero, i.e. which omits a subset of the explanatory variables from the model.

For example, if we fit a cubic polynomial to a set of n responses, we might test the hypothesis that a linear polynomial is adequate, putting the explanatory variables in the natural order x, x^2, x^3, and calculating the extra SS for the quadratic and cubic terms as the sum of their sequential SS.

4.5.4 Analysis of variance tables for tests of linear hypotheses

An analysis of variance (or ANOVA) table can be used to present the decomposition of the total SS about the mean into the regression SS and residual SS. This sort of table can be expanded to display the calculations required for a test of a linear hypothesis. The table below shows the general form for a test of such a hypothesis when (as is usual) the model includes an intercept which is not constrained by the hypothesis; the hypothesis is assumed to impose c linear constraints on $\dot{\beta}$ given by $\mathbf{C}\dot{\beta} = \mathbf{d}$, so that there are effectively only $q-c$ parameters in $\dot{\beta}$ to be estimated. Below the dashed line, the total SS about the mean is split into the usual regression SS and residual SS from fitting the unconstrained model. Above the line, this regression SS is itself split into the SS for fitting the model under the hypothesis and the extra SS arising from it. [Here $\widehat{\dot{\beta}}_c$ denotes the vector of estimates under the hypothesis.] The hypothesis can be tested by comparing

$$\frac{\text{hypothesis MS}}{\text{residual MS}} \tag{4.5.6}$$

with the distribution $F(c, n-q-1)$.

Source	DF	SS	MS
Regression under hypothesis	$q-c$	$\widehat{\dot{\beta}}_c^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \widehat{\dot{\beta}}_c$	
Deviations from hypothesis	c	extra SS	hypothesis MS
Regression for unconstrained model	q	$\widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \widehat{\dot{\beta}}$	
Residual for unconstrained model	$n-q-1$	$S_{yy} - \widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \widehat{\dot{\beta}}$	residual MS ($\widehat{\sigma}^2$)
Total about mean	$n-1$	S_{yy}	

Under the more general linear model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$ with $\mathbf{X}$ $n \times p$ and of rank p, there is a similar analysis of variance table for a test of a hypothesis $\mathbf{C}\beta = \mathbf{0}$ in which $\mathbf{C}$ is $c \times p$ and has rank c. The raw total SS (with n df) is first decomposed into the residual SS (with $n-p$ df) and the model SS. The model SS is then split into the SS for the model under the hypothesis (with $p-c$ df) and the extra SS for the hypothesis (with c df).

For an example of this type of table, consider again the example of §4.5.3 in which a cubic polynomial is fitted to n responses, and we want to test the hypothesis that the data follow a linear polynomial. The analysis of variance table would then have the following form.

Source	DF	SS	MS
Linear regression on x	1	S_{xy}^2/S_{xx}	
Deviations from linear regression on x	2	extra SS	hypothesis MS
Cubic regression on x	3	$\widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \widehat{\dot{\beta}}$	
Residual from cubic regression on x	$n-4$	$S_{yy} - \widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \widehat{\dot{\beta}}$	residual MS ($\widehat{\sigma}^2$)
Total about mean	$n-1$	S_{yy}	

4.6 Testing the fit of a linear regression model with replicate data (‘Lack of Fit’ and ‘Pure Error’)

To further illustrate the test of a linear hypothesis and its analysis of variance table, suppose that responses Y_{jk} are independent with distributions $N(\mu_j, \sigma^2)$ $(j = 1, \ldots, g\,; k = 1, 2, \ldots, n_j)$ (one-way ANOVA model) and that there is an explanatory variable x taking the value x_j for group j. The hypothesis to be tested is that the responses have linear regression on x, so that, for some unknown values β_0 and β_1, the expectations $\mu_1, \ldots, \mu_g$ are related by

$$\mu_j = \beta_0 + \beta_1 x_j \quad (j = 1, \ldots, g)\,. \tag{4.6.1}$$

The residual SS under the full model is the within-groups SS. The residual SS under the constraints (4.6.1) equals the total SS about the mean minus the regression SS. The extra SS is thus

$$\begin{aligned} & (\text{within-groups SS} + \text{between-groups SS} - \text{regression SS}) - \text{within-groups SS} \\ = \; & \text{between-groups SS} - \text{regression SS}\,. \end{aligned}$$

This extra SS, with $g-2$ degrees of freedom, measures how much of the variation in the response is *not* explained by linear regression, i.e. the extent of deviations from linearity. The analysis of variance table has the following form.

Source	DF	SS	MS
Linear regression on x	1	$\left\{\sum_j n_j(x_j-\overline{x})\overline{y}_j\right\}^2/\sum_j n_j(x_j-\overline{x})^2$	
Deviations from linearity	$g-2$	extra SS for non-linearity	hypothesis MS
Between groups	$g-1$	$\sum_j n_j(\overline{y}_j-\overline{y})^2$	
Within groups	$n-g$	$\sum_j\sum_k(y_{jk}-\overline{y}_j)^2$	within-groups MS
Total about mean	$n-1$	$\sum_j\sum_k(y_{jk}-\overline{y})^2$	

The hypothesis of linear regression is tested by comparing the ratio of the hypothesis MS to the within-groups MS with $F(g-2,n-g)$. Note that

(a) In this context the within-groups SS is called 'pure error', since its distribution does not depend on the linear regression assumption in (4.6.1).

(b) To apply this test in R, we can use `anova(lm(y~x),lm(y~as.factor(x)))`.

(c) The n_j responses at each x_j are called 'replicates'.

(d) The null hypothesis being tested is that the responses have linear regression on the x_j (with the alternative of arbitrary dependence on j) *not* that there is no dependence on x_j.

(e) Assuming that the regression model in (4.6.1) is true gives more precise inferences about the μ_j, but introduces bias if the model is wrong.

5 Generalized linear models

5.1 Definition of a generalized linear model

Definition: A generalized linear model has the following three components:

- **Model matrix:**
$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$
of known constants, with associated parameters $\beta = (\beta_1, \dots, \beta_p)^T$.

