APTS – Statistical Modelling – Preliminary Material

Helen Ogden[†]
University of Southampton

In order to get the most out of the APTS module on Statistical Modelling, students should have, at the start of the module, a sound knowledge of the principles of statistical inference and the theory of linear and generalised linear models. Students should also have some experience of practical statistical modelling in `R`.

The following reading and activities are recommended for students to (re)-familiarise themselves with these areas.

Statistical inference: It is recommended that students (re)-read the notes of the APTS module on Statistical Inference, available from `www.apts.ac.uk`, and complete the assessment exercise (if they have not already done so). No further material is provided here.

Linear and generalised linear models: A student who has covered Chapters 8 and 10.1-10.4 of *Statistical Models* by A. C. Davison (Cambridge University Press, 2003) will be more than adequately prepared for the APTS module. For students without access to this book, the main theory is repeated below. The inference methodology described is largely based on classical statistical theory. Although prior experience of Bayesian statistical modelling would be helpful, it will not be assumed.

Preliminary exercises: Eight relatively straightforward exercises are included throughout these notes.

Practical statistical modelling in `R`: Some short practical exercises are also provided at the end of these notes to enable students to familiarise themselves with statistical modelling in `R`.

1 Linear Models

1.1 Introduction

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. We denote the n observations of the response variable by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. In a statistical model, these are assumed to be observations of *random variables* $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Associated with each y_i is a vector $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables.

Linear models are those for which the relationship between the response and explanatory variables is of the form

$$\begin{aligned} E(Y_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \\ &= \sum_{j=0}^{p} x_{ij}\beta_j \qquad \text{(where we define } x_{i0} \equiv 1\text{)} \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ &= [\boldsymbol{X}\boldsymbol{\beta}]_i, \qquad i = 1, \ldots, n \end{aligned} \tag{1}$$

[†] Preliminary material originally provided by Jon Forster, Anthony Davison, Dave Woods and Antony Overstall

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on $\boldsymbol{x}_i$. The four ways of describing the linear model in (1) are equivalent, but the most ecoonomical is the matrix form

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{2}$$

The $n \times (p+1)$ matrix $\boldsymbol{X}$ consists of known (observed) constants and is called the *design matrix* . The ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$, the explanatory data corresponding to the ith observation of the response. The jth column of $\boldsymbol{X}$ contains the n observations of the jth explanatory variable.

Example 1 The null model

$$E(Y_i) = \beta_0 \qquad i = 1, \ldots, n$$

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \boldsymbol{\beta} = (\,\beta_0\,)$$

Example 2 Simple linear regression

$$E(Y_i) = \beta_0 + \beta_1 x_i \qquad i = 1, \ldots, n$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Example 3 Polynomial regression

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p \qquad i = 1, \ldots, n$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Example 4 Multiple regression

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{i\,p} \qquad i = 1, \ldots, n$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Strictly, the only requirement for a model to be linear is that the relationship between the response variables, $\boldsymbol{Y}$, and any explanatory variables can be written in the form (2). No further specification of the joint distribution of $Y_1, \ldots, Y_n$ is required. However, the linear model is more useful for statistical analysis if we can make three further assumptions:

1. $Y_1, \ldots, Y_n$ are independent random variables.
2. $Y_1, \ldots, Y_n$ are normally distributed.
3. $Var(Y_1) = Var(Y_2) = \cdots = Var(Y_n)$ ($Y_1, \ldots, Y_n$ are homoscedastic). We denote this common variance by σ^2

With these assumptions the linear model completely specifies the distribution of $\boldsymbol{Y}$, in that $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim N\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\ ,\ \sigma^2\right) \qquad i = 1, \ldots, n.$$

Another way of writing this is

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}\ +\ \epsilon_i \qquad i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$ random variables.

A linear model can now be expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\boldsymbol{I}$, (because all $Var(\epsilon_i) = \sigma^2$ and $\epsilon_1, \ldots, \epsilon_n$ are independent implies all $Cov(\epsilon_i, \epsilon_j) = 0$). It follows from (3) that the distribution of $\boldsymbol{Y}$ is multivariate normal with mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and variance covariance matrix $\sigma^2\boldsymbol{I}_n$, *i.e.* $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$.

1.2 Least squares estimation

The regression coefficients $\beta_0, \ldots, \beta_p$ describe the pattern by which the response depends on the explanatory variables. We use the observed data $y_1, \ldots, y_n$ to *estimate* this pattern of dependence.

In least squares estimation, roughly speaking, we choose $\hat{\boldsymbol{\beta}}$, the estimates of $\boldsymbol{\beta}$ to make the estimated means $\hat{E}(\boldsymbol{Y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ *as close as possible* to the observed values $\boldsymbol{y}$, *i.e.* $\hat{\boldsymbol{\beta}}$ minimises the sum of squares

$$\begin{aligned}\sum_{i=1}^{n}[y_i - E(Y_i)]^2 &= \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2\end{aligned}$$

as a function of $\beta_0, \ldots, \beta_p$.

Exercise 1: By differentiating the sum of squares above w.r.t. $\beta_k,\ k = 0, \ldots, p$, show that

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

The least squares estimates $\hat{\boldsymbol{\beta}}$ are the solutions to this set of $p+1$ simultaneous linear equations, which are known as the *normal equations*. If $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (as it usually is) then the least squares estimates are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

The corresponding fitted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$\Rightarrow \qquad \hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \qquad i = 1, \ldots, n.$$

We define the *hat* matrix by $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, so $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. The residuals are

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$$
$$\Rightarrow \qquad e_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \qquad i = 1, \ldots, n.$$

The residuals describe the variability in the observed responses $y_1, \ldots, y_n$ which has not been explained by the linear model. The *residual sum of squares* or *deviance* for a linear model is defined to be

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)^2.$$

It is the actual minimum value attained in the least squares estimation.

