On apsects of statistical modelling

Preliminary material

April 1, 2025

Ioannis Kosmidis
University of Warwick
ioannis.kosmidis@warwick.ac.uk

Table of contents

Introduction

In order to get the most out of the AUEB MSc short course "On aspects of 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 statistical modelling in R.

The following reading and activities are recommended to all students to (re)-familiarise themselves with those topics.

Linear and generalised linear models: A student who has covered Davison (2003, Chapter 8 and 10.1-10.4) will be more than adequately prepared for the short course. For students without access to this book, the main theory is repeated in the current set of preliminary notes. 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 material exercises: Nine exercises are included in the current preliminary material.

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

Typos and issues

You can report and suggest fixes to typos and issues by email to ioannis.kosmidis@warwick.ac.uk.

Chapter 1

Linear models

1.1 Introduction

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables (or *covariates*). 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 $y = (y_1, y_2, \dots, y_n)^\top$. In a statistical model, these are assumed to be observations of *random variables* $Y = (Y_1, Y_2, \dots, Y_n)^\top$. Associated with each y_i is a vector $x_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^\top$ of values of p explanatory variables.

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

$$\begin{aligned} \mathrm{E}(Y_i \mid x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \\ &= \sum_{j=0}^{p} x_{ij}\beta_j \quad \text{(where we define } x_{i0} = 1\text{)} \\ &= x_i^\top \beta \\ &= [X\beta]_i \qquad (i = 1, \dots, n)\,, \end{aligned} \tag{1.1}$$

where

$$\mathrm{E}(Y \mid X) = \begin{bmatrix} \mathrm{E}(Y_1 \mid x_1) \\ \vdots \\ \mathrm{E}(Y_n \mid x_n) \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},$$

and $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is a vector of fixed but unknown parameters describing the dependence of Y_i on x_i. The four ways of describing the linear model in (1.1) are equivalent, but the most economical is the matrix form

$$\mathrm{E}(Y \mid X) = X\beta\,. \tag{1.2}$$

The $n \times (p+1)$ matrix X consists of known (observed) constants and is called the *model matrix*. The ith row of X is $x_i^\top$, the explanatory data corresponding to the ith observation of the response. The jth column of X contains the n observations of the jth explanatory variable.

Example 1.1. The null model

$$\mathrm{E}(Y_i) = \beta_0 \qquad (i = 1, \dots, n)$$

has

$$X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Example 1.2. The simple linear regression

$$\mathrm{E}(Y_i \mid x_i) = \beta_0 + \beta_1 x_i \qquad (i = 1, \dots, n)$$

has

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Example 1.3. The polynomial regression model

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

has

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Example 1.4. The multiple regression model

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

has

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Strictly, the only requirement for a model to be linear is that the relationship between the response variables, Y, and any explanatory variables can be written in the form (1.2). No further specification of the joint distribution of $Y_1, \ldots, Y_n$ is required. However, statistical inference about the model parameters is conveniently performed under the *normal linear model*, which involves three further assumptions:

1. $Y_1, \ldots, Y_n$ are independent random variables conditionally on the covariates vectors $x_1, \ldots, x_n$.
2. Conditionally on the covariates vectors $x_1, \ldots, x_n$, $Y_1, \ldots, Y_n$ are normally distributed.
3. $\mathrm{var}(Y_1 \mid x_1) = \mathrm{var}(Y_2 \mid x_1) = \cdots = \mathrm{var}(Y_n) = \sigma^2$ or, equivalently, $Y_1, \ldots, Y_n$ are homoscedastic.

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

$$Y_i \mid x_i \sim \mathrm{N}\left(x_i^\top \beta, \sigma^2\right) \qquad (i = 1, \ldots, n).$$

Another way of writing this is

$$Y_i = x_i^\top \beta + \epsilon_i \qquad (i = 1, \ldots, n),$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (IID) random variables with $\epsilon_1 \sim \mathrm{N}(0, \sigma^2)$.

The normal linear model can now be expressed in matrix form as

$$Y = X\beta + \epsilon, \tag{1.3}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ has a multivariate normal distribution with mean vector 0 and variance covariance matrix $\sigma^2 I_n$, where I_n is the $n \times n$ identity matrix

$$I_n = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}.$$

This is because $\mathrm{var}(\epsilon_i) = \sigma^2$, and $\epsilon_1, \ldots, \epsilon_n$ are independent, which implies that $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ $(i, j = 1, \ldots, n)$.

