M3S2/M4S2 Statistical Modelling II

Recommended Literature

All the material used in this module can be found in various textbooks.

- McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall,
- Lee, Y., Nelder, J. A. & Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. CRC Press
- Hastie, T. & Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library
- Venables, W. N. & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media
- Dobson, A. J. & Barnett, A. (2008). *An introduction to generalized linear models*. CRC press
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press
- Ravishanker, N. & Dey, D. K. (2001). *A first course in linear model theory*. CRC Press
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, vol. 124. CRC press
- Fox, J. & Weisberg, S. (2011). *An R companion to applied regression.* Sage Publications

These books can be found (some electronically) in the college library.

Chapter 1. Introduction to Statistical Models

Consider situations where we observe some data y from experiments, events etc. We interpret the observation y as a realisation of some random variable Y.

A statistical model is the specification of the distribution of Y up to an unknown parameter θ. Often, the observations $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is a vector and $Y = (Y_1, \ldots, Y_n)$ is a random vector. In this case a statistical model is the specification of the joint distribution of $Y_1, \ldots, Y_n$ up to an unknown parameter θ.

Example Mobile Phones

Observations: $y_i =$ student i looks at their mobile in first 10 mins of the lecture.

Model: $Y_1, \ldots, Y_n$ iid, $P(Y_i = \text{true}) = 1 - P(Y_i = \text{false}) = \theta, \quad \theta \in [0, 1]$.

■

In many experiments and situations, the observations $Y_1, \ldots, Y_n$ do not have the same distribution. The distribution of $Y_1, \ldots, Y_n$ may depend on non-random quantities $x_1, \ldots, x_n$ called covariates.

Example Do students who attend more lectures get higher marks in Maths exams?

$y_i =$ average mark, $x_i =$ number of lectures attended

Model: $Y_i = a + bx_i + \epsilon_i,\ i = 1, \ldots, n,\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),\ \theta = (a, b, \sigma^2)$.

■

Model Fitting

We shall describe the process of the model fitting as follows:

1. Model specification — specify the distribution of the observations $Y_1, \ldots, Y_n$ up to unknown parameters.
2. Estimation of the unknown parameters of the model.
3. Inference — this involves constructing confidence intervals and testing hypotheses about the parameters.
4. Diagnostics — to check how well the model fits the data.

An "ideal" model should

- agree with the observed data reasonably well.
- not contain more parameters than are necessary.
- be easy to interpret.

Chapter 2. Normal Linear Models

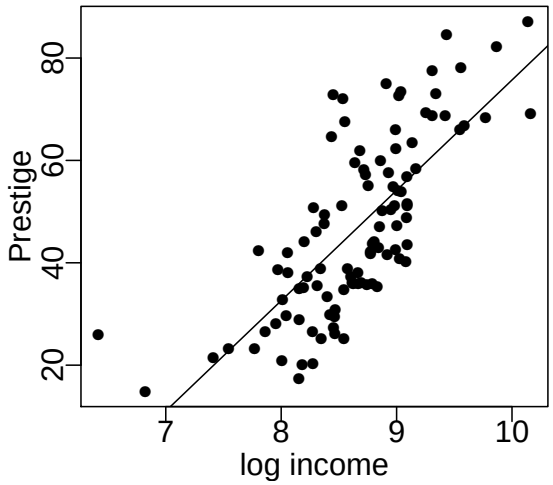
We start with a revision of linear models. We shall see how this simple model leads to some closed form results. However, we shall also point out that linear models have their limitations.

The intuition behind the fit

`prestige.R prestige-data.RData`

Consider the following dataset relating to the prestige of different occupations. Each row, relates to a different occupation, contains a prestige score and the log income. Can the (log) income be used to measure the prestige of occupations?

Occupation	Prestige	Log Income
gov.administrators	68.8	9.421
general.managers	69.1	10.16
accountants	63.4	9.135
purchasing.officers	56.8	9.09
chemists	73.5	9.036
physicists	77.6	9.308
biologists	72.6	9.019
architects	78.1	9.558
civil.engineers	73.1	9.339
mining.engineers	68.8	9.308
⋮	⋮	⋮