- **Link function:** A link function $g(\cdot)$ which links together the mean
$$\mu_i = \mathrm{E}(Y_i),$$
and the **linear component** $\mathbf{x}_i^T\beta$,
$$g(\mu_i) = \mathbf{x}_i^T\beta.$$

- **Exponential family:** Each response Y_i has a distribution that is from a member of the exponential family with pdf
$$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\}.$$

5.1.1 Canonical link functions

Often, the natural parameter $b(\theta)$ in

$$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\},$$

is used to link the mean μ_i to the linear component $\eta_i = \mathbf{x}_i^T\beta$:

$$g(\mu_i) = b(\theta_i) = \eta_i = \mathbf{x}_i^T\beta$$

This is known as the **canonical link** function. This may or may not provide a satisfactory model. However, it is often used, at least as a starting point, in data analysis.

Family	**Response**	**Canonical link**		**Range**
$Y_i \sim$ Normal (μ, σ_0^2)	Y_i	$g(\mu) = \mu$	identity link	$-\infty < \mu < \infty$
$Y_i \sim$ Poisson (μ)	Y_i	$g(\mu) = \log\mu$	log link	$\mu > 0$
$Y_i \sim$ Binomial (m_i, π)	Y_i/m_i	$g(\pi) = \log(\frac{\pi}{1-\pi})$	logit link	$0 < \pi < 1$

Note: As we shall see in Section 5.6, for a binomial distribution we model the expected proportion, and denote this by $\pi_i = \mathrm{E}(Y_i/m_i)$.

5.2 Estimation

We use MLE to fit a GLM. Unfortunately, there is usually no explicit solution for the maximum likelihood estimates of the elements of β. Therefore, generally, we need an iterative procedure, i.e. **Fisher's method of scoring**, to determine the MLEs. In fact we can show that this is equivalent to an **iterative weighted least squares** procedure.

As in Question 5 on Problem Sheet 4, the idea of weighted least squares is that if the responses Y_i have non-constant variance then we want to weight the contributions of

$$(Y_i - \underbrace{\mu_i(\beta)}_{\text{fitted values}})^2,$$

in the least squares sum by including weighting factors w_i. Thus, the problem is to minimise

$$\sum w_i(Y_i - \mu_i(\beta))^2,$$

for appropriate weights w_i. This leads to the weighted least squares estimator

$$\begin{aligned}\widehat{\beta}_{\text{WLS}} &= (X^T W X)^{-1} X^T W Y,\\ W &= \text{diag}(w_i) \quad \text{where} \quad w_i = (\text{var}(Y_i))^{-1},\end{aligned}$$

if responses are independent. Unfortunately, W often depends on the coefficients β, and thus cannot be used directly. Also in fitting a GLM we need to include the contribution of the link function.

5.2.1 Fisher's method of scoring: iterative weighted least squares

The MLE of β can be obtained as follows. The log likelihood for independent observations $y_1, \dots, y_n$ is

$$l(\beta) = \log\left\{\prod_{i=1}^n f(y_i, \theta_i)\right\} = \sum_{i=1}^n \{y_i b(\theta_i) + c(\theta_i) + d(y_i)\}$$

Define $\mu_i = E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}$ (mean), and $\eta_i = g(\mu_i) = \mathbf{x}_i^T\beta$ (linear component).

$\mathbf{x}_i^T \quad = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})$ ith values of explanatory variables (i.e. ith row of X).

To obtain the MLE we require the solution of

$$U_j = \frac{\partial l}{\partial \beta_j} = 0 \qquad (j = 1, \dots, p) \quad \text{i.e.} \quad U = \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Consider the log likelihood for y_i.

$$l_i = \log f(y_i; \theta_i) = y_i b(\theta_i) + c(\theta_i) + d(y_i)$$

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

since

$$\begin{aligned}\frac{\partial l_i}{\partial \theta_i} &= y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i)\\ \frac{\partial \mu_i}{\partial \theta_i} &= -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{b'(\theta_i)^2} = b'(\theta_i)\text{var}(Y_i)\\ \frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i}.\end{aligned}$$

This leads to

$$U_j = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right).$$

Since solution of $U_j = 0$, $j = 1, \ldots, p$, is often intractable, we use Fisher's method of scoring. This is a modification of the multiparameter Newton-Raphson

$$\beta_r = \beta_{r-1} - \left\{ \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) \right\}^{-1} U(\beta_{r-1}).$$

Fisher's method of scoring replaces the matrix of second derivatives by its expectation. This gives the iterative scoring formula as

$$I_{r-1}\beta_r = I_{r-1}\beta_{r-1} + U_{r-1}.$$

Therefore, the (j,k)th element of I is

$$\begin{aligned} I_{jk} &= \sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial l_i}{\partial \beta_j} \cdot \frac{\partial l_i}{\partial \beta_k} \right) \\ &= \sum_{i=1}^{n} \mathrm{E}\left\{ \frac{(Y_i - \mu_i)^2}{(\mathrm{var}(Y_i))^2} x_{ij} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right\} \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \end{aligned}$$

In matrix notation, this is

$$I = X^T W X \quad \text{with } W = \text{diag } (w_{ii}) \quad \text{where } w_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

Therefore, the jth element of $(I\beta + U)$ is

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \beta_k + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

$$= \sum_{i=1}^{n} \frac{x_{ij}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \left\{ \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \right\}.$$

Thus, the MLE is determined by the solution of the system of p equations

$$X^T W X \widehat{\beta} = X^T W z, \quad \text{where } z \text{ has elements } z_i = \eta_i + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right),$$

which is of weighted least squares form. However, if W, z depend on β, then we must iterate to obtain MLE of β.

Example

- Fitting Poisson (x_i, Y_i) regression model. Y_i independent $\sim Po(\mu_i = \beta_1 + \beta_2 x_i)$. The model assumes the **identity link**, i.e. $g(\mu_i) = \mu_i$.

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

This model has $\mu_i = E(Y_i) = \mathrm{var}(Y_i)$, and $g(\mu_i) = \mu_i = \eta_i = \beta_1 + \beta_2 x_i$, i.e. $\frac{\partial \mu_i}{\partial \eta_i} = 1$. Thus, $w_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\beta_1 + \beta_2 x_i} = \mu_i^{-1}$ (depends on β). Thus, we have

$$I = X^T W X = \begin{pmatrix} \sum_{i=1}^n \mu_i^{-1} & \sum_{i=1}^n x_i \mu_i^{-1} \\ \sum_{i=1}^n x_i \mu_i^{-1} & \sum_{i=1}^n x_i^2 \mu_i^{-1} \end{pmatrix} \quad \text{and} \quad X^T W z = \begin{pmatrix} \sum_{i=1}^n y_i \mu_i^{-1} \\ \sum_{i=1}^n x_i y_i \mu_i^{-1} \end{pmatrix}.$$

Iterate $\widehat{\beta} = (X^T W X)^{-1} X^T W z$ to find MLE of β.