Properties of the least squares estimator

1. (**Exercise 2**) Show that $\hat{\boldsymbol{\beta}}$ is multivariate normal with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

2. Assuming that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$ the least squares estimate $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate. This is obvious when one considers the likelihood for a linear model

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right). \tag{4}$$

1.3 Estimation of σ^2

In addition to the linear coefficients $\beta_0, \ldots, \beta_p$ estimated using least squares, we also need to estimate the *error variance* σ^2, representing the variability of observations about their mean.

We can estimate σ^2 using maximum likelihood. Maximising (4) with respect to $\boldsymbol{\beta}$ and σ^2 gives

$$\hat{\sigma}^2 = \frac{D}{n} = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

If the model is correct, then D is independent of $\hat{\boldsymbol{\beta}}$ and

$$\frac{D}{\sigma^2} \sim \chi^2_{n-p-1}$$
$$\Rightarrow E(\hat{\sigma}^2) = \frac{n-p-1}{n}\sigma^2,$$

so the maximum likelihood estimator is biased for σ^2 (although still asymptotically unbiased as $\frac{n-p-1}{n} \to 1$ as $n \to \infty$). We usually prefer to use the unbiased estimator of σ^2

$$s^2 = \frac{D}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2.$$

The denominator $n - p - 1$, the number of observations minus the number of coefficients in the model is called the *degrees of freedom* of the model. Therefore, we estimate the error variance by the deviance divided by the degrees of freedom.

1.4 Inference

It follows from the distribution of $\hat{\boldsymbol{\beta}}$ that

$$\frac{\hat{\beta}_k - \beta_k}{\sigma[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{kk}^{1/2}} \sim N(0,1), \qquad k = 0, \ldots, p.$$

The dependence on unknown σ can be eliminated by replacing σ with its estimate s, in which case it can be shown that

$$\frac{\hat{\beta}_k - \beta_k}{s.e.(\hat{\beta}_k)} \sim t_{n-p-1},$$

where the standard error $s.e.(\hat{\beta}_k)$ is given by

$$s.e.(\hat{\beta}_k) = s[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{kk}^{1/2}.$$

This enables confidence intervals for any β_k to be calculated, or hypotheses of the form H_0: $\beta_k = 0$ to be tested.

The sampling distributions of the fitted values and residuals can be obtained, straightforwardly as

$$\hat{\boldsymbol{y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})$$

and

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2[\boldsymbol{I}_n - \boldsymbol{H}]).$$

The latter expression allows us to calculate *standardised* residuals, for comparison purposes, as

$$r_i = \frac{e_i}{s(1-h_{ii})^{1/2}}.$$

1.5 Prediction

We estimate the mean, $\boldsymbol{x}_+^T\boldsymbol{\beta}$, for Y at values of the explanatory variables given by $\boldsymbol{x}_+^T = (\,1 \quad x_{+1} \quad \ldots \quad x_{+p}\,)$, which may or may not match a set of values observed in the data, using

$$\hat{Y}_+ = \boldsymbol{x}_+^T\hat{\boldsymbol{\beta}}.$$

Then

$$\hat{Y}_+ \sim N(\boldsymbol{x}_+^T\boldsymbol{\beta}, \sigma^2 h_{++})$$

where $h_{++} = \boldsymbol{x}_+^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_+$. Hence confidence intervals for predictive means can be derived using

$$\frac{\hat{Y}_+ - \boldsymbol{x}_+^T\boldsymbol{\beta}}{sh_{++}^{1/2}} \sim t_{n-p-1}.$$

For predicting the actual value $Y_+ = \boldsymbol{x}_+^T\boldsymbol{\beta} + \epsilon_+$, the predictor $\hat{Y}_+$ is also sensible, as $E(\hat{Y}_+ - Y_+) = 0$. Now

$$\hat{Y}_+ - Y_+ \sim N(0, \sigma^2(1 + h_{++})).$$

as $\hat{Y}_+$ and Y_+ are independent. Hence predictive confidence intervals can be derived using

$$\frac{\hat{Y}_+ - Y_+}{s(1+h_{++})^{1/2}} \sim t_{n-p-1}.$$

1.6 Comparing linear models [1]

A pair of *nested* linear models can be compared pairwise using a generalised likelihood ratio test. Nesting implies that the simpler model (H_0) is a special case of the more complex model (H_1). In practice, this usually means that the explanatory variables present in H_0 are a subset of those present in H_1. Let $\Theta^{(1)}$ be the unrestricted parameter space under H_1 and $\Theta^{(0)}$ be the parameter space corresponding to model H_0, *i.e.* with the appropriate coefficients constrained to zero.

Without loss of generality, we can think of H_1 as the model

$$E(Y_i) = \sum_{j=0}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$

with H_0 being the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Now, a *generalised likelihood ratio test* of H_0 against H_1 has a test statistic of the form

$$T = \frac{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(1)}} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\beta},\sigma^2)}{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(0)}} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\beta},\sigma^2)}$$

and rejects H_0 in favour of H_1 when $T > k$, where where k is determined by α, the size of the test.

For a linear model,

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right).$$

This is maximised with respect to $(\boldsymbol{\beta},\sigma^2)$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\sigma^2 = \hat{\sigma}^2 = D/n$. Therefore

$$\begin{aligned}\max_{\boldsymbol{\beta},\sigma^2} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2D}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2\right) \\ &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)\end{aligned}$$

Exercise 3: Let the deviances under models H_0 and H_1 be denoted D_0 and D_1 respectively. Show that the likelihood ratio test statistic T above can be written as

$$T = \left(1 + \frac{p-q}{n-p-1}F\right)^{\frac{n}{2}},$$

where

$$F = \frac{(D_0 - D_1)/(p-q)}{D_1/(n-p-1)}.$$

Hence, the simpler model H_0 is rejected in favour of the more complex model H_1 if F is 'too large'.