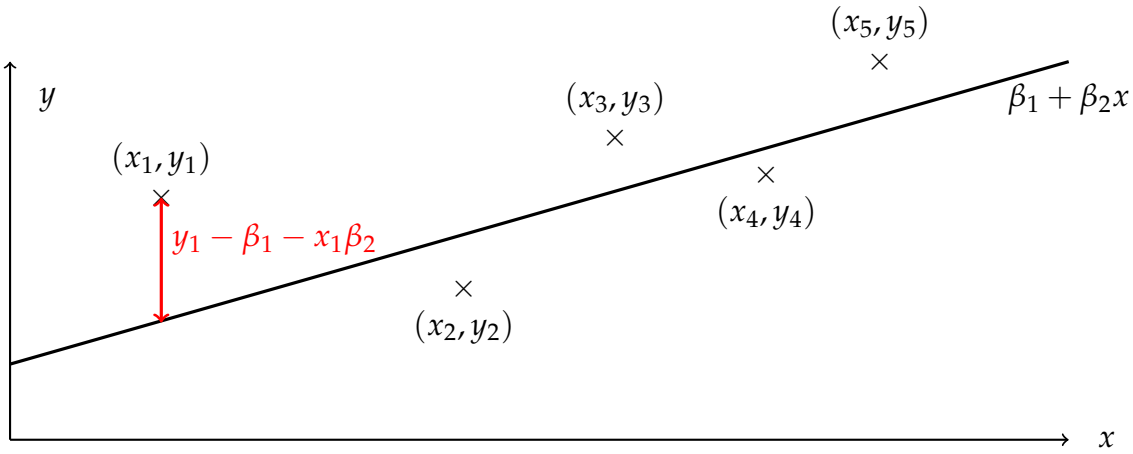


A possible linear model:

$$Y_i = \beta_1 + x_i\beta_2 + \epsilon_i, \quad i = 1, \ldots, n,$$

where

- Y_i is the outcome or response of interest (random variable) — observed realisations y_i.
- x_i is a covariate — observed constant
- β_1, β_2 are unknown parameters.
- $\epsilon_1, \ldots, \epsilon_n$ are iid error (random variables) with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$. Note that σ^2 is another unknown parameter — not observed.



The least-squares estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are defined as the minimisers of

$$\sum_{i=1}^{n}(y_i - \beta_1 - x_i\beta_2)^2.$$

2.1 Specification of Normal Linear Models

Definition. A linear model is defined as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- $\boldsymbol{Y}$ is the n-dimensional random vector of observations, called the response.
- $X \in \mathbb{R}^{n\times p}$ is the design matrix (known) which contains the covariates (predictors). Let $r := \text{rank}(X)$.
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter vector.
- $\boldsymbol{\epsilon}$ is the n-variate unobserved error vector such that $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$

All vectors are column vectors and are presented in bold.

Further if, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for some $\sigma^2 > 0$, then the linear model is normal. The normal linear model can be written as

$$\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n),$$

where I_n is the $n \times n$ identity matrix.

The assumption that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ is called the Normal Theory Assumption (NTA) and has two parts:

- the observations error are uncorrelated; and
- the variance of each Y_i is the same, namely, σ^2.

Examples of normal linear models

Example Consider the model where $Y_1, \ldots, Y_n$ are independent normally distributed random variables with variance σ^2 and $\mathrm{E}(Y_i) = \beta_1 + x_i\beta_2$ for $i = 1, \ldots, n$. This is a normal linear model as it can be written in the form

■

Example Consider the model where $Y_i \sim N(\mu_i, \sigma^2)$ independently with $\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}x_{i4}\beta_3$ for $i = 1, \ldots, n$. This is a normal linear model as it can be written as

■

Example Consider two school classes each consisting of 10 children. Denote the two classes as class A and class B. Suppose we model the observations of all children as normally distributed random variables with variance σ^2. Further suppose for child i from class A, denoted as Y_{iA}, that $\mathrm{E}(Y_{iA}) = \mu_A$ for all i and analogously the observation for child i from class B, denoted by Y_{iB}, that $\mathrm{E}(Y_{iB}) = \mu_B$. Then this is a normal linear model as it can be written as

■

Therefore, the linear model accommodates a large variety of different models, i.e. those with and without a constant intercept term and those structured according to classes/groups. However, ...