5.3 Fitting models in R/S-PLUS

R/S-PLUS are powerful **statistical** environments for data analysis, and may be used to fit and analyse GLMs. R has almost identical commands to S-PLUS but is free to download[1].

In R/S-PLUS, if we select a distributional **family** for the response variable, then we are automatically given the **canonical link**. So to use the **canonical link** function we just need to give name of distributional family, e.g. for the Poisson distribution

```
glm (y ~ x, family = poisson)
```

fits the model with **canonical** link $\log \mu_i = \eta_i$ (i.e. g is the log link function). However, to use a **non-canonical** link we need to specify the link argument in family, e.g. for the Poisson distribution

```
glm(y ~ x, family = poisson(identity))
```

fits a model with a **non-canonical** link $\mu_i = \eta_i$ (i.e. g is the identity function).

Example

- Textile data — Poisson regression with identity link.

*****Example — Poisson Regression in R/S-PLUS*****

5.4 Analysis of deviance

Definition of deviance: The **deviance** associated with a model ω is given by

$$D = -2 \log LR = -2 \log \frac{\max L(\text{under model } \omega)}{\max L(\text{under model with } n \text{ parameters, } \Omega)}$$

Use ω to denote the model under consideration, and use Ω to denote the **maximal** or **saturated** model with n parameters. The maximal model Ω has $\widehat{\mu}_i = y_i$, i.e. the model gives a perfect fit since there are n parameters and n observations. The statistic D is the (general) likelihood ratio test statistic. We could also write D in terms of the difference of log likelihoods:

$$D = -2\left\{l(\text{fitted model}, \omega) - l(\text{maximal model or saturated model}, \Omega)\right\}$$

[1] See http://www.r-project.org/

Example

- Deviance for Poisson family, log link. $Y_1, \ldots, Y_n$ independent with $Y_i \sim Po(\lambda_i)$.

$$l(\beta; y_1, \ldots, y_n) = \sum y_i \log \lambda_i - \sum \lambda_i - \sum \log(y_i!) = \sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!)$$

Model Ω: For the saturated (maximal) model we have $\widehat{\mu}_i = y_i$, i.e. best possible fit, and the log likelihood for this maximal model is

$$l(\text{maximal model}, \Omega; y_1, \ldots, y_n) = \sum y_i \log y_i - \sum y_i - \sum \log(y_i!)$$

Model ω: For the fitted model, $g(\widehat{\mu}_i) = \mathbf{x}_i^T \widehat{\beta}$ with $g \equiv \log$, as the **log link function** is assumed, i.e. $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T \widehat{\beta}) = \exp\{\mathbf{x}_i^T \widehat{\beta}\}$, and

$$l(\widehat{\beta}; y_1, \ldots, y_n) = \sum y_i \log \widehat{\mu}_i - \sum \widehat{\mu}_i - \sum \log(y_i!)\ .$$

Thus, the deviance for model ω is

$$D = -2\left\{l(\widehat{\beta}) - l(\text{maximal model}, \Omega)\right\} = -2\sum_{i=1}^{n}\left\{y_i \log\frac{\widehat{\mu}_i}{y_i} + (y_i - \widehat{\mu}_i)\right\}.$$

5.4.1 Testing subsets of parameters

Consider two models:

$$\begin{aligned} \omega &: \quad \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q \\ \Omega &: \quad \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q + \ldots + \beta_p x_p \quad (p > q). \end{aligned}$$

To test

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0,$$

we use the distribution of the **change in deviance** under H_0:

$$D_\omega - D_\Omega \sim \chi^2_{p-q} \quad (p > q).$$

This is obtained by applying the LR test of

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0 \text{ for at least one } q < i \leq p,$$

since

$$D_\omega - D_\Omega = -2\{l(\widehat{\beta}_\omega) - l(\text{maximal model})\} + 2\{l(\widehat{\beta}_\Omega) - l(\text{maximal model})\} = -2\log LR,$$

where

$$LR = \frac{\max L(\text{under restricted model } \omega)}{\max L(\text{under full model } \Omega)}.$$

Example

- Poisson Deviance. Testing constant mean.

  Assume a log link, $\log\lambda_i = \eta_i$ where $\lambda_i = E(Y_i) = \mu_i$

$$\begin{aligned} \omega &: \eta_i = \alpha \quad \text{(common mean response)} \\ \Omega &: \eta_i = \alpha_i. \end{aligned}$$

  Under ω: $\log\lambda_i = \alpha$ or $\lambda_i = e^{\alpha} = \gamma$ say
  Fit model by MLE, i.e. $\widehat{\gamma} = \frac{\sum y_i}{n} = \bar{y}$, and $\widehat{\mu}_i = \widehat{\gamma} = \bar{y}$. Therefore, the deviance for model ω is

$$D_{\omega} = -2\{l(\widehat{\gamma}) - l(\text{maximal model})\} \quad \text{where} \quad l(\widehat{\gamma}) = \sum y_i \log\widehat{\mu}_i - \sum \widehat{\mu}_i.$$

  Under Ω: $\widehat{\mu}_i = y_i$.

  Change in deviance is given by

$$D_{\omega} - D_{\Omega} = -2\left(\sum y_i \log\bar{y} - \sum \bar{y} - \sum y_i \log y_i + \sum y_i\right) = 2\sum y_i \log\frac{y_i}{\bar{y}}.$$

  This may be written in the form of o (observed) and e (expected under H_0):

$$-2\log LR = 2\sum o \log\left(\frac{o}{e}\right).$$

  Distribution of change in deviance, under ω (H_0), is $D_{\omega} - D_{\Omega} \sim \chi^2_{n-1}$.

5.5 Residuals

Ordinary residuals $y_i - \widehat{\mu}_i$ are not used in a GLM (generally) since they have non-constant variance. Two widely used residuals are:

(i) Pearson residuals

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}},$$

where $V(\mu_i) = \text{var}(Y_i)$ is the variance function in terms of μ_i.

Example

- Poisson family.

  Since $V(\mu_i) = \text{var}(Y_i) = \mu_i$, the ith Pearson residual is given by $r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$.