It follows from (1.3) that the distribution of Y is multivariate normal with mean vector $X\beta$ and variance covariance matrix $\sigma^2 I_n$, that is $Y \sim \mathrm{N}(X\beta, \sigma^2 I_n)$.

1.2 Least squares estimation

The regression coefficients $\beta_0, \ldots, \beta_p$ describe the pattern by which the response is associated with the explanatory variables. We use the observed response values $y_1, \ldots, y_n$ to *estimate* that association.

In least squares estimation, roughly speaking, we choose $\hat{\beta}$, the estimates of β, to make the estimated means $\widehat{\mathrm{E}(Y \mid X)} = X\hat{\beta}$ *as close as possible* to the observed values y, where closeness is determined in terms of the sum of squared errors. In other words, we seek $\hat{\beta}$ that minimises the sum of squares

$$\begin{aligned} \sum_{i=1}^{n} \{y_i - \mathrm{E}(Y_i \mid x_i)\}^2 &= \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 , \end{aligned} \tag{1.4}$$

with respect to $\beta = (\beta_0, \ldots, \beta_p)^\top$.

Exercise 1.1 (Normal equations)**.** Differentiate the sum of squares in (1.4) with respect to β_k ($k = 0, \ldots, p$), to show that $\hat{\beta}$ should satisfy

$$X^\top X\hat{\beta} = X^\top y\,. \tag{1.5}$$

The least squares estimates $\hat{\beta}$ are the solutions to the set of $p+1$ simultaneous linear equations in (1.5), which are known as the *normal equations*. If $X^\top X$ is invertible (as it usually is) then the least squares estimates are given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y\,.$$

The corresponding fitted values are

$$\begin{aligned} & \hat{y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top y \\ \Rightarrow \quad & \hat{y}_i = x_i^\top \hat{\beta} \qquad (i = 1, \ldots, n)\,. \end{aligned}$$

The matrix $H = X(X^\top X)^{-1}X^\top$ is typically called the *hat* matrix, because $\hat{y} = Hy$, that is H "puts a hat" on y. The *residuals* are

$$\begin{aligned} & e = y - \hat{y} = y - X\hat{\beta} = (I_n - H)y \\ \Rightarrow \quad & e_i = y_i - x_i^\top \hat{\beta} \qquad (i = 1, \ldots, n)\,. \end{aligned}$$

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

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - x_i^\top \hat{\beta} \right)^2 ,$$

and is the minimum value that the sum of squared errors in (1.4) attains.

Exercise 1.2 (Properties of the least squares estimator)**.**

1. Show that $\hat{\beta}$ is multivariate normal with mean $\mathrm{E}(\hat{\beta} \mid X) = \beta$, and variance covariance matrix $\mathrm{var}(\hat{\beta} \mid X) = \sigma^2(X^\top X)^{-1}$.
2. Assuming that $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with $\epsilon_1 \sim \mathrm{N}(0, \sigma^2)$, show that the least squares estimate $\hat{\beta}$ is also the maximum likelihood estimate. To do that start by showing that the likelihood for the normal linear model is

$$f_Y(y \mid X\,;\, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \right) . \tag{1.6}$$

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, which represents the variability of the observations about their mean.

We can estimate σ^2 using maximum likelihood. Maximising (1.6) with respect to β and σ^2 gives

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

Under the assumptions of the normal linear model, D is independent of $\hat{\beta}$ and

$$\frac{D}{\sigma^2} \sim \chi^2_{n-p-1} \,.$$

Hence,

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n-p-1}{n}\sigma^2 \,.$$

As a result, the maximum likelihood estimator for σ^2 is biased for fixed p, and is only asymptotically unbiased because $(n-p-1)/n \to 1$ as $n \to \infty$. For this reason, 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 $n-p-1$ is the number of observations minus the number of coefficients in the model, and is called the *degrees of freedom* of the model. We estimate the error variance by the deviance divided by the degrees of freedom.

1.4 Inference

From the distribution of $\hat{\beta}$ under the normal linear model (see Exercise 1.2), it follows that

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

Replacing the unknown parameter σ with its estimate s, the definition of the t distribution gives that

$$T_k = \frac{\hat{\beta}_k - \beta_k}{s[(X^\top X)^{-1}]_{kk}^{1/2}} \sim t_{n-p-1} \,.$$

Hence, T_k is a pivotal quantity (function of the random variables and parameters, whose distribution does not depend on the parameters), and can be used for constructing inferences about β_k in the form of confidence intervals and test of hypotheses of the form $H_0 : \beta_k = b$.