Example Suppose the independent observations $Y_1, \ldots, Y_n$ are normal random variables with $\mathrm{E}(Y_i) = \beta_1 + x_{i1}^{\beta_2}$ and $\mathrm{var}(Y_i) = \sigma^2$ for $i = 1, \ldots, n$. ■

... is *not* a linear model, as it cannot be written in the form $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$

2.2 Estimation

So far we have just introduced the normal linear model, where $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$. The joint probability density function of $\boldsymbol{Y}$ is

We now estimate the unknown parameter $\boldsymbol{\beta}$ using the maximum likelihood approach; that is, we want to find (keeping σ^2 fixed)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}),$$

where $L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})$ is the likelihood function. Typically, it is easier to obtain the MLE by maximising $\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \log L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})$.*

First, the log-likelihood of the data is

To find the maximum likelihood estimator (MLE), $\widehat{\boldsymbol{\beta}}$, we differentiate with respect to $\boldsymbol{\beta}$ (keeping σ^2 fixed and treating it as a constant) to get

$$-2X^T\boldsymbol{Y} + 2X^TX\boldsymbol{\beta}.$$

Solving this equation equal to zero give the so-called normal equations

$$X^TX\widehat{\boldsymbol{\beta}} = X^T\boldsymbol{Y}.$$

If X^TX is invertible (iff X has full rank), the normal equations can be solved directly giving the MLE as

$$\widehat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1} X^T\boldsymbol{Y}. \tag{2.1}$$

The MLE for σ^2 can be obtained using the same method.

*The probability density function (pdf) $f(y;\theta)$ is considered a function of y for fixed or given θ whereas the likelihood $L(\theta;y)$, and therefore the log-likelihood $\ell(\theta;y) = \log L(\theta;y)$, is considered a function of θ for a particular data y observed.

Properties of the Maximum Likelihood Estimator

The MLE of $\boldsymbol{\beta}$ has the following properties.

- $\widehat{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$.
- $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$
- $\text{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^TX)^{-1}$.

The Gauss-Markov theorem asserts that for the linear model, the MLE $\widehat{\boldsymbol{\beta}}$ is in a certain sense optimal: it is the *best linear unbiased estimator* (BLUE). An estimator $\widehat{\gamma}$ is called linear if there exists $\boldsymbol{L} \in \mathbb{R}^n$ such that $\widehat{\gamma} = \boldsymbol{L}^T\boldsymbol{Y}$.

Theorem 1. (The Gauss Markov Theorem for full rank linear models) Assume a linear model with unknown parameter $\boldsymbol{\beta}$, a design matrix of full rank and $\text{cov}(\epsilon) = \sigma^2 I_n$. Fix any $\boldsymbol{c} \in \mathbb{R}^p$ and let $\widehat{\boldsymbol{\beta}}$ be the least squares estimator of the linear model. Then the estimator $\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators for $\boldsymbol{c}^T\boldsymbol{\beta}$.

For Normal linear models, the least squares estimator i.e. the $\boldsymbol{\beta}$ that minimises:

$$S(\boldsymbol{\beta}) := (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})$$

and the MLE, $\widehat{\boldsymbol{\beta}}$, are the same (see problem sheet).

Before we present an estimator of the variance σ^2, it is necessary to introduce the notion of projection matrices.

Projection Matrices

Let L be a linear subspace of $\mathbb{R}^n$, where $\dim L = p \leq n$.

Definition. $P \in \mathbb{R}^{n\times n}$ is projection matrix onto L, if

1. $P\boldsymbol{x} = \boldsymbol{x}$ for all $\boldsymbol{x} \in L$,
2. $P\boldsymbol{x} = \boldsymbol{0}$ for all $\boldsymbol{x} \in \underbrace{L^\perp = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^T\boldsymbol{y} = \boldsymbol{0}\ \forall \boldsymbol{y} \in L\}}_{\text{orthogonal complement}}$

It follows that

Lemma 2. P is a projection matrix $\iff$ $\overbrace{P^T = P}^{\text{symmetric}}$ and $\overbrace{P^2 = P}^{\text{idempotent}}$.

Definition. $\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the MLE, is called the vector of fitted values. In the case where X has full rank, $\widehat{\boldsymbol{Y}} = X(X^TX)^{-1}X^T\boldsymbol{Y}$.

It turns out that

$$P := X(X^TX)^{-1}X^T$$

is a projection matrix; since (by Lemma 2)

where the first equation holds because $(A^{-1})^T = (A^T)^{-1}$ for a non-singular matrix A.

We now define the residuals and the residual sum of squares.

Definition. $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is called the vector of residuals.

Notice that the residuals can be written as

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - P\boldsymbol{Y} = (I_n - P)\boldsymbol{Y}.$$

It can be shown that $I_n - P$ is also a projection matrix (see problem sheet). The projection matrix P is sometimes referred to as the hat matrix, as it puts the hat on Y.