(ii) Deviance residuals

$$d_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{\text{deviance associated with } y_i}$$

In R/S-PLUS these may be obtained from a `glm` object using:

1. `residuals(object.glm, type = 'pearson')` for Pearson residuals.
2. `residuals(object.glm, type ='deviance')` for Deviance residuals.

As in regression, residuals may be used to check the data for outliers, and/or model adequacy.

5.6 Logistic models for binomial data

5.6.1 Tolerance distributions: link functions

Suppose that

$$Y_i \sim \text{Binomial}(m_i, \pi_i) \qquad (i = 1, \ldots, n),$$

where the probability of 'success' for an observation in the ith group, π_i, is given by

$$g(\pi_i) = \eta_i = \mathbf{x}_i^T \beta$$

and $\mathbf{x}_i^T$ is the value of the ith explanatory variable, i.e. ith row of model matrix X.

The curve for π is of sigmoid form and therefore it is natural to model it by a cdf F — this cdf is called a **tolerance distribution**.

Some possibilities for F are:

	Model	$F(\eta)$	F cdf distribution
(a)	Probit	$\pi = \Phi(\eta)$	N(0,1)
** (b)	Logistic	$\pi = \frac{1}{1+e^{-\eta}}$	logistic
(c)	Complementary log-log	$\pi = 1 - \exp(-\exp(\eta))$	extreme value

In the GLM framework for each of these tolerance distributions we have a corresponding link function:

$$\begin{aligned}
\text{(a)}\quad g(\pi) &= \Phi^{-1}(\pi) = \eta \quad \textbf{Probit link} \\
{**}\text{(b)}\quad g(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \eta \quad \textbf{Logit link} \\
\text{(c)}\quad g(\pi) &= \log(-\log(1-\pi)) = \eta \quad \textbf{Complementary log} - \textbf{log link}
\end{aligned}$$

Link (b) is the canonical link function, and we consider this particular form of the binomial model in the following section.
In R/S-PLUS, to fit a binomial model with link (b) use:

```
glm(cbind(y,m-y) ~ x, family = binomial (logit))
```

`cbind(y,m-y)` is a two column matrix with number of successes in first column and number of failures in second column.

5.6.2 Logistic model: logit link function

Suppose now that $Y_i \sim \text{Binomial}(m_i, \pi_i) \quad (i = 1, \ldots, n)$ where

$$\text{logit}\, \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T \beta$$

This is a GLM since it has three elements:

- Model matrix containing rows $\mathbf{x}_i^T$, and coefficients β.
- The link function is the logit function, i.e. $\text{logit}\, \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$, where $\eta_i = \mathbf{x}_i^T \beta$ is the linear component of the model, and $\pi_i = \text{E}(Y_i/m_i)$. This is the 'systematic part' of the model.

- $Y_i \sim \text{Binomial}(m_i, \pi_i)$ the 'random part' of the model, and binomial is a member of the exponential family.

The log likelihood function is

$$l(\pi_1, \ldots, \pi_n; y_1, \ldots, y_n) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \text{constant} \right\}$$

where π_i is given by logit $\pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T \beta$

The MLEs of π_i, for the saturated model are $p_i = \frac{Y_i}{m_i}$, i.e. the observed proportion of 'success' responses associated with the ith row $\mathbf{x}_i^T$ of $\mathbf{X}$. Thus, for the saturated model $\widehat{\pi}_i = p_i$.

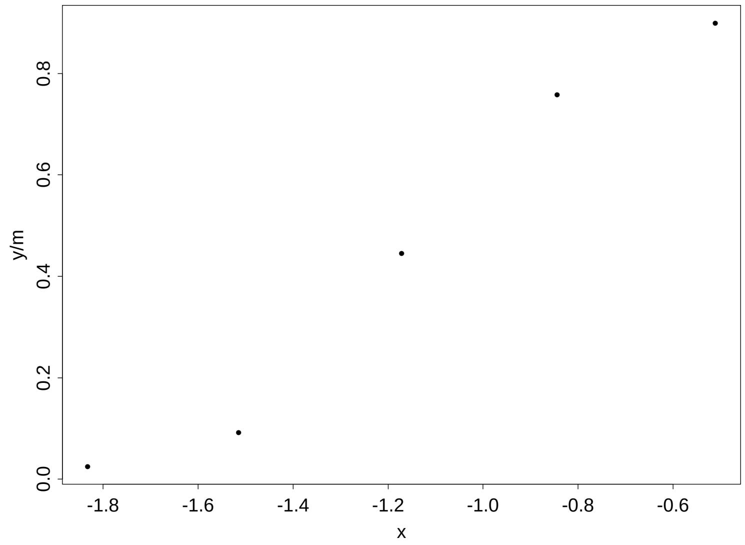
Example

- Dose-response curve. Weevil data set.

  Five doses of an insecticide (*Malathion*) were applied to granary weevils. For each dose (d_i), the number of insects (m_i) receiving that level of dose and the number killed (y_i) were recorded.

Group	Dose	m_i	y_i	p_i
1	0.16	120	3	3/120
2	0.22	120	11	11/120
3	0.31	119	53	53/119
4	0.43	120	91	91/120
5	0.60	119	107	107/119

  Take $x = \log(\text{Dose})$ as the explanatory variable.



$\widehat{\pi}_i$ = estimate of prob $\pi_i = p_i$
Number killed $Y_i \sim \text{Binomial}(m_i, \pi_i)$
where $\pi_i = \Pr$ (insect killed $|x_i$).

5.6.3 Estimation

The binomial model implies that for the ith observation the log likelihood contribution is

$$l(y_i;\pi_i) = y_i\eta_i + m_i\log(1-\pi_i) \quad \text{(``the random part'')}$$

where

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = g(\pi_i)$$

if there are m_i trials and y_i successes. The systematic or regression part of the model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \mathbf{x}_i^T\beta = \sum_{j=1}^{p} x_{ij}\beta_j$$

or $\eta = \mathbf{X}\beta$. This implies that

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}$$

$$\begin{aligned} l(\beta) &= \sum_i \left[y_i(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p) - m_i\log(1+e^{\mathbf{x}_i^T\beta})\right] \\ \frac{dl(\beta)}{d\beta_j} &= \sum_i \left[y_i x_{ij} - m_i\frac{x_{ij}e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\right] \\ &= \sum_i [y_i - m_i\pi_i(\beta)]x_{ij} \end{aligned}$$

The ML equations for $\widehat{\beta}$ are

$$\frac{dl(\beta)}{d\beta_j} = 0$$

Use iterative weighted least squares to estimate β by MLE:

$$\widehat{\beta} = (X^TWX)^{-1}X^TWz$$

Here, using the general theory developed above, we have $W = \text{diag}(w_{ii})$ where

$$w_{ii} = \frac{m_i}{\pi_i(1-\pi_i)}\left(\frac{\partial\pi_i}{\partial\eta_i}\right)^2,$$

and $z = (z_1,\ldots,z_n)$ with

$$z_i = \eta_i + \left(\frac{y_i - m_i\pi_i}{m_i}\right)\left(\frac{\partial\eta_i}{\partial\pi_i}\right).$$

Note we are using Y_i/m_i as the response (not Y_i).