As we have required H_0 to be nested in H_1 then, under H_0, F has an F distribution with $p-q$ degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator. To see this, note the analysis of variance decomposition

$$\frac{D_0}{\sigma^2} = \frac{D_0 - D_1}{\sigma^2} + \frac{D_1}{\sigma^2}.$$

[1] It should be noted that this section describes just one method for comparing models. General principles and other methods will be discussed in detail in the APTS module itself.

We know (§1.3) that, under H_0, D_1/σ^2 has a χ^2_{n-p-1} distribution and D_0/σ^2 has a χ^2_{n-q} distribution. It is also true (although we do not show it here) that under H_0, $(D_0 - D_1)/\sigma^2$ and D_0/σ^2 are independent. Therefore, from the properties of the chi-squared distribution, it follows that under H_0, $(D_0 - D_1)/\sigma^2$ has a χ^2_{p-q} distribution, and F has a $F_{p-q,\,n-p-1}$ distribution.

Therefore, H_0 is rejected in favour of H_1 when $F > k$ where k is the $100(1-\alpha)\%$ point of the $F_{p-q,\,n-p-1}$ distribution.

1.7 Model checking

Confidence intervals and hypothesis tests for linear models may be unreliable if all the model assumptions are not justified. In particular, we have made four assumptions about the distribution of $Y_1, \ldots, Y_n$.

1. The model correctly describes the relationship between $E(Y_i)$ and the explanatory variables
2. $Y_1, \ldots, Y_n$ are normally distributed.
3. $Var(Y_1) = Var(Y_2) = \cdots = Var(Y_n)$.
4. $Y_1, \ldots, Y_n$ are independent random variables.

These assumptions can be checked using plots of raw or standardised residuals.

1. If a plot of the residuals against the values of a potential explanatory variable reveals a pattern, then this suggests that the explanatory variable, or perhaps some function of it, should be included in the model.
2. A simple check for non-normality is obtained using a normal probability plot of the ordered residuals. The plot should look like a straight line, with obvious curves suggesting departures from normality.
3. A simple check for non-constant variance is obtained by plotting the residuals $r_1, \ldots, r_n$ against the corresponding fitted values $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \quad i = 1, \ldots, n$. The plot should look like a random scatter. In particular, check for any behaviour which suggests that the error variance increases as a function of the mean ('funnelling' in the residual plot).
4. In general, independence is difficult to validate, but where observations have been collected in serial order, serial correlation may be detected by a lagged scatterplot or correlogram.

Another place where residual diagnostics are useful is in assessing *influence*. An observation is influential if deleting it would lead to estimates of model parameters being substantially changed. Cook's distance C_j is a measure of the change in $\hat{\boldsymbol{\beta}}$ when observation j is omitted from the dataset.

$$C_j = \frac{\sum_{i=1}^{n}\left(\hat{y}_i^{(j)} - \hat{y}_i\right)^2}{ps^2}$$

where $\hat{y}_i^{(j)}$ is the fitted value for observation i, calculated using the least squares estimates obtained from the modified data set with the jth observation deleted. A rule of thumb is that values of C_j greater than $8/(n-2p)$ indicate influential points. It can be shown that

$$C_j = \frac{r_j^2 h_{jj}}{p(1-h_{jj})}$$

so influential points have either a large standardised residual (unusual Y value) or large h_{jj}. The quantity h_{jj} is called the *leverage* and is a measure of how unusual (relative to the other values in the data set) the explanatory data for the jth observation are.

1.8 Bayesian inference for linear models

Bayesian inference for the parameters $(\boldsymbol{\beta}, \sigma^2)$ of a linear model requires computation of the posterior density. Bayes theorem gives us

$$f(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}) \;\propto\; f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2)\, f(\boldsymbol{\beta}, \sigma^2)$$

where the likelihood $f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2)$ is given by (4) as

$$f(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right).$$

Posterior computation is straightforward if the prior density $f(\boldsymbol{\beta}, \sigma^2)$ is conjugate to the likelihood, which, for a linear model, is achieved by the prior decomposition

$$\sigma^{-2} \;\sim\; \text{Gamma}(a_0, b_0), \qquad \boldsymbol{\beta} \,|\, \sigma^2 \;\sim\; N(\boldsymbol{m}_0, \sigma^2\boldsymbol{V}_0)$$

where $(a_0, b_0, \boldsymbol{m}_0, \boldsymbol{V}_0)$ are hyperparameters, whose values are chosen to reflect prior uncertainty about the linear model parameters $(\boldsymbol{\beta}, \sigma^2)$.

With this prior structure, the corresponding posterior distributions are given by

$$\sigma^{-2} \;\sim\; \text{Gamma}(a_0 + n/2, b), \qquad \boldsymbol{\beta} \,|\, \sigma^2 \;\sim\; N(\boldsymbol{m}, \sigma^2\boldsymbol{V})$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}_0^{-1})^{-1}$, $\boldsymbol{m} = \boldsymbol{V}(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{V}_0^{-1}\boldsymbol{m}_0)$ and

$$\begin{aligned} b \;&=\; b_0 + \frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{m}_0\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 - \boldsymbol{m}\boldsymbol{V}^{-1}\boldsymbol{m}\right) \\ &=\; b_0 + \frac{1}{2}\left\{[n-p-1]s^2 + \left[\boldsymbol{m}_0 - \hat{\boldsymbol{\beta}}\right]^T\left[\boldsymbol{V}_0 + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]^{-1}\left[\boldsymbol{m}_0 - \hat{\boldsymbol{\beta}}\right]\right\} \end{aligned}$$

if $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular, where $\hat{\boldsymbol{\beta}}$ and s^2 are the classical unbiased estimators given above.