Definition. $\text{RSS} = \boldsymbol{e}^T\boldsymbol{e}$ is called the residual sum of squares.

The RSS quantifies the departure of the data from the model - ideally we want a small RSS value. It is also the minimum of $S(\boldsymbol{\beta})$.

Theorem 3. $\widehat{\sigma}^2 := \frac{\text{RSS}}{n-r}$ is an unbiased estimator of σ^2.

Proof. See problem sheet □

If we derived the MLE for σ^2 in the usual way — maximising the log-likelihood and solving equal to zero, we would obtain a biased estimator. To see this:

$$\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{Y}) =$$

$$\frac{\partial \ell}{\partial(\sigma^2)} =$$

Solving $\frac{\partial \ell}{\partial(\sigma^2)} = 0$ gives:

$$\widetilde{\sigma}^2 =$$

Notice that if we plug in $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, we have $\widetilde{\sigma}^2 = \text{RSS}/n$. Since $\text{E}(\text{RSS}) = \sigma^2(n-r)$, it follows that $\text{E}(\widetilde{\sigma}^2) = \sigma^2(n-r)/n$. This is why we use the estimator in Theorem 3.

Example In the following, we consider a normal linear model where the MLE of β can be derived explicitly. Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \sim N(\mu_i, \sigma^2)$ with

$$\mu_i = \beta_1 + \beta_2 a_i,$$

where $a_1, \ldots, a_n$ are known deterministic constants. Take $n \geq 2$. Then

Assume that not all a_is are equal to ensure X has full rank. Then

Now we can write the MLE $\widehat{\boldsymbol{\beta}}$ explicitly as $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$:

■

Identifiability

In this section we have discussed how to estimate the unknown parameter, $\boldsymbol{\beta}$, in the normal linear model. It can be the case that two parameter values lead to the same distribution for the observed data. This occurs when $r < p$ and we say that $\boldsymbol{\beta}$ is not identifiable.

Example in R

`gas.R gas-data.RData`

Suppose we have data consisting of household gas consumption and average external temperature (see table below). How is the external temperature related to the amount of gas consumed by a household?

#	Temp	Gas	#	Temp	Gas
1	-0.8	7.2	14	6.0	4.4
2	-0.7	6.9	15	6.2	4.5
3	0.4	6.4	16	6.3	4.6
4	2.5	6.0	17	6.9	3.7
5	2.9	5.8	18	7.0	3.9
6	3.2	5.8	19	7.4	4.2
7	3.6	5.6	20	7.5	4.0
8	3.9	4.7	21	7.5	3.9
9	4.2	5.8	22	7.6	3.5
10	4.3	5.2	23	8.0	4.0
11	5.4	4.9	24	8.5	3.6
12	6.0	4.9	25	9.1	3.1
13	6.0	4.3	26	10.2	2.6

Table 2.1: Gas consumption data

We take Gas as the response, Y_i, and Temp as the covariate x_i. Suppose that we use a linear normal model to explain the data; where Y_i are independent $N(\mu_i, \sigma^2)$ with $\mu_i = \beta_1 + \beta_2 x_i$ for $i = 1, \ldots, 26$. For this model, we have

Then to compute the MLE for $\widehat{\boldsymbol{\beta}}$ with the following commands. First, we assemble the vectors of observations $\boldsymbol{Y}$ and the design matrix X:

```
> Y <- dat$Gas
> X <- cbind(1,dat$Temp)
```

As $X^T X$ has full rank, we can compute its inverse:

```
> XtXinv <- solve(t(X)%*%X)
```

The MLE $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$ can be computed as follows:

```
> betahat <- XtXinv%*%t(X)%*%Y
> betahat
```

```
           [,1]
[1,]  6.8538277
[2,] -0.3932388
```