The (asymptotic) estimated variance-covariance matrix for the MLEs is given by

$$\text{var}(\widehat{\beta}) = (X^TWX)^{-1}$$

with the (j,k)th element of (X^TWX) given by

$$(X^TWX)_{jk} = \sum_{i=1}^{n} m_i\pi_i(1-\pi_i)x_{ij}x_{ik}$$

5.6.4 Analysis of deviance

$$\begin{aligned} &\text{Consider the full model } \Omega \quad g(\pi) = \mathbf{X}\beta \quad \text{with } \widehat{\pi}_\Omega = g^{-1}(\mathbf{X}\widehat{\beta}_\Omega) \\ &\text{and the reduced model } \omega \quad g(\pi) = \mathbf{X}_\omega \beta_\omega \quad \text{with } \widehat{\pi}_\omega = g^{-1}(\mathbf{X}_\omega \widehat{\beta}_\omega) \end{aligned}$$

$$\begin{aligned} l(\widehat{\pi}_\Omega) &= \sum\{y \log \widehat{\pi}_\Omega + (m-y)\log(1-\widehat{\pi}_\Omega)\} \\ l(\widehat{\pi}_\omega) &= \sum\{y \log \widehat{\pi}_\omega + (m-y)\log(1-\widehat{\pi}_\omega)\} \end{aligned}$$

Therefore, the change in deviance (or LRT) statistic is

$$\lambda = -2\log LR = 2[l(\widehat{\pi}_\Omega) - l(\widehat{\pi}_\omega)] = 2\sum\left\{y \log \frac{\widehat{\pi}_\Omega}{\widehat{\pi}_\omega} + (m-y)\log \frac{1-\widehat{\pi}_\Omega}{1-\widehat{\pi}_\omega}\right\}$$

In the special case when the full model is the **saturated** model with the # parameters = # observed values of y. Then clearly $\widehat{\pi}_\Omega = \mathbf{p}$, i.e. the MLE of the true probabilities = the observed proportions. The **deviance** for any reduced model is defined as the -2 log(likelihood ratio) statistic for comparing the reduced model with the saturated model.

$$\text{Deviance} = D(\mathbf{p}, \pi_\omega) = 2\sum\left\{y \log\left(\frac{p}{\widehat{\pi}_\omega}\right) + (m-y)\log\left(\frac{1-p}{1-\widehat{\pi}_\omega}\right)\right\}$$

This can be used directly (as in Poisson case), since it does not involve nuisance parameters, for **analysis of deviance**:

$$D_\omega - D_\Omega \sim \chi^2_{p-q} \qquad (p > q)$$

to test $\beta_{q+1} = \ldots = \beta_p = 0$

$$\begin{aligned} D_\Omega &: \text{deviance for model } \eta_i = \beta_1 x_1 + \ldots + \beta_p x_p \\ D_\omega &: \text{deviance for model } \eta_i = \beta_1 x_1 + \ldots + \beta_q x_q \end{aligned}$$

We can also test for **nonlinearity** using

$$D_\Omega \sim \chi^2_{n-p}$$

D_Ω : deviance of model under consideration.
— cf lack of fit in regression. However, see the comments below on the (asymptotic) distribution of deviance.

Example

- Binomial $Y_i \sim Bin(m_i, \pi_i)$, $i = 1, \ldots, n$, with logit link, i.e.

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i.$$

In this example we are modelling $\pi_i = P(\text{killed}|x_i)$.

*****Example — R/S-PLUS Logistic Regression*****

ML estimates of (α, β) are: $\widehat{\alpha} = 4.889407$ and $\widehat{\beta} = 4.538052$.

To test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ use analysis of deviance:

$$D_\omega - D_\Omega \sim \chi^2_1$$

D_ω deviance for model logit $(\pi_i) = \alpha$
D_Ω deviance for model logit $(\pi_i) = \alpha + \beta x_i$

Change in deviance is

$$D_\omega - D_\Omega = 341.5 \quad \text{(given in R/S-PLUS output)}$$

$$\chi^2_1(5\%) = 3.84$$

$$\Rightarrow \text{reject } H_0 : \beta = 0$$

Consider the $2 \times n$ table

					Regressors
	#"Successes"	#"Failures"	Total	Proportion	$\mathbf{x}_1 \cdots \mathbf{x}_p$
1	y_1	$m_1 - y_1$	m_1	p_1	
2	y_2	$m_2 - y_2$	m_2	p_2	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	y_n	$m_n - y_n$	m_n	p_n	

For each of the $2n$ cells we can call the observed frequency o $(= y_i$ or $m_i - y_i)$ and the estimated expected frequency e $(= m_i\widehat{\pi}_i$ or $m_i(1 - \widehat{\pi}_i)$ where $\widehat{\pi}_i$ is a function of $\widehat{\beta}$)

Then Deviance can be written

$$D = 2\sum o \log\left(\frac{o}{e}\right)$$

where the summation is now over all $2n$ cells of the table. It can easily be shown that when all m_i are large, this is approximately equal to the X^2 statistic

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Both of these statistics have distributions which are asymptotically (as the $m_i \to \infty$) χ^2_{n-q}. Here $n =$ dimension of the saturated model and $q =$ dimension of the reduced model $=$ rank of the $\mathbf{X}$ matrix.

In the special case when $\mathbf{X}$ consists of the single column $\mathbf{1}$, we have the homogeneity model $\pi = \pi_0\mathbf{1}$, or $\eta = \beta_0\mathbf{1}$ where $\beta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right)$.