In applications where prior information about the model parameters $(\boldsymbol{\beta}, \sigma^2)$ is weak, it is conventional to use the vague prior specification given by the improper prior density

$$f(\boldsymbol{\beta}, \sigma^2) \;\propto\; \sigma^{-2}.$$

This corresponds to the conjugate prior above with $a_0 = -(p+1)$, $b_0 = 0$ and $\boldsymbol{V}_0^{-1} = \boldsymbol{0}$.

Exercise 4: Show that, for this choice of hyperparameters, the posterior mean of $\boldsymbol{\beta}$ is given by the least squares estimator $\hat{\boldsymbol{\beta}}$. Show also that, a posteriori, $1/\sigma^2$ has the distribution of $X^2/[s^2(n-p-1)]$ where X^2 has a χ^2_{n-p-1} distribution. Hence show that posterior probability intervals for σ^2 are equivalent to confidence intervals based on the sampling distribution of s^2.

For a longer exercise, show that $(\boldsymbol{\beta} - \boldsymbol{m})/\sigma$ has a multivariate normal posterior marginal distribution, independent of σ^2, and hence that posterior probability intervals for a coefficient β_k are equivalent to the confidence intervals based on the sampling distribution of $(\hat{\beta}_k - \beta_k)/s.e.(\hat{\beta}_k)$ derived in Section 1.4 above.

2 Generalised linear models

2.1 Introduction

The generalised linear model extends the linear model defined in §1.1 to allow a more flexible family of probability distributions.

In a generalised linear model (GLM) the n observations of the response $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ are assumed to be observations of *independent* random variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, which take the same distribution from the exponential family. Hence,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\left(\sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i) \right) \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is the collection of canonical parameters and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ is the collection of dispersion parameters (where they exist). Commonly, the dispersion parameters are known up to, at most, a single common unknown σ^2, and we write $\phi_i = \sigma^2/m_i$ where the m_i represent known weights.

The distribution of the response variable Y_i depends on the explanatory data $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ through the *linear predictor* η_i where

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \\ &= \sum_{j=0}^{p} x_{ij} \beta_j \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ &= [\boldsymbol{X}\boldsymbol{\beta}]_i, \qquad i = 1, \ldots, n \end{aligned}$$

in an exactly analagous fashion to the linear model in §1.1.

The link between the distribution of $\boldsymbol{Y}$ and the linear predictor $\boldsymbol{\eta}$ is provided by the *link function* g,

$$\eta_i = g(\mu_i) \qquad i = 1, \ldots, n$$

where $\mu_i \equiv E(Y_i)$, $i = 1, \ldots, n$. Hence, the dependence of the distribution of the response on the explanatory variables is established as

$$g(E[Y_i]) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n$$

In principle, the link function g can be any one-to-one differentiable function. However, we note that η_i can in principle take any value in $\mathcal{R}$ (as we make no restriction on possible values taken by explanatory variables or model parameters). However, for some exponential family distributions μ_i is restricted. For example, for the Poisson distribution $\mu_i \in \mathcal{R}_+$; for the Bernoulli distribution $\mu_i \in (0, 1)$. If g is not chosen carefully, then there may exist a possible $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ such that $\eta_i \neq g(\mu_i)$ for any possible value of μ_i. Therefore, most common choices of link function map the set of allowed values for μ_i onto $\mathcal{R}$.

Recall that for a random variable Y with a distribution from the exponential family, $E(Y) = b'(\theta)$. Hence, for a generalised linear model

$$\mu_i = E(Y_i) = b'(\theta_i) \qquad i = 1, \ldots, n.$$

Therefore

$$\theta_i = b'^{-1}(\mu_i) \qquad i = 1, \ldots, n$$

and as $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, then

$$\theta_i = b'^{-1}(g^{-1}[\boldsymbol{x}_i^T \boldsymbol{\beta}]) \qquad i = 1, \ldots, n. \tag{6}$$

Hence, we can express the joint density (5) in terms of the coefficients $\boldsymbol{\beta}$, and for observed data $\boldsymbol{y}$, this is the likelihood $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\phi})$ for $\boldsymbol{\beta}$.

Note that considerable simplification is obtained in (5) and (6) if the functions g and b'^{-1} are identical. Then

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n.$$

The link function

$$g(\mu) \equiv b'^{-1}(\mu)$$

is called the *canonical* link function. Under the canonical link, the canonical parameter is equal to the linear predictor.

Distribution	Normal	Poisson	Bernoulli Binomial
$b(\theta)$	$\frac{1}{2}\theta^2$		$\log(1 + \exp\theta)$
$b'(\theta) \equiv \mu$	θ		$\dfrac{\exp\theta}{1+\exp\theta}$
$b'^{-1}(\mu) \equiv \theta$	μ		$\log\dfrac{\mu}{1-\mu}$
Link	$g(\mu) = \mu$	$g(\mu) = \log\mu$	$g(\mu) = \log\dfrac{\mu}{1-\mu}$
	Identity link	Log link	Logit link

Table 1: Canonical link functions

Exercise 5: Complete the table above.