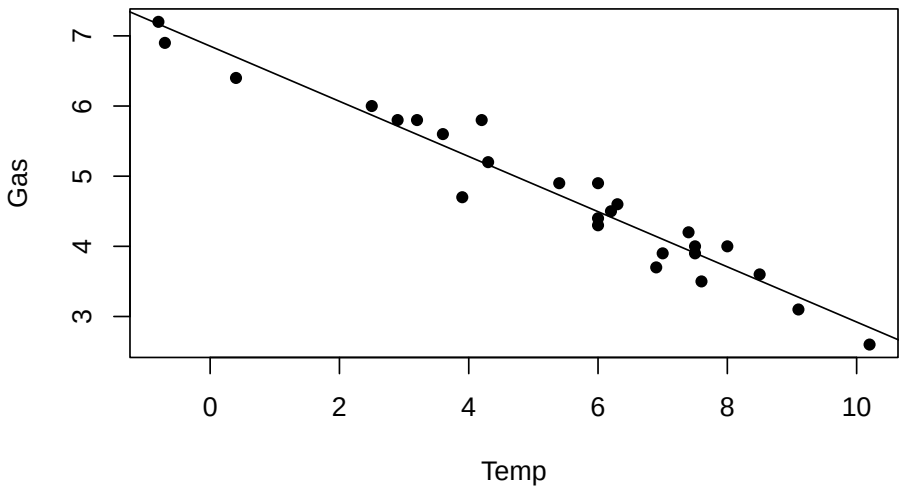
Note that we can determine $\widehat{\boldsymbol{\beta}}$ without explicitly computing $(X^T X)^{-1}$,

```
> betahat <- solve(t(X)%*%X, t(X)%*%Y)
> betahat
```

```
           [,1]
[1,]  6.8538277
[2,] -0.3932388
```

Then to plot the model fit we can use

```
> plot(dat,pch=16) #pch=16 selected for better presentation
> xs <- seq(-2,12,by=0.5) #dummy x values
> lines(x=xs,y=betahat[1]+xs*betahat[2])
```



Further, the residual sum of squares (RSS) is

```
> ehat <- Y-X%*%betahat
> RSS <- t(ehat)%*%ehat
> RSS
```

```
         [,1]
[1,] 1.899568
```

Lastly, we can compute $\widehat{\sigma}^2$:

```
> sig2hat <- RSS/(26-2)
> sig2hat
```

```
           [,1]
[1,] 0.07914867
```

■

2.3 Inference

In section 2.1 we specified a normal linear model. Then in section 2.2 we discussed how to estimate the parameters of the linear model.

In this section we shall consider two main tools of statistical inference

1. **Confidence Intervals** The width of these intervals indicate uncertainty of parameter estimates and inferences.

2. **Hypothesis Testing** These are used to compare how two related models fit the data. To compare models we require a measure of their **goodness of fit**. We shall present goodness of fit statistics based on the log-likelihood function.

In order to construct confidence intervals and measure the goodness of fit of normal linear models, we require sampling distributions. Note that the derivations of these sampling distributions are omitted — see Statistical Modelling I for details.

Before we begin, here is an example of constructing confidence interval and performing a hypothesis test as a reminder:

Example

Confidence Interval
Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables with $\sigma^2 > 0$ known. We know that

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This is our pivotal quantity for μ. We can construct a $1 - \alpha$ confidence interval for μ as follows:

■

Notes

- $\left(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}\right)$ is a random interval.
- $\left(\overline{x} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}\right)$ is the observed interval.
- We could use asymmetric values, but symmetrical values i.e. using $\pm z_{\alpha/2}$ gives the shortest intervals in this example.

Definition. A $(1-\alpha)$ confidence interval for θ is a random interval I that contains the true parameter with probability $\geq 1-\alpha$, i.e.

$$P_\theta(\theta \in I) \geq 1-\alpha \quad \forall \theta \in \Theta$$

The interval, I, can be any type of interval. For example, if L and U are random variables with $L \leq U$ then I could be the open interval (L, U), the closed interval $[L, U]$, the unbounded interval $[L, \infty)$, etc.

Example

Hypothesis Test (Continuing the previous example)
If we wish to test the null hypothesis $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ a test rejection region is $\{x : |\overline{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$. So, the null hypothesis, H_0, is not rejected for samples with

$$\overline{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu_0 < \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}.$$

This test is constructed to have significance level α i.e. $\mathrm{P}(H_0 \text{ is not rejected} | \mu = \mu_0) = 1-\alpha$. ■

Example

Pipes
In a manufacturing process that produces pipes, historical data suggests that the diameter of the pipes are $N(30, 1.5^2)$. The process is modified but the variance remains the same (at 1.5^2). After the modification $n = 13$ pipes are produced and the average diameter of the pipes is 32.05; so that the new diameter $\sim N(\mu, 1.5^2)$. Let's construct a 95% confidence interval for μ:

$$n = \quad , \quad \sigma = \quad , \quad \overline{x} = \quad , \quad \alpha = \quad , \quad z_{\alpha/2} =$$

which gives the 95% confidence interval: . If we were to test the hypothesis $H_0 : \mu = 30$ versus $H_1 : \mu \neq 30$; that is the null is that the process has not changed, we would reject the null hypothesis as 30 is not contained in the interval.