The MLE of π_0 under this reduced model is

$$\widehat{\pi}_0 = \frac{\sum y_i}{\sum m_i} = \frac{y_0}{m_0}$$

and the estimated expected frequencies in the ith row are

$$m_i\widehat{\pi}_0 = \frac{m_i y_0}{m_0} \text{ and } m_i(1 - \widehat{\pi}_0) = \frac{m_i(m_0 - y_0)}{m_0}$$

Then the χ^2 statistic is the familiar contingency table statistic for testing the hypothesis

$$\pi_1 = \pi_2 = \cdots = \pi_n$$

This and the deviance $2\sum o \log\left(\frac{o}{e}\right)$ both $\sim \chi^2_{n-1}$ (approximately if the m_i's are large).

Distribution of the deviance

In general -2 log(likelihood ratio) is approximately χ^2_{p-q} as $n \to \infty$. However, the deviance is defined as the -2 log(likelihood ratio) statistic for the special case of testing the reduced model against the **saturated** model with $p = n$. In considering the asymptotic distribution of the deviance two very different kinds of limit can be considered

1. Keep n fixed and let each $m_i \to \infty$
2. Let $n \to \infty$ with m_i not necessarily large

Under (1) our general result will hold and the deviance $\sim \chi^2_{n-q}$. Under (2) the number of parameters in the saturated model $\to \infty$ as $n \to \infty$ and so general ML theory does not hold: Deviance does not have a χ^2 distribution. As an extreme case, consider all $m_i = 1$. Then it is easy to show that

$$D = -2\sum \{\widehat{\pi} \log \widehat{\pi} + (1-\widehat{\pi})\log(1-\widehat{\pi})\}$$

Since it depends only on the fitted probabilities $\widehat{\pi}$, and not on the differences $y_i - \widehat{\pi}_i$, it is clear that D cannot tell us anything about goodness of fit.

So for logistic regression models tests of goodness of fit using the statistic

$$D = 2\sum o \log\left(\frac{o}{e}\right)$$

or the approximate equivalent

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

and comparing these with χ^2_{n-q} are only valid if the m_i are large.

On the other hand the **change** in deviance between 2 non-saturated models of rank p and q will have a χ^2_{p-q} distribution by the general theorem for likelihood ratio tests. So starting with a given maximal model of rank p say, model selection strategies analogous to those for normal regression models (Backwards Elimination, etc) can be followed, using a sequence of χ^2 tests instead of F tests. However once an acceptable model has been chosen its goodness of fit may have to be assessed using graphical methods for residuals, rather than a single ‘omnibus’ test of deviance against χ^2_{n-q}.

5.6.5 Residuals for logistic regression

If we fit a model we can calculate the fitted values:

$$\widehat{\pi}_i = g^{-1}(\mathbf{x}_i^T\widehat{\beta}) \quad (i = 1,\ldots,n)$$

From these fitted values we can calculate residuals. Residuals for logistic regression models can be defined in terms of the contribution to either χ^2 or D of the ith row of the $2 \times n$ table of successes and failures.

$$\text{‘Pearson residual’} \quad r_i = \frac{y_i - m_i\widehat{\pi}_i}{\sqrt{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

$$\text{‘Deviance residual’} \quad d_i = \operatorname{sign}(y_i - m_i\widehat{\pi}_i)\sqrt{2\left\{y_i \log\left(\frac{p_i}{\widehat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{1-p_i}{1-\widehat{\pi}_i}\right)\right\}}$$

5.7 Log-linear models for Poisson data

An application of equal importance to logistic regression for binomial random variable is log-linear regression for Poisson random variables. Here we have for a single y

$$l(\mu;y) = y\log(\mu) - \mu - \log(y!)$$

The canonical link is $\theta = \log(\mu)$. Therefore, if we use this in a GLM to link the mean μ_i to the linear component η_i we have the log link

$$\eta = \log(\mu) = \mathbf{X}\beta$$

Note that for the log link function we have fitted values

$$\widehat{\mu_i} = g^{-1}(\mathbf{x}_i^T\widehat{\beta}) = \exp(\mathbf{x}_i^T\widehat{\beta}).$$

5.7.1 Estimation

The method of scoring iterative equation, given above is

$$X^TWX\beta = X^TW\mathbf{z}$$

where $W = \text{diag}(\mu_i)$ because $\text{var}(Y_i) = \mu_i$ for the Poisson distribution and $\mathbf{z}$ has elements

$$z_i = \log(\mu_i) + \frac{y_i - \mu_i}{\mu_i}$$

5.7.2 Analysis of deviance

We obtained the deviance for Poisson family above as

$$D = -2\left\{l(\widehat{\beta}) - l(\text{maximal model})\right\} = -2\sum_{i=1}^{n}\left\{y_i\log\frac{\widehat{\mu}_i}{y_i} + (y_i - \widehat{\mu}_i)\right\}$$

Example

- Contingency table, r rows and c columns, with the cell entries following a Poisson distribution.

  Model ω: $\eta_i = \mu + \alpha_i + \beta_j$

  Model Ω: $\eta_i = \mu + \alpha_i + \beta_j + \gamma_{ij}$

  To test independence, we test the null hypothesis $H_0 : \gamma_{ij} = 0$.

  This is similar to the Poisson example given above. The test statistic for analysis of deviance is of the form

  $$D = 2\sum o\log(\frac{o}{e})$$

  where o denotes the observed value, and e denotes the expected value under model ω.

  Compare with a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

5.7.3 Residuals for Poisson regression

The Pearson residuals are given by

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}},$$

where $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T \widehat{\beta})$.

UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Generalized Linear Mixed Models (GLMM) **Semester 1, 2022–2023**

1 Introduction

We consider an extension of linear and generalized linear models to include **random effects**. Up to now we have considered all the random variation in the response $\mathbf{y}$ to result from the 'family' assumed. However, it is often useful to regard some of the unknown parameters in a linear model or generalized linear model as random, e.g. regard some of the unknown parameters as being chosen at random from a larger population. Models containing both **random** and **fixed** parameters are called **mixed models**. The 'lme4' package in R provides software for fitting these models.

Example Variation in the yield of a dyestuff
An experiment was carried out to investigate how much of the variation in yield in the manufacture of a dyestuff was due to the variation between batches in one of the raw materials. Five laboratory determinations of the yield were made for each of six randomly chosen batches of raw material, with the results given below.