Clearly the linear model considered in §1 is also a generalised linear model where $Y_1, \ldots, Y_n$ are independent normally distributed, the explanatory variables enter a linear model through the linear predictor

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n.$$

and the link between $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and the linear predictor $\boldsymbol{\eta}$ is through the (canonical) identity link function

$$\mu_i = \eta_i \qquad i = 1, \ldots, n.$$

2.2 Maximum likelihood estimation

As usual, we maximise the log likelihood function which, from (2), can be written

$$\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i) \tag{7}$$

and depends on $\boldsymbol{\beta}$ through

$$\begin{aligned}
&\mu_i = b'(\theta_i) \qquad i = 1, \ldots, n \\
&g(\mu_i) = \eta_i \qquad i = 1, \ldots, n \\
&\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=0}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n.
\end{aligned}$$

To find $\hat{\boldsymbol{\beta}}$, we solve the equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ where $\boldsymbol{u}$ is the *score* vector whose components are given by

$$\begin{aligned}
u_k(\boldsymbol{\beta}) &\equiv \frac{\partial}{\partial \beta_k} \log f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left[\frac{y_i\theta_i - b(\theta_i)}{\phi_i} \right] \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \left[\frac{y_i\theta_i - b(\theta_i)}{\phi_i} \right] \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \qquad k = 0, \ldots, p. \\
&= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i} \frac{x_{ik}}{b''(\theta_i) g'(\mu_i)} \\
&= \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \qquad k = 0, \ldots, p
\end{aligned} \tag{8}$$

which depends on $\boldsymbol{\beta}$ through $\mu_i \equiv E(Y_i)$ and $Var(Y_i)$, $i = 1, \ldots, n$.

In practice, these equations are usually non-linear and have no analytic solution. Therefore, we rely on numerical methods to solve them.

First, we note that the Hessian and Fisher information matrices can be derived directly from (8), as

$$\begin{aligned}
[\boldsymbol{H}(\boldsymbol{\beta})]_{jk} &= \frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}) \\
&= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \\
&= \sum_{i=1}^{n} \frac{-\frac{\partial \mu_i}{\partial \beta_j}}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} + \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_j} \left[\frac{x_{ik}}{Var(Y_i) g'(\mu_i)} \right]
\end{aligned}$$

and

$$[\mathcal{I}(\boldsymbol{\beta})]_{jk} = E[-\boldsymbol{H}(\boldsymbol{\beta})]_{jk} = \sum_{i=1}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i) g'(\mu_i)^2}.$$

Exercise 6: Show that we we can write

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \tag{9}$$

where $\boldsymbol{X}$ is the design matrix and

$$\boldsymbol{W} = \text{diag}(\boldsymbol{w}) = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}$$

where

$$w_i = \frac{1}{Var(Y_i)g'(\mu_i)^2} \qquad i = 1, \ldots, n.$$

The Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ depends on $\boldsymbol{\beta}$ through $\boldsymbol{\mu}$ and $Var(Y_i),\ i = 1, \ldots, n$.

The scores in (8) may now be written as

$$\begin{aligned} u_k(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \mu_i) x_{ik} w_i g'(\mu_i) \\ &= \sum_{i=1}^{n} x_{ik} w_i z_i \qquad k = 0, \ldots, p \end{aligned}$$

where

$$z_i = (y_i - \mu_i) g'(\mu_i) \qquad i = 1, \ldots, n.$$

Therefore

$$\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}. \tag{10}$$

One possible method to solve the p simultaneous equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ that give $\hat{\boldsymbol{\beta}}$ is the (multivariate) Newton-Raphson method. If $\boldsymbol{\beta}^t$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \boldsymbol{H}(\boldsymbol{\beta}^t)^{-1} \boldsymbol{u}(\boldsymbol{\beta}^t). \tag{11}$$

In practice, an alternative to Newton-Raphson replaces $\boldsymbol{H}(\boldsymbol{\theta})$ in (11) with $E[\boldsymbol{H}(\boldsymbol{\theta})] \equiv -\mathcal{I}(\boldsymbol{\theta})$. Therefore, if $\boldsymbol{\beta}^t$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \mathcal{I}(\boldsymbol{\beta}^t)^{-1} \boldsymbol{u}(\boldsymbol{\beta}^t). \tag{12}$$

The resulting iterative algorithm is called *Fisher scoring*. Notice that if we substitute (9) and (10) into (12) we get

$$\begin{aligned} \boldsymbol{\beta}^{t+1} &= \boldsymbol{\beta}^t + [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{z}^t \\ &= [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X}]^{-1} [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X} \boldsymbol{\beta}^t + \boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{z}^t] \\ &= [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{W}^t [\boldsymbol{X} \boldsymbol{\beta}^t + \boldsymbol{z}^t] \\ &= [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{W}^i [\boldsymbol{\eta}^t + \boldsymbol{z}^t] \end{aligned}$$

where $\boldsymbol{\eta}^t$, $\boldsymbol{W}^t$ and $\boldsymbol{z}^t$ are all functions of $\boldsymbol{\beta}^t$.

This is a weighted least squares equation, as $\boldsymbol{\beta}^{t+1}$ minimises the weighted sum of squares

$$(\boldsymbol{\eta} + \boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{\eta} + \boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(\eta_i + z_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2$$

as a function of $\boldsymbol{\beta}$ where $w_1, \ldots, w_n$ are the weights and $\boldsymbol{\eta} + \boldsymbol{z}$ is called the *adjusted dependent variable*.

Therefore, the Fisher scoring algorithm proceeds as follows.

1. Choose an initial estimate $\boldsymbol{\beta}^t$ for $\hat{\boldsymbol{\beta}}$ at t=0.
2. Evaluate $\boldsymbol{\eta}^t$, $\boldsymbol{W}^t$ and $\boldsymbol{z}^t$ at $\boldsymbol{\beta}^t$.
3. Calculate $\boldsymbol{\beta}^{t+1} = [\boldsymbol{X}^T \boldsymbol{W}^t \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{W}^t [\boldsymbol{\eta}^t + \boldsymbol{z}^t]$.
4. If $||\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t|| >$ some prespecified (small) tolerance then set $t \to t+1$ and go to 2.
5. Use $\boldsymbol{\beta}^{t+1}$ as the solution for $\hat{\boldsymbol{\beta}}$.