The denominator $s.e.(\hat{\beta}_k) = s[(X^\top X)^{-1}]_{kk}^{1/2}$ is called the estimated standard error for $\hat{\beta}_k$.

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

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

and

$$e \mid X \sim \mathrm{N}(0, \sigma^2(I_n - H)) \,.$$

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

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

where h_{ii} is the ith diagonal element of the hat matrix H.

1.5 Prediction

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

$$\hat{Y}_+ = x_+^\top\hat{\beta}\,.$$

Then,

$$\hat{Y}_+ \mid X, x_+ \sim N(x_+^\top\beta, \sigma^2 h_{++})\,,$$

where $h_{++} = x_+^\top(X^\top X)^{-1}x_+$. Hence, confidence intervals for predictive means can be derived using

$$\frac{\hat{Y}_+ - x_+^\top\beta}{sh_{++}^{1/2}} \sim t_{n-p-1}\,.$$

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

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

because $\hat{Y}_+$ 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

A pair of *nested* linear models can be compared 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, which sets some of the coefficients to zero.

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

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

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

A *generalised likelihood ratio test* of H_0 against H_1 uses a test statistic of the form

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

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

For a normal linear model,

$$f_Y(y \mid X\,;\, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2\right).$$

This is maximised with respect to (β, σ^2) for $\beta := \hat{\beta}$ and $\sigma^2 := \hat{\sigma}^2 = D/n$. So,

$$\begin{aligned}\max_{\beta,\sigma^2} f_Y(y \mid X\,;\, \beta, \sigma^2) &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2D}\sum_{i=1}^{n}(y_i - x_i^\top\hat{\beta})^2\right) \\ &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)\end{aligned}$$

Exercise 1.3 (Likelihood ratio statistic and F tests in the normal linear model)**.** Denote the deviances under models H_0 and H_1 as D_0 and D_1, respectively. Show that the likelihood ratio test statistic T above can be written as

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

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

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

We know from Section 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. 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.

Hence, 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 normal linear models may be unreliable if some of 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 $\mathrm{E}(Y_i)$ and the explanatory variables.
2. $Y_1, \ldots, Y_n$ are normally distributed.
3. $\mathrm{var}(Y_1) = \mathrm{var}(Y_2) = \cdots = \mathrm{var}(Y_n)$.
4. $Y_1, \ldots, Y_n$ are independent random variables.

Evidence of departures from the above assumptions can be explored 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 $x_i^\top \widehat{\beta}$ $(i = 1, \ldots, n)$. The plot should look like a random scatter. If any patterns are apparent, for example increasing or decreasing variance as the fitted values increase ('funnelling' in the residual plot), then this is evidence against the homoscedasticity assumptions.
4. Independence is typically difficult to validate. Nevertheless, if observations have been collected in serial order, serial correlation may be detected by a lagged scatterplot or correlogram of the residuals.

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

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

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.

Exercise 1.4 (Basic properties of the hat matrix and the leverage)**.**

1. Show that H is idempotent.
2. Show that $h_{ii} \in (0,1)$ and that $\text{tr}(H) = \sum_{i=1}^{n} h_{ii} = p$, where $\text{tr}(H)$ denotes the trace of the matrix H.

1.8 Bayesian inference for linear models

Bayesian inference for the parameters β and σ^2 of the normal linear model requires computation of the posterior density. Bayes theorem gives us

$$f(\beta, \sigma^2 \mid y, X) \propto f(y \mid X, \beta, \sigma^2) f(\beta, \sigma^2)\,,$$

where the likelihood $f(y \mid X, \beta, \sigma^2)$ is given by (1.6) as

$$f(y \mid X, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2\right).$$

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

$$\sigma^{-2} \sim \text{Gamma}(a_0, b_0) \quad \text{and} \quad \beta \,|\, \sigma^2 \sim \text{N}(\mu_0, \sigma^2 V_0)\,,$$

where a_0, b_0, μ_0, and V_0 are hyperparameters, whose values are chosen to reflect prior uncertainty about the linear model parameters β and σ^2.