	Batch					
	1	2	3	4	5	6
Yields	1545	1540	1595	1445	1595	1520
	1440	1555	1550	1440	1630	1455
	1440	1490	1605	1595	1515	1450
	1520	1560	1510	1465	1635	1480
	1580	1495	1560	1545	1625	1445

For this study, we may not be interested in the particular six batches used; they are merely representatives of a larger population about which we want to make inferences. We can model the effects of the batches as forming a random sample from some distribution, usually a Normal distribution.

If y_{jk} denotes the kth yield measurement for the jth batch then we might assume a model

$$y_{jk} = \mu + a_j + e_{jk} \quad (j = 1, \ldots, g; k = 1, \ldots, m_j), \tag{1}$$

with $g = 6$ and $m_j = 5$. Here μ and a_j denote respectively the expected value of the yield (over batches as well as determinations) and the effect of the jth batch. The a_j are called *random effects*, while μ is a *fixed effect*. [We follow a convention of using Roman and Greek letters for random and fixed effects respectively.] The random variables a_j and e_{jk} might be assumed uncorrelated with expectations zero and variances σ_a^2 and σ^2 respectively. To make inferences about the parameters μ, σ_a^2 and σ^2, we might also take the a_j and e_{jk} to be jointly Normally distributed.

Under this *one-way random-effects* model, we have (conditional on μ, but not on the a_j)

$$\operatorname{var}(y_{jk}) = \sigma_a^2 + \sigma^2, \quad \operatorname{cov}(y_{jk}, y_{jk'}) = \sigma_a^2 \quad (k \neq k'), \quad \operatorname{var}(\bar{y}_{j\cdot}) = \sigma_a^2 + \sigma^2/m_j. \tag{2}$$

2 Which effects should be treated as random?

It is not always clear which of the effects in a statistical model should be treated as random: the decision may depend on the purpose of the analysis as well as the nature of the study. The published advice on when effects should be taken to be random is contradictory. To decide which effects to treat as random, it can be useful to imagine repeating the study: the 'fixed' effects are those whose levels would be the same in the new experiment, and the 'random' effects are those whose levels would be different from before. Thus factors that would be treated as fixed include sex, age groupings, disease types, medical treatments, and measurement times in a repeated-measures experiment. Factors that would usually be treated as random include experimental animals (including humans), years (for studies of annual crops) and batches of raw material used in an experiment. Some that could be treated as either fixed or random are crop varieties and expert assessors.

Some other examples of where random effects might arise are as follows.

1. In educational research, sources of variation in examination performance may include pupils, classes, schools and local authorities. Suitable models would be *hierarchical* or *multi-level* or have *nested* effects, and might include random effects at each level. Fixed effects might include factors relating to teaching methods or class organisation (such as age groups), and covariates (such as pupils' birth dates) could be incorporated at the appropriate level.

2. Similarly, sample surveys may be organised by choosing a sample of local authorities, then parts of the local authority areas, such as post-code sectors, and then individual households. The inferences from such surveys should take account of the possible sources of variation.

3. If measurements are made on related animals, such as litter-mates, we expect that they will be positively correlated within the groups because of the effects of shared genes and shared maternal environment. Conversely, competition for food within litters may lead to a negative correlation between body weights: assuming additive litter effects is then not appropriate.

4. When blocking is used in an experimental design, it may be reasonable to take block effects as random. With incomplete blocks, estimates of treatment differences can be based on the totals over the blocks as well as on differences within blocks. The best combination of these *interblock* and *intrablock* estimates depends on the ratio of the residual and block variances.

3 Linear component of Mixed Models

Linear Models: The usual fixed-effects linear model for an n-vector $\mathbf{y}$ of responses is $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$ given the design matrix $\mathbf{X}$. This model can be generalized in various ways to a *mixed* model, that is one containing both fixed and random effects. One of the simpler ways is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \ldots + \mathbf{Z}_q\mathbf{u}_q + \mathbf{e}, \tag{3}$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_q$ are vectors of random effects with corresponding design matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_q$. Note that the residual vector $\mathbf{e}$ can be included as a vector of random effects (with design matrix $\mathbf{I}_n$); other $\mathbf{u}$-vectors might represent main effects and interaction effects. If $\mathbf{Z}_s$ is $n \times p_s$ $(s =$

$1,\ldots,q$) then $\mathbf{u}_1,\ldots,\mathbf{u}_q$ are assumed to be uncorrelated with each other and with $\mathbf{e}$, and to have zero expectations and variance matrices $\sigma_1^2\mathbf{I}_{p_1},\ldots,\sigma_q^2\mathbf{I}_{p_q}$.

Under (3), the response vector $\mathbf{y}$ has expectation $\mathbf{X}\beta$, but its variance matrix becomes, instead of $\sigma^2\mathbf{I}_n$,

$$\Sigma = \mathrm{var}(\mathbf{y}\,|\,\mathbf{X}) = \sigma_1^2\mathbf{Z}_1\mathbf{Z}_1^T + \ldots + \sigma_q^2\mathbf{Z}_q\mathbf{Z}_q^T + \sigma^2\mathbf{I}_n\,. \tag{4}$$

Model (3) includes the model assumed in the example considered above, as well as more complex factorial models. Also included are regression models in which the slopes and intercepts are random: including an interaction between a random factor and a covariate is interpreted as allowing the regression coefficients to vary between levels of the factor. Such models are used with time as a covariate to analyse repeated-measures data.

Unless the data are balanced, the estimate of a fixed effect depends on whether other effects are treated as fixed or random.

Generalized Linear Models: In the GLM context we have a **linear component**

$$\eta = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \ldots + \mathbf{Z}_q\mathbf{u}_q\,, \tag{5}$$

with a link function $g(\mu) = \eta$, where the response Y has a distribution that is a member of the exponential family.

4 ML and REML estimation

Under the assumption of Normality, the maximum likelihood (ML) estimates of β, $\sigma_1^2,\ldots,\sigma_q^2$ and σ^2 in (3) maximize the log-likelihood, which is

$$-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\Sigma^{-1}(\mathbf{y}-\mathbf{X}\beta)\,. \tag{6}$$

For given σ, the ML estimate $\widehat{\beta}$ of β satisfies the (weighted least-squares) equation

$$\mathbf{X}^T\Sigma^{-1}\mathbf{X}\widehat{\beta} = \mathbf{X}^T\Sigma^{-1}\mathbf{y}\,,$$

but maximizing (6) with respect to $\sigma_1^2,\ldots,\sigma_q^2$ and σ^2 usually requires iterative estimation (including checking for negative variance estimates).