As this algorithm involves iteratively minimising a weighted sum of squares, it is sometimes known as *iteratively (re)weighted least squares.*

Notes

1. Recall that the canonical link function is $g(\mu) = b'^{-1}(\mu)$ and with this link $\eta_i = g(\mu_i) = \theta_i$. Then
$$\frac{1}{g'(\mu_i)} = \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \qquad i = 1, \ldots, n.$$
Therefore $Var(Y_i)g'(\mu_i) = \phi_i$ which does not depend on $\boldsymbol{\beta}$, and hence $\frac{\partial}{\partial \beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right] = 0$ for all $j = 0, \ldots, p$. It follows that $\boldsymbol{H}(\boldsymbol{\theta}) = -\mathcal{I}(\boldsymbol{\theta})$ and, for the canonical link, Newton-Raphson and Fisher scoring are equivalent.
2. (**Exercise 7**) The linear model is a generalised linear model with identity link, $\eta_i = g(\mu_i) = \mu_i$ and $Var(Y_i) = \sigma^2$ for all $i = 1, \ldots, n$. For this model, show that $w_i = \sigma^{-2}$ and $z_i = y_i - \eta_i$, $i = 1, \ldots, n$. Hence, show that the Fisher scoring algorithm converges in a single iteration, from any starting point, to the usual least squares estimate.
3. Estimation of an unknown dispersion parameter σ^2 is discussed later. A common σ^2 has no effect on $\hat{\boldsymbol{\beta}}$.

2.3 Inference

Subject to standard regularity conditions, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$. For 'large enough n' we treat this distribution as an approximation.

Therefore, standard errors are given by

$$s.e.(\hat{\beta}_k) = [\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}]^{\frac{1}{2}}_{kk} = [(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}]^{\frac{1}{2}}_{kk} \qquad k = 0, \ldots, p,$$

where the diagonal matrix $\hat{\boldsymbol{W}} = \text{diag}(\hat{\boldsymbol{w}})$ is evaluated at $\hat{\boldsymbol{\beta}}$, that is $\hat{w}_i = (\widehat{Var}(Y_i)g'(\hat{\mu}_i)^2)^{-1}$ where $\hat{\mu}_i$ and $\widehat{Var}(Y_i)$ are evaluated at $\hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$. Furthermore, if $Var(Y_i)$ depends on an unknown dispersion parameter, then this too must be estimated in the standard error.

The asymptotic distribution of the maximum likelihood estimator can be used to provide approximate large sample confidence intervals, using

$$\frac{\hat{\beta}_k - \beta_k}{s.e.(\hat{\beta}_k)} \overset{\text{asymp}}{\sim} N(0,1).$$

2.4 Comparing generalised linear models [1]

As with linear models, we can proceed by comparing nested models H_0 and H_1 pairwise using a generalised likelihood ratio test where 'nested' means that H_0 and H_1 are based on the same exponential family distribution, have the same link function, but $\Theta^{(0)}$, the set of values of the canonical parameter $\boldsymbol{\theta}$ allowed by H_0, is a subset of $\Theta^{(1)}$, the set of values allowed by H_1.

[1] It should be noted that this section describes just one method for comparing models. General principles and other methods will be discussed in detail in the APTS module itself.

Without loss of generality, we can think of H_1 as the model

$$\eta_i = \sum_{j=0}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$

and H_0 is the same model with $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$.

Now, the log likelihood ratio statistic for a test of H_0 against H_1 is

$$\begin{aligned} L_{01} &\equiv 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\right) \\ &= 2\log f_{\boldsymbol{Y}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}^{(1)}) - 2\log f_{\boldsymbol{Y}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}^{(0)}) \end{aligned} \tag{13}$$

where $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ follow from $b'(\hat{\theta}_i) = \hat{\mu}_i,\ g(\hat{\mu}_i) = \hat{\eta}_i,\ i = 1, \ldots, n$ where $\hat{\boldsymbol{\eta}}$ for each model is the linear predictor evaluated at the corresponding maximum likelihood estimate for $\boldsymbol{\beta}$. Here, we assume that $\phi_i,\ i = 1, \ldots, n$ are known; unknown ϕ is discussed in §2.5.

We reject H_0 in favour of H_1 when $L_{01} > k$ where k is determined by α, the size of the test. Under H_0, L_{01} has an asymptotic chi-squared distribution with $p - q$ degrees of freedom.

The *saturated* model is defined to be the model where the canonical parameters $\boldsymbol{\theta}$ (or equivalently $\boldsymbol{\mu}$ or $\boldsymbol{\eta}$) are unconstrained, and the parameter space is n-dimensional. For the saturated model, we can calculate the m.l.es $\hat{\boldsymbol{\theta}}$ directly from their likelihood (5) by differentiating with respect to $\theta_1, \ldots, \theta_n$ to give

$$\frac{\partial}{\partial\theta_k}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) = \frac{y_k - b'(\theta_k)}{\phi_k} \qquad k = 1, \ldots, n.$$

Therefore $b'(\hat{\theta}_k) = y_k,\ k = 1, \ldots, n$, and it follows immediately that $\hat{\mu}_k = y_k,\ k = 1, \ldots, n$. Hence the saturated model fits the data perfectly, as the *fitted values* $\hat{\mu}_k$ and observed values y_k are the same for every observation $k = 1, \ldots, n$.

The saturated model is rarely of any scientific interest in its own right. It is highly parameterised, having as many parameters as there are observations. However, every other model is necessarily nested in the saturated model, and a test comparing a model H_0 against the saturated model can be interpreted as a goodness of fit test. If the saturated model, which fits the observed data perfectly, does not provide a significantly better fit than model H_0, we can conclude that H_0 is an acceptable fit to the data.