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

$$\sigma^{-2} \sim \text{Gamma}(a_0 + n/2, b) \quad \text{and} \quad \beta \,|\, \sigma^2 \sim \text{N}(\mu, \sigma^2 V)\,,$$

where $V = (X^\top X + V_0^{-1})^{-1}$, $\mu = V(X^\top y + V_0^{-1}\mu_0)$ and

$$\begin{aligned} b &= b_0 + \frac{1}{2}\left(y^\top y + \mu_0 V_0^{-1}\mu_0 - \mu V^{-1}\mu\right) \\ &= b_0 + \frac{1}{2}\left\{(n-p-1)s^2 + \left[\mu_0 - \hat{\beta}\right]^\top \left[V_0 + (X^\top X)^{-1}\right]^{-1}\left[\mu_0 - \hat{\beta}\right]\right\}, \end{aligned}$$

if $X^\top X$ is non-singular, where $\hat{\beta}$ and s^2 are the classical unbiased estimators for β and σ^2.

In applications where prior information about the model parameters β and σ^2 is weak, it is conventional to use the vague prior specification given by the improper prior density

$$f(\beta, \sigma^2) \propto \sigma^{-2}. \tag{1.7}$$

This corresponds to the conjugate prior structure above with $a_0 = -(p+1)$, $b_0 = 0$ and $V_0^{-1} = 0$.

Exercise 1.5 (Links between Bayesian and frequentist inference for the normal linear model)**.**

1. Show that, for the vague prior specification in (1.7), the posterior mean of β is the least squares estimator $\hat{\beta}$. 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.
2. For a longer exercise, show that $(\beta-\mu)/\sigma$ 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 T_k derived in Section 1.4 above.

Chapter 2

Generalized linear models

2.1 Introduction

The generalized linear model extends the normal linear model defined in Section 1.1 to allow a more flexible family of probability distributions.

Suppose that $y_1, \dots, y_n$ are observations on random variables $Y_1, \dots, Y_n$ that are conditionally independent given $x_1, \dots, x_n$, where x_i is a p-vector of covariates. A generalized linear model (GLM) assumes that, conditionally on x_i, Y_i has an exponential family distirbution with density or probability mass function

$$f_Y(y \mid X\,;\, \theta, \phi) = \exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i) \right) , \tag{2.1}$$

where $\theta = (\theta_1, \dots, \theta_n)^\top$ is the collection of canonical parameters and $\phi = (\phi_1, \dots, \phi_n)^\top$ 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 $x_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^\top$ through the *linear predictor* η_i, where

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

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

The distribution of Y is linked to the linear predictor η through the *link function* g as

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

where $\mu_i = E(Y_i \mid x_i)$.

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 $\mathfrak{R}$, because 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 $\mu_i \in \mathfrak{R}^+$; for the Bernoulli distribution $\mu_i \in (0, 1)$. If g is not chosen carefully, then there may exist a combination of x_i and β such that $\eta_i \neq g(\mu_i)$ for any possible value of μ_i. Most common choices of link function map the set of allowed values for μ_i onto $\mathbb{R}$.

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

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

So,

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

and, because $g(\mu_i) = \eta_i = x_i^\top \beta$,

$$\theta_i = b'^{-1}(g^{-1}(x_i^\top \beta)) \qquad (i = 1, \dots, n)\,. \tag{2.2}$$

Hence, we can express the joint density (2.1) in terms of the coefficients β, and for observed data y, this is the likelihood $f_Y(y; \beta, \phi)$ about β.

Note that considerable simplification is obtained in (2.1) and (2.2) if the functions g and b'^{-1} are identical. Then,

$$\theta_i = x_i^\top \beta \qquad (i = 1, \dots, n)\,.$$

The link function

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

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

Table 2.1: Canonical link functions

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

Exercise 2.1 (GLM characteristics)**.** Complete Table 2.1.

Clearly the linear model considered in Chapter 1 is also a generalized linear model where $Y_1, \dots, Y_n$ are independent and normally distributed, the explanatory variables enter the model through the linear predictor

$$\eta_i = x_i^\top \beta \qquad (i = 1, \dots, n)\,,$$

and the link between $E(Y) = \mu$ and the linear predictor η is through the (canonical) identity link function

$$\mu_i = \eta_i \qquad (i = 1, \dots, n)\,.$$

2.2 Maximum likelihood estimation

As usual, we maximize the log likelihood function which, from (2.1), can be written as

$$\log f_Y(y \mid X\,;\, \beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i)\,, \tag{2.3}$$

and depends on β through expression (2.2) for the canonical parameters.

The maximum likelihood estimate $\hat{\beta}$ satisfies $u(\hat{\beta}) = 0$ where u is the *score* vector whose components are

given by

$$\begin{aligned} u_k(\beta) &\equiv \frac{\partial}{\partial \beta_k} \log f_Y(y;\beta) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi_i} \right] \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi_i} \right] \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\ &= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i} \frac{x_{ik}}{b''(\theta_i) g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{var}(Y_i \mid x_i)} \frac{x_{ik}}{g'(\mu_i)} \qquad (k = 0, \dots, p)\,, \end{aligned} \tag{2.4}$$

which depends on β through $\mu_i = E(Y_i)$ and $\mathrm{var}(Y_i \mid x_i)$ $(i = 1, \dots, n)$.