Approximate tests of hypotheses about the *fixed* effects in (3) can be based on differences in the maximized log-likelihood (i.e. on the χ^2 approximation for differences between deviances).

For REML estimation, the likelihood is based not on $\mathbf{y}$ but on a vector of linear functions of $\mathbf{y}$ chosen to have zero expectations under the model. If $\mathbf{X}$ has rank r then an $n\times(n-r)$ matrix $\mathbf{K}$ of full rank can be found satisfying $\mathbf{K}^T\mathbf{X} = \mathbf{O}$. The $(n-r)$-vector $\mathbf{K}^T\mathbf{y}$ then has distribution $N_{n-r}(\mathbf{0},\mathbf{K}^T\Sigma K)$, which does not depend on β. The REML estimates maximize the corresponding likelihood, but do not depend on the particular choice of $\mathbf{K}$.

Under REML, tests of hypotheses about *fixed* effects cannot use the maximized log-likelihood directly because omitting elements of β changes $\mathbf{X}$ and hence changes the matrix $\mathbf{K}$ used to construct the REML likelihood. For the same reason, an AIC criterion based on the REML likelihood cannot be used to compare models having different fixed effects. However, inferences about individual coefficients in a model can be based on the approximate variance matrix of the coefficients: individual estimates divided by their estimated standard errors are compared with $N(0,1)$.

5 The R functions `lmer` and `glmer`

The function lmer can be used to estimate fixed effects (including regression coefficients) and variance components in mixed linear models with the form defined in (3). The method of estimation used is REML (by default) or ML (using the option REML=FALSE). Similarly, the function glmer can be used to estimate fixed effects (including regression coefficients) in GLMMs with the form defined in (5). These function are provided in the lme4 package, described at http://lme4.r-forge.r-project.org/[1].

The model formulae used in lmer include fixed effects as in lm or aov. A simple random-effect term corresponding to a factor randfactor, say, is denoted by (1 | randfactor) in an lmer formula.

Optional arguments for lmer include data, subset and na.action. The result of using lmer is an object of class mer: methods for mer objects include summary, fitted, resid, coef and anova. Note that the statistics such as AIC displayed when anova is used are based on ML estimates, even if REML has been used for estimation: see the final paragraph of Section 4.

Example Variation in the yield of a dyestuff (continued)
The file dyestuff.txt contains columns with yield and batch. To fit model (1) we could use

```
dyestuff.mer <- lmer(yield ~ 1 + (1 | as.factor(batch)))
```

The output from summary(dyestuff.mer) includes estimates of the variance components σ^2 and σ_a^2, and of the fixed effect, the overall expected yield μ, as follows.

```
Linear mixed model fit by REML
Formula: yield ~ 1 + (1 | as.factor(batch))

   AIC   BIC logLik deviance REMLdev
 325.7 329.9 -159.8    327.4   319.7

Random effects:
 Groups           Name        Variance Std.Dev.
 as.factor(batch) (Intercept) 1764.0   42.001
 Residual                     2451.3   49.510
Number of obs: 30, groups: as.factor(batch), 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38   78.81
```

[1] R TIP: you can use install.packages('lme4') to install the 'lme4' package if not already installed, and then library(lme4) before using the functions lmer/glmer.

Example Contagious bovine pleuropneumonia (CBPP)

Contagious bovine pleuropneumonia (CBPP) is a major disease of cattle in Africa, caused by a mycoplasma. This dataset describes the serological incidence of CBPP in zebu cattle during a follow-up survey implemented in 15 commercial herds located in the Boji district of Ethiopia. The goal of the survey was to study the within-herd spread of CBPP in newly infected herds. Blood samples were quarterly collected from all animals of these herds to determine their CBPP status. These data were used to compute the serological incidence of CBPP (new cases occurring during a given time period). Some data are missing (lost to follow-up).

Below is an analysis with `glmer`.

```
library(lme4)
## response as a matrix
(m1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
family = binomial, data = cbpp))
## response as a vector of probabilities and usage of argument "weights"
m1p <- glmer(incidence / size ~ period + (1 | herd), weights = size,
family = binomial, data = cbpp)
## Confirm that these are equivalent:
stopifnot(all.equal(fixef(m1), fixef(m1p), tolerance = 1e-5),
all.equal(ranef(m1), ranef(m1p), tolerance = 1e-5))checkConv
13
## GLMM with individual-level variability (accounting for overdispersion)
cbpp$obs <- 1:nrow(cbpp)
(m2 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd) + (1|obs),
family = binomial, data = cbpp))
```

Results of analysis

Output from the `glmer` analysis is given below.

```
> library(lme4)
Loading required package: Matrix

> (m1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
+ family = binomial, data = cbpp))

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: cbind(incidence, size - incidence) ~ period + (1 | herd)
   Data: cbpp
     AIC      BIC   logLik deviance df.resid
194.0531 204.1799 -92.0266 184.0531       51

Random effects:
 Groups Name        Std.Dev.
 herd   (Intercept) 0.6421
Number of obs: 56, groups:  herd, 15

Fixed Effects:
(Intercept)      period2      period3      period4
    -1.3983      -0.9919      -1.1282      -1.5797
```

```
> m1p <- glmer(incidence / size ˜ period + (1 | herd), weights = size,
+ family = binomial, data = cbpp)

> stopifnot(all.equal(fixef(m1), fixef(m1p), tolerance = 1e-5),
+ all.equal(ranef(m1), ranef(m1p), tolerance = 1e-5))checkConv

> 13
[1] 13

> cbpp$obs <- 1:nrow(cbpp)
> (m2 <- glmer(cbind(incidence, size - incidence) ˜ period + (1 | herd) +
                  (1|obs+ family = binomial, data = cbpp))

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: cbind(incidence, size - incidence) ˜ period + (1 | herd) + (1 |
    obs)
   Data: cbpp

     AIC      BIC   logLik deviance df.resid
186.6383 198.7904 -87.3192 174.6383       50

Random effects:
 Groups Name        Std.Dev.
 obs    (Intercept) 0.8911
 herd   (Intercept) 0.1840

Number of obs: 56, groups:  obs, 56; herd, 15
Fixed Effects:
(Intercept)      period2      period3      period4
     -1.500       -1.226       -1.329       -1.866
```