■

Confidence Intervals and Regions

We now discuss how to construct confidence regions for the model parameter $\boldsymbol{\beta}$. First we require a pivotal quantity for σ^2. It turns out that

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-r}, \tag{2.2}$$

where $r = \text{rank}(X)$ and χ^2_k denotes the chi-squared distribution with k degrees of freedom. Then it follows that, for any $\boldsymbol{c} \in \mathbb{R}^p$, and when X has full rank i.e. $\text{rank}(X) = p$

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}\frac{\text{RSS}}{n-p}}} \sim t_{n-p}, \tag{2.3}$$

where t_k denotes the student-t distribution with k degrees of freedom. This follows from

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}\sigma^2}} \sim N(0,1) \tag{2.4}$$

and recalling that:

By selecting $\boldsymbol{c}$ appropriately, we can use (2.3) to construct confidence intervals for components of $\boldsymbol{\beta}$ e.g. β_1

In fact, we can go further and construct confidence regions for the vector $\boldsymbol{\beta}$ by using

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\text{RSS}} \frac{n-p}{p} \sim F_{p,n-p}, \tag{2.5}$$

where $F_{a,b}$ denotes the F-distributions with degrees of freedom a and b.

Example Suppose we use a normal linear model with known $\sigma^2 = 1$ and design matrix

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

A hypothesis test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ is constructed as follows:

■

Note

- If σ is known, use the normal distribution to form confidence intervals. If σ is unknown, use the t-distribution.
- We can now construct confidence intervals (and regions) and conduct hypothesis test for $\boldsymbol{\beta}$ — see examples at the beginning of this section.

Warning

Following from (Wasserman, 2013, p. 111):

"There is sometimes much confusion over the interpretation of a confidence interval. A confidence interval is not a probability statement about μ (or some other parameter of interest) since μ is a fixed quantity, not a random variable. Some texts interpret confidence intevals as follows":

if I repeat the experiment over and over again, the interval will contain the true parameter value 95% *of the time.*

"This is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this":

On day 1, you collect data and construct a 95% confidence interval for a parameter θ_1. *On day 2, you collect new data and construct a 95% confidence interval for an unrelated parameter* θ_2. *On day 3, you collect new data and construct a 95% confidence interval for an unrelated parameter* θ_3. *You continue this way constructing confidence interval for a sequence of unrelated parameters* θ_1,, θ_2, ... *Then the 95% of the time your intervals will contain the true parameter. There is no need to introduce the idea of repeating the same experiment over and over.*

p-values

Sometimes one does not want to specify the significance level α in advance. In this case, the so-called p-value is reported.

$$p = \sup_{\theta \in \Theta_0} P_\theta(\text{observing something at least as extreme as the observation}).$$

Then we reject H_0 iff the p-value $p \leq \alpha$.

Hypothesis Testing: The F-test

Suppose we wish to test whether a sub-model of a linear model $\mathrm{E}(\boldsymbol{Y}) = X\boldsymbol{\beta}$ is better i.e. we wish to test

$$H_0 : \mathrm{E}(\boldsymbol{Y}) \in \mathrm{span}(X_0) \quad \text{vs} \quad H_1 : \mathrm{E}(\boldsymbol{Y}) \notin \mathrm{span}(X_0),$$

for some matrix X_0 with $\mathrm{span}(X_0) \subset \mathrm{span}(X)$. The span is the linear space spanned by the column vectors of the matrix.

Example Consider comparing the model $H_0 : Y_i = \beta_1 + \beta_2 x_{1,i} + \epsilon_i$ against $H_1 : Y_i = \beta_1 + \beta_2 x_{1,i} + \beta_3 x_{2,i} + \epsilon_i$. ■

Theorem 4. Under $H_0 : \mathrm{E}(\boldsymbol{Y}) \in \mathrm{span}(X_0)$,

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}} \frac{n-r}{r-s} \sim F_{r-s,n-r},$$

where $r = \mathrm{rank}(X)$ and $s = \mathrm{rank}(X_0)$.

Think: what is the difference between the F-test and the t-test?