The equations $u(\hat{\beta}) = 0$ are usually non-linear and have no analytic solution. For that reason, we rely on numerical methods to solve them.

First, we note that the Hessian and Fisher information matrices can be derived directly from (2.4), as

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

and

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

Exercise 2.2 (Fisher information matrix)**.** Show that the Fisher information matrix can be written as

$$I(\beta) = X^\top W X\,, \tag{2.5}$$

where X is the model matrix and

$$W = \mathrm{diag}(w) = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{bmatrix},$$

where

$$w_i = \frac{1}{\mathrm{var}(Y_i \mid x_i) g'(\mu_i)^2} \qquad (i = 1, \dots, n)\,.$$

The Fisher information matrix $I(\beta)$ depends on β through μ_i and $\mathrm{var}(Y_i \mid x_i)$ $(i = 1, \dots, n)$.

The scores in (2.4) may now be written as

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

where

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

Hence,

$$u(\beta) = X^\top W z\,. \tag{2.6}$$

One possible method to solve the p simultaneous equations $u(\beta) = 0$ is the Newton-Raphson method. If β^t is the current estimate of β, then the next estimate is

$$\beta^{t+1} = \beta^t - H(\beta^t)^{-1}u(\beta^t)\,. \tag{2.7}$$

A popular alternative to Newton-Raphson replaces $H(\beta)$ in (2.7) with $E(H(\beta)) = -I(\beta)$. If β^t is the current estimate of β, the next estimate is

$$\beta^{t+1} = \beta^t + I(\beta^t)^{-1}u(\beta^t)\,. \tag{2.8}$$

The resulting iterative algorithm is called *Fisher scoring*. Notice that if we substitute (2.5) and (2.6) into (2.8) we get

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

where η^t, W^t, and z^t are η, W and z evaluated at β^t.

As is clear, β^{t+1} are estimates from a weighted linear regression model of the, so called, *working variates* $\eta^t + z^t$ on X with weights W^t. Equivalently, β^{t+1} minimizes the weighted sum of squares

$$(\eta^t + z^t - Xb)^\top W^t (\eta^t + z^t - Xb) = \sum_{i=1}^{n} w_i^t \left(\eta_i^t + z_i^t - x_i^\top b\right)^2\,,$$

with respect to b.

The Fisher scoring algorithm proceeds as follows:

0. Choose an initial estimate β^0 for \$ and a small constant $\epsilon > 0$

For $t = 0, 1, ...,$ do:

1. Evaluate η^t, W^t and z^t at β^t.
2. Calculate $\beta^{t+1} = [X^\top W^t X]^{-1} X^\top W^t [\eta^t + z^t]$.
3. If $||\beta^{t+1} - \beta^t|| > \epsilon$ then set $t \to t + 1$ and go to 2.
4. Use β^{t+1} as the value for $\hat{\beta}$.

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

Recall that the canonical link function is $g(\mu) = b'^{-1}(\mu)$ and with this link $\eta_i = g(\mu_i) = \theta_i$. Then,

$$\frac{1}{g'(\mu_i)} = \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \qquad i = 1, \dots, n.$$

As a result, $\text{var}(Y_i \mid x_i)g'(\mu_i) = \phi_i$, which does not depend on β. It follows that $\frac{\partial}{\partial \beta_j}\left[\frac{x_{ik}}{\text{var}(Y_i|x_i)g'(\mu_i)}\right] = 0$ $(j = 0, \dots, p)$. It follows that $H(\beta) = -I(\beta)$ and, for the canonical link, Newton-Raphson and Fisher scoring are equivalent.

Exercise 2.3 (IWLS for the normal linear model)**.** For the normal linear model, show that $w_i = \sigma^{-2}$ and $z_i = y_i - \eta_i$ $(i = 1, \dots, n)$. Hence, show that the Fisher scoring algorithm converges in a single iteration, from any starting point, to the usual least squares estimate.

2.3 Inference

Subject to standard regularity conditions, $I(\beta)^{1/2}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean 0 and variance covariance matrix I_p. So, we can treat the normal distribution with mean β and variance $I(\beta)^{-1}$ as the approximate distribution of $\hat{\beta}$.