The log likelihood ratio statistic for a test of H_0 against the saturated alternative is, from (13)

$$L_0 = 2\log f_{\boldsymbol{Y}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}^{(s)}) - 2\log f_{\boldsymbol{Y}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}^{(0)})$$

where $\hat{\boldsymbol{\theta}}^{(s)}$ follows from $b'(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\mu}} = \boldsymbol{y}$. However, calibrating L_0 is not straightforward. In some circumstances (typically those where the response distribution might be adequately approximated by a normal) L_0 has an asymptotic chi-squared distribution with $n - q - 1$ degrees of freedom, under H_0. Therefore, if L_0 is 'too large' then we reject H_0 as a plausible model for the data, as it does not fit the data adequately. However, in other situations, for example, for Bernoulli data, this approximation breaks down.

The *degrees of freedom* of model H_0 is defined to be the degrees of freedom for this test, $n - q - 1$, the number of observations minus the number of linear parameters of H_0. We call L_0 the *scaled deviance* of model H_0.

From (7) and (13) we can write the scaled deviance of model H_0 as

$$L_0 = 2\sum_{i=1}^{n} \frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{\phi_i}. \tag{14}$$

which can be calculated using the observed data, provided that $\phi_i,\ i = 1, \ldots, n$ is known.

Notes

1. The log likelihood ratio statistic (13) for testing H_0 against a nonsaturated alternative H_1 can be written as

$$\begin{aligned} L_{01} &= 2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(1)}) - 2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(0)}) \\ &= [2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(s)}) - 2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(0)})] - [2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(s)}) - 2\log f_{\boldsymbol{Y}}(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(1)})] \\ &= L_0 - L_1. \end{aligned} \tag{15}$$

   Therefore the log likelihood ratio statistic for comparing two nested models is the difference between their scaled deviances. Furthermore, as $p - q = (n - q - 1) - (n - p - 1)$, the degrees of freedom for the test is the difference in degrees of freedom of the two models.

2. An alternative goodness of fit statistic for a model H_0 is Pearson's X^2 given by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\widehat{Var}(Y_i)}. \tag{16}$$

   X^2 is small when the squared differences between observed and fitted values (scaled by variance) is small. Hence, large values of X^2 correspond to poor fitting models. In fact, X^2 and L_0 are asymptotically equivalent. However, the asymptotic χ^2_{n-q-1} approximation associated with X^2 is often more reliable.

2.5 Models with an unknown dispersion parameter

Thus far, we have assumed that the ϕ_i are known. This is the case for both the Poisson and Bernoulli distributions ($\phi = 1$). Neither the scaled deviance (14) nor Pearson X^2 statistic (16) can be evaluated unless $a(\phi)$ is known. Therefore, when ϕ_i are not known, we cannot use the scaled deviance as a measure of goodness of fit, or to compare models using (15).

Progress is possible if we assume that $\phi_i = \sigma^2/m_i,\ i = 1, \ldots, n$ where σ^2 is a common unknown dispersion parameter and $m_1, \ldots, m_n$ are known weights. (A normal linear model takes this form, if we assume that $Var(Y_i) = \sigma^2,\ i = 1, \ldots, n$, in which case $m_i = 1,\ i = 1, \ldots, n$). Under this assumption

$$\begin{aligned} L_0 &= \frac{2}{\sigma^2} \sum_{i=1}^{n} m_i y_i [\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - m_i [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})] \\ &\equiv \frac{1}{\sigma^2} D_0 \end{aligned} \tag{17}$$

where D_0 can be calculated using the observed data. We call D_0 the *deviance* of the model.

In order to compare nested models H_0 and H_1, one might calculate the test statistic

$$F = \frac{L_{01}/(p-q)}{L_1/(n-p-1)} = \frac{(L_0 - L_1)/(p-q)}{L_1/(n-p-1)} = \frac{(D_0 - D_1)/(p-q)}{D_1/(n-p-1)}. \tag{18}$$

This statistic does not depend on the unknown dispersion parameter σ^2, so can be calculated using the observed data. Asymptotically, under H_0, L_{01} has a χ^2_{p-q} distribution and L_{01} and L_1 are independent (not proved here). Assuming that L_1 has an approximate χ^2_{n-p-1} distribution, then F has an approximate F distribution with $p - q$ degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator. Hence, we compare nested generalised linear models by calculating F and rejecting H_0 in favour of H_1 if F is too large.

The dependence of the maximum likelihood equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ on σ^2 (where $\boldsymbol{u}$ is given by (8)) can be eliminated by multiplying through by σ^2. However, inference based on the maximum likelihood estimates, as described in §2.3 does require knowledge of σ^2. This is because asymptotically $Var(\hat{\boldsymbol{\beta}})$ is the inverse of the Fisher information matrix $\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, and this depends on $w_i = \frac{1}{Var(Y_i)g'(\mu_i)^2}$ where $Var(Y_i) = \phi_i b''(\theta_i) = \sigma^2 b''(\theta_i)/m_i$ here.

Therefore, to calculate standard errors and confidence intervals, we need to supply an estimate $\hat{\sigma}^2$ of σ^2. Generally, rather than use the maximum likelihood estimate, it is more common to base an estimator of σ^2 on the Pearson X^2 statistic. As $Var(Y_i) = \phi_i V(\mu_i) = \sigma^2 V(\mu_i)/m_i$ here (where the *variance function* $V(\mu)$ is defined as $b''(\theta)$, written in terms of μ), then from (16)

$$X^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{m_i(y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i)}. \tag{19}$$

Exercise 8: Making the assumption that, if H_0 is an adequate fit, X^2 has an chi-squared distribution with $n - q - 1$ degrees of freedom, show that

$$\hat{\sigma}^2 \equiv \frac{1}{n-q-1}\sum_{i=1}^{n}\frac{m_i(y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i)}$$

is an approximately unbiased estimator of σ^2. Suggest an alternative estimator based on the deviance D_0.