A possible use of this F-test is the ANOVA test, which we now discuss.

One-Way Analysis of Variance (ANOVA)

Suppose we have m groups of observations, each group consisting of k observation. Label Y_{ij} as the jth observation from ith group. Assume the model is

$$\mathrm{E}(Y_{ij}) = \mu + \beta_i \quad i = 1, \ldots, m; j = 1, \ldots, k,$$

and that $\mathrm{var}(Y_{ij}) = \sigma^2$ with all observations independent. We want to test the null hypothesis that all the observations Y_{ij} come from the same population; that is

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_m,$$

or equivalently

$$H_0 : \ \mathrm{E}(Y_{ij}) = \mu \quad i = 1, \ldots, m; j = 1, \ldots, k.$$

We can test this hypothesis using the F-test (Theorem 4).

2.4 Prediction

Once we have fitted a model, we may use it to predict outcomes. Prediction, as interpreted here, means determining plausible values for unknown values of the response variable, rather than prediction of the future (forecasting).

Suppose we have fitted a linear model with design matrix X to obtain the estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$. For given covariates, $\boldsymbol{x}_\star$, the predicted (expected) response is $\widehat{y}_\star = \boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}}$. To ascertain the uncertainty in this prediction, we need to clear about the type of prediction we are making.

Suppose we have fitted a linear regression model that predicts the selling price of homes in a given area that is based on predictors such as the number of bedrooms and closeness to a major highway. There are two kinds of predictions that can be made for a given $x_\star$:

- Suppose a specific house comes on the market with characteristics $x_\star$. Its selling price will be $\boldsymbol{x}_\star^T\boldsymbol{\beta} + \epsilon$. Since $\mathrm{E}(\epsilon) = 0$, the predicted price is $\boldsymbol{x}_\star^T\widehat{\beta}$, but in determining the variance of this prediction, we must include the variance of ϵ.

- Suppose we ask the question: "What would a house with characteristics $\boldsymbol{x}_\star$ sell for on average?" This selling price is $\boldsymbol{x}_\star^T\boldsymbol{\beta}$ and is again predicted by $\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}}$ but now only the variance in $\widehat{\boldsymbol{\beta}}$ needs to be taken into account.

We have that $\mathrm{var}(\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}}) = \boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star\sigma^2$. So the variance of the prediction for a single unseen observation is

$$\mathrm{var}(\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}} + \epsilon) = \mathrm{var}(\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}}) + \mathrm{var}(\epsilon) = \boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star\sigma^2 + \sigma^2,$$

since ϵ is independent of the random variable $\widehat{\boldsymbol{\beta}}$.

So a $100(1-\alpha)\%$ confidence interval for a single unseen observation is

$$\widehat{y}_\star \pm t_{n-p}^{(\alpha/2)}\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star},$$

where $P(T \leq t_{n-p}^{(\alpha/2)}) = 1 - \alpha/2$, $T \sim t_{n-p}$. The confidence interval for the mean response for given $\boldsymbol{x}_\star$ is

$$\widehat{y}_\star \pm t_{n-p}^{(\alpha/2)}\widehat{\sigma}\sqrt{\boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star}.$$

These confidence intervals are sometimes referred to as prediction intervals.

2.4.1 Worked Example

We now show how predictions and their confidence intervals can be computed in R. Consider the Galapagos dataset (which is in the faraway library).

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58.00	23.00	25.09	346.00	0.60	0.60	1.84
Bartolome	31.00	21.00	1.24	109.00	0.60	26.30	572.33
Caldwell	3.00	3.00	0.21	114.00	2.80	58.70	0.78
Champion	25.00	9.00	0.10	46.00	1.90	47.40	0.18
Coamano	⋮	⋮	⋮	⋮	⋮	⋮	⋮

This dataset consists of 30 Islands and seven variables. We shall be interested in Species which is the number of species of tortoise found on the island. Lets start by fitting a linear model:

```
> library("faraway")
> X <- cbind(1,gala$Area,gala$Elevation,
+            gala$Nearest,gala$Scruz,gala$Adjacent)
> y <- gala$Species
> XTX <- solve(t(X)%*%X)
> betahat <- XTX%*%t(X)%*%y
> ehat <- y-X%*%betahat
> RSS <- sum(ehat^2)
> sig2hat <- RSS/(30-6)
```

Now suppose we want to predict the number of species of tortoise on an island with predictors $0