Hence, standard errors can be estimated as

$$[I(\hat{\beta})^{-1}]^{\frac{1}{2}}_{kk} = [(X^\top \hat{W} X)^{-1}]^{\frac{1}{2}}_{kk} \qquad (k = 0, \ldots, p)\,,$$

where $\hat{W}$ is W evaluated at $\hat{\beta}$ and $\hat{\phi}_i$, if $\text{var}(Y_i \mid x_i)$ depends on an unknown dispersion parameter. Section 2.5 discusses the estimation of ϕ_i in models with unknown dispersion parameter.

The asymptotic distribution of the maximum likelihood estimator can be used to provide asymptotically valid confidence intervals, and hypeotheris testing procedures, using

$$\frac{\hat{\beta}_k - \beta_k}{[(X^\top \hat{W} X)^{-1}]^{\frac{1}{2}}_{kk}} \overset{\text{asymp}}{\sim} \text{N}(0, 1)\,.$$

2.4 Comparing generalized linear models

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

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

$$\begin{aligned} L_{01} &\equiv 2\log\left(\frac{\max_{\theta\in\Theta^{(1)}} f_Y(y \mid X\,;\,\theta)}{\max_{\theta\in\Theta^{(0)}} f_Y(y \mid X\,;\,\theta)}\right) \\ &= 2\log f_Y(y \mid X\,;\,\hat{\theta}^{(1)}) - 2\log f_Y(y \mid X\,;\,\hat{\theta}^{(0)}) \end{aligned} \tag{2.9}$$

where $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(0)}$ result from $b'(\hat{\theta}_i^{(j)}) = \hat{\mu}_i^{(j)}$, $g(\hat{\mu}_i^{(j)}) = \hat{\eta}_i^{(j)}$ $(i = 1, \ldots, n)$ where $\hat{\eta}^{(j)}$ is the linear predictor evaluated at the maximum likelihood estimate for β under hypothesis H_j $(j = 0, 1)$. Here, we assume that ϕ_i $(i = 1, \ldots, n)$ are known; the case of unknown ϕ is discussed in Section 2.5.

We reject H_0 in favour of H_1 when $L_{01} > k$ where k is determined by the size α 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 θ (or equivalently μ or η) are unconstrained, and the parameter space is n-dimensional. For the saturated model, we can calculate the maximum likelihood estimators $\hat{\theta}$ directly from their likelihood (2.1) by differentiating with respect to $\theta_1, \ldots, \theta_n$ to give

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

Therefore $b'(\hat{\theta}_k) = y_k$ $(k = 1, \ldots, n)$, and, hence, $\hat{\mu}_k = y_k$ $(k = 1, \ldots, n)$. 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. The saturated model is rarely of any scientific interest in its own right. It is overly parameterized, 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 there is no significant evidence that the saturated model — which fits the observed data perfectly — provides a better fit than model H_0, we can conclude that H_0 is an acceptable fit to the data.

From (2.9), the log likelihood ratio statistic for a test of H_0 against the saturated alternative is

$$L_0 = 2\log f_Y(y \mid X\,;\,\hat{\theta}^{(s)}) - 2\log f_Y(y \mid X\,;\,\hat{\theta}^{(0)})$$

where $\hat{\theta}^{(s)}$ is such that $b'(\hat{\theta}^{(s)}) = y$. 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. Large values of L_0 is evidence against H_0 as a plausible model for the data. However, in other situations, for example Bernoulli response distributions, the χ^2 approximation to L_0 may be poor.

The *degrees of freedom* of model H_0 and for this test is $n - q - 1$, whicn is 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 (2.3) and (2.9) we can write the scaled deviance of model H_0 as

$$L_0 = 2\sum_{i=1}^{n} \frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{\phi_i}\,, \tag{2.10}$$

which is easily computed using the observed data, provided that ϕ_i $(i = 1, \ldots, n)$ is known.

Remark 2.1. The log likelihood ratio statistic (2.9) for testing H_0 against a non-saturated alternative H_1 can be written as

$$\begin{aligned} L_{01} &= 2\log f_Y(y;\hat{\theta}^{(1)}) - 2\log f_Y(y;\hat{\theta}^{(0)}) \\ &= [2\log f_Y(y;\hat{\theta}^{(s)}) - 2\log f_Y(y;\hat{\theta}^{(0)})] - [2\log f_Y(y;\hat{\theta}^{(s)}) - 2\log f_Y(y;\hat{\theta}^{(1)})] \\ &= L_0 - L_1\,. \end{aligned} \tag{2.11}$$

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)$, that is the degrees of freedom for the test is the difference in degrees of freedom of the two models.