2.6 Residuals and Model Checking

Recall that for linear models, we define the residuals to be the differences between the observed and fitted values $y_i - \hat{\mu}_i^{(0)},\ i = 1, \ldots, n$. In fact, both the scaled deviance and Pearson X^2 statistic for a normal GLM are the sum of the squared residuals divided by σ^2. We can generalise this to define residuals for other generalised linear models in a natural way.

For any GLM we define the *Pearson residuals* to be

$$e_i^P = \frac{y_i - \hat{\mu}_i^{(0)}}{\widehat{Var}(Y_i)^{\frac{1}{2}}} \qquad i = 1, \ldots, n.$$

Then, from (16), X^2 is the sum of the squared Pearson residuals.

For any GLM we define the *deviance residuals* to be

$$e_i^D = \mathrm{sign}(y_i - \hat{\mu}_i^{(0)})\left[\frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{\phi_i}\right]^{\frac{1}{2}}, \qquad i = 1, \ldots, n,$$

where $\mathrm{sign}(x) = 1$ if $x > 0$ and -1 if $x < 0$. Then, from (14), the scaled deviance, L_0, is the sum of the squared deviance residuals.

When $\phi_i = \sigma^2/m_i$ and σ^2 is unknown, as in §2.5, the expressions above need to be multiplied through by σ^2 to eliminate dependence on the unknown dispersion parameter. Therefore, for a normal GLM the Pearson and deviance residuals are both equal to the usual residuals, $y_i - \hat{\mu}_i^{(0)},\ i = 1, \ldots, n$.

Both the Pearson and deviance residuals can be standardised by dividing through by $(1 - h_{ii})^{1/2}$, as in §1.4. The derived residuals

$$r_i^* = r_i^D + \frac{1}{r_i^D}\log(r_i^P/r_i^D)$$

are close to normal for a wide range of models, where r_i^D and r_i^P are the standardised deviance and Pearson residuals, respectively.

Generalised linear model checking, using residuals, is based on the same kind of diagnostic plots, as were suggested for linear models in §1.7. Similarly, the Cook's distance C_j for linear models can be adapted for GLMs by using Pearson residuals.

3 Practical statistical modelling in R

The following exercises are adapted from *Practicals to Accompany Statistical Models* by A. C. Davison, available at `http://statwww.epfl.ch/davison/SM/`. You will need to install the libraries `ellipse` and `SMPracticals` containing the data sets, and other useful functions and load them using

```
library(ellipse)
library(SMPracticals)
```

3.1 `trees` contains data on the volume, height and girth (diameter) of 31 felled black cherry trees; girth is measured four feet six inches above ground. The problem is to find a simple linear model for predicting volume from height and girth.

```
pairs(trees,panel = panel.smooth)
pairs(log(trees), panel = panel.smooth)
```

`coplot` generates conditioning plots, in which the relationship between two variables is displayed conditional on subsets of values of other variables. This is useful to see if the relationship is stable over the range of other variables. To assess this for the relationship of log volume and log girth, conditional on height:

```
attach(trees)
coplot(log(Volume) ~ log(Girth) | Height, panel = panel.smooth)
```

Try this on the orginal scale also. For an initial fit, we take a linear model and assess model fit using diagnostic plots:

```
fit <- glm(Volume ~ Girth + Height)
summary(fit)
plot.glm.diag(fit)
```

What do you make of the fit?

To assess the possibility of transformation:

```
library(MASS)
boxcox(fit)
```

Both $\lambda = 1$ and $\lambda = 0$ lie outside the confidence interval, though the latter is better supported. One possibility is to take $\lambda = 1/3$, corresponding to response $\texttt{Volume}^{1/3}$. What transformations for `Girth` and `Height` are then needed for dimensional compatibility? Fit this model, give interpretations of the parameter estimates, and discuss its suitability.

An alternative is to suppose that a tree is conical in shape, in which case $\texttt{Volume} \propto \texttt{Height} \times \texttt{Girth}^2$. Equivalently, we fit

```
fit <- glm(log(Volume) ~ log(Girth) + log(Height)))
summary(fit)
```

```
plot.glm.diag(fit)
```

Are the parameter estimates consistent with this model? Does it fit adequately? What advantage has it over the others for prediction of future volumes?
(Chapter 8; Atkinson, 1985, p. 63)

3.2 `salinity` contains $n = 28$ observations on the salinity of water in Pamlico Sound, North Carolina. The response `sal` is the bi-weekly average of salinity. The next three columns contain values of the covariates, respectively a lagged value of salinity `lag`, a trend indicator `trend`, and the river discharge `dis`. Using the techniques of the previous problem as a guide, find a model suitable for prediction of salinity from the covariates. The data contain at least one outlier.
(Chapter 8; Ruppert and Carroll, 1980; Atkinson, 1985, p. 48)

3.3 `shuttle` contains the data in Table 1.3 of Davison (2003) on O-ring failures for the space shuttle. To fit a binomial logistic regression model with covariate temperature:

```
fit <- glm(cbind(r, m-r) ~ temperature, data = shuttle, binomial)
anova(fit)
summary(fit)
```

Try fitting with and without both covariates. To assess model fit, try `plot.glm.diag(fit)`. Do you find these diagnostics useful?
(Sections 10.1-10.4; Dalal et al., 1989)

3.4 `bliss` contains data on mortality of flour-beetles as a function of dose of a poison. To plot the death rates:

```
plot(log(dose), r/m, ylim = c(0, 1), ylab = "Proportion dead")
fit <- glm(cbind(r, m-r) ~ log(dose), binomial)
summary(fit)
points(log(dose), fitted(fit), pch = "L")
```

Does the fit seem good to you? Try again with the `probit` and `cloglog` link functions, for example:

```
fit <- glm(cbind(r, m-r) ~ log(dose), binomial(cloglog))
points(log(dose), fitted(fit), pch = "C")
```

Which fits best? Give a careful interpretation of the resulting model.
(Sections 10.1-10.4; Bliss, 1935)