Remark 2.2. An alternative goodness of fit statistic for a model H_0 is Pearson's X^2 given by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\widehat{\mathrm{var}(Y_i \mid x_i)}}\,. \tag{2.12}$$

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

2.5.1 Model comparison

Thus far, we have assumed that $\phi_1, \ldots, \phi_n$ are known. This is the case for both the Poisson and Bernoulli distributions, where $\phi_i = 1$. When ϕ_i are not known, we can evaluate neither the scaled deviance (2.10) nor the Pearson X^2 statistic (2.12), and hence we cannot directly construct inferences based on them, or compare models using (2.11).

Progress can be made 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 (this form is present in a normal linear model, where $\mathrm{var}(Y_i \mid x_i) = \sigma^2$). Under this assumption

$$\begin{aligned} L_0 &= \frac{2}{\sigma^2}\sum_{i=1}^{n}\left[m_i y_i(\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}) - m_i\{b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})\}\right] \\ &\equiv \frac{1}{\sigma^2} D_0\,, \end{aligned} \tag{2.13}$$

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

$$F = \frac{L_{01}/(p-q)}{L_1/(n-p-1)} = \frac{(L_0 - L_1)/(p-q)}{L_1/(n-p-1)} = \frac{(D_0 - D_1)/(p-q)}{D_1/(n-p-1)}. \tag{2.14}$$

This statistic does not depend on the unknown dispersion parameter σ^2, so it 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, large values of F is evidence against H_0 in favour of H_1.

2.5.2 Inference about model parameters

The dependence of the maximum likelihood equations $u(\hat{\beta}) = 0$ on σ^2 (where u is given by (2.4)) can be eliminated by multiplying through by σ^2. However, inference based on the maximum likelihood estimates requires knowledge of σ^2. This is because asymptotically the variance covariance matrix of $\hat{\beta}$ is the inverse of the Fisher information matrix $I(\beta) = X^\top W X$, and this depends on $w_i = 1/\{\text{var}(Y_i \mid x_i) g'(\mu_i)^2\}$ where $\text{var}(Y_i \mid x_i) = \phi_i b''(\theta_i) = \sigma^2 b''(\theta_i)/m_i$.

To calculate standard errors and confidence intervals, we need to supply an estimate $\hat{\sigma}^2$ of σ^2. Despite that the maximum likelihood estimator of σ^2 is well-defined, it is more common to base an estimator of σ^2 on the Pearson X^2 statistic. The variance of Y_i can be written as $\text{var}(Y_i \mid x_i) = \phi_i V(\mu_i) = \sigma^2 V(\mu_i)/m_i$, where $V(\mu_i) = b''(\theta_i)$ and $\theta_i = {b'}^{-1}(\mu_i)$ (see Section 2.1). Hence, (2.12) can be written as

$$X^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{2.15}$$

Exercise 2.4. Suppose that H_0 is an adequate fit and that X^2 has an chi-squared distribution with $n-q-1$ degrees of freedom.

1. Show that
$$\hat{\sigma}^2 = \frac{1}{n-q-1} \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$
is an approximately unbiased estimator of σ^2.
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 $y_i - \hat{\mu}_i$ $(i = 1, \ldots, n)$. In fact, both the scaled deviance and Pearson X^2 statistic for a normal linear model (which is a GLM with normal distirbution and identity link function) are the sum of the squared residuals divided by σ^2. We can build on that observation to define residuals for other generalized linear models in a natural way.

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

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\widehat{\text{var}(Y_i \mid x_i)}^{1/2}} \qquad (i = 1, \ldots, n)\,.$$

Then, from (2.12), the statistic X^2 is the sum of the squared Pearson residuals.

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

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

where $\text{sign}(x) = 1$ if $x > 0$ and -1 if $x < 0$. Then, from (2.10), 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 Section 2.5, the expressions above are typically multiplied through by σ^2 to eliminate dependence on the unknown dispersion parameter.

So, 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 standardized by dividing through by $(1 - h_{ii})^{1/2}$, as in Section 1.4. If the model is adequate, the derived residuals

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

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

Checking GLMs using residuals is based on the same kind of diagnostic plots suggested for linear models in Section 1.7. Similarly, the Cook's distance C_j for linear models can be adapted for GLMs by using Pearson residuals.

Chapter 3

R practicals

3.1 Getting started

For running the code below, you will need the R packages `MASS` and `SMPracticals`. The `MASS` package (Venables & Ripley, 2002) is one of the recommended R package and is included with the binary distributions of R, so you should have it. The `SMPracticals` package (Davison, 2024) is the R package providing the datasets and a few functions for use with the practicals outlined in Davison (2003, Appendix A), and can be installed by running

```
install.packages("SMPracticals")
```

The packages can be loaded and attached by doing

```
library("MASS")
library("SMPracticals")
```

3.2 `trees` data

`trees` contains data on the volume of timber, height and girth (diameter) of 31 felled black cherry trees; girth is measured four feet six inches above ground (Atkinson, 1985, p. 63). The problem is to find a simple linear model for predicting volume from height and girth. See `?trees` for more details.

```
data("trees", package = "datasets")
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. The plots should be read from left to right, starting from the bottom row, and each plot corresponds to the ranges of values (from left to right) shown on the top plot for the conditioning variable. For the relationship of log volume and log girth, conditional on height we get:

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

Produce and interpret the conditioning plots on the orginal scale.

For an initial fit, we take a linear model and assess model fit using diagnostic plots:

```
m_trees <- glm(Volume ~ Girth + Height, data = trees)
summary(m_trees)
plot.glm.diag(m_trees)
```

> What do you make of the `m_trees` fit?

To assess the possibility of transformation:

```
boxcox(m_trees)
```

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

```
m_trees_log <- glm(log(Volume) ~ log(Girth) + log(Height), data = trees)
summary(m_trees_log)
plot.glm.diag(m_trees_log)
```

> Are the parameter estimates consistent with this model? Does it fit adequately? What advantage has it over the others for prediction of future volumes?

3.3 `salinity` data

`salinity` contains $n = 28$ observations on the salinity of water in Pamlico Sound, North Carolina (Atkinson, 1985, p. 48; Ruppert & Carroll, 1980). The response `sal` is the bi-weekly average of salinity. The other 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 from the analysis of the `tree` data set as a guide, find a model suitable for prediction of salinity from the covariates. The data contain at least one outlier.

3.4 `shuttle` data

`shuttle` contains the data in Davison (2003, Table 1.3) on O-ring failures for the space shuttle (Dalal et al., 1989).

```
data("shuttle", package = "SMPracticals")
row.names(shuttle) <- NULL
```

To fit a binomial logistic regression model with covariate temperature, we do:

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

> Try fitting with and without both covariates. To assess model fit, try
>
> ```
> plot.glm.diag(m_shuttle)
> ```
>
> Do you find these diagnostics useful?

3.5 `bliss` data

`bliss` provides data on mortality of flour-beetles as a function of dose of a poison (Bliss, 1935). To plot the death rates and fit a logistic regression model, we do:

```
data("bliss", package = "SMPracticals")
m_bliss_logit <- glm(cbind(r, m - r) ~ log(dose), family = binomial(), data = bliss)
summary(m_bliss_logit)

with(bliss, {
    plot(log(dose), r/m, ylim = c(0, 1), ylab = "Proportion dead")
    points(log(dose), fitted(m_bliss_logit), pch = 3, col = 2)
})
```

> Does the fit seem reasonble to you?
> Try again with the probit and cloglog link functions.

For example, for the cloglog link function we have:

```
m_bliss_cloglog <- glm(cbind(r, m-r) ~ log(dose), family = binomial("cloglog"), data = bliss)
with(bliss, {
    plot(log(dose), r/m, ylim = c(0, 1), ylab = "Proportion dead")
    points(log(dose), fitted(m_bliss_logit), pch = 3, col = 2)
    points(log(dose), fitted(m_bliss_cloglog), pch = 3, col = 3)
})
```

> Which link function fits best? Give a careful interpretation of the resulting model.

Bibliography

Atkinson, A. C. (1985). *Plots, transformations,and regression: An introduction to graphical methods of diagnostic regression analysis*. Clarendon Press.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, *22*(1), 134–167. https://doi.org/10.1111/j.1744-7348.1935.tb07713.x

Dalal, S. R., Fowlkes, E. B., & Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, *84*(408), 945–957. https://doi.org/10.2307/2290069

Davison, A. C. (2003). *Statistical models*. Cambridge University Press.

Davison, A. C. (2024). *SMPracticals: Practicals for use with davison (2003) statistical models*. https://CRAN.R-project.org/package=SMPracticals

Ruppert, D., & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, *75*(372), 828–838. https://doi.org/10.2307/2287169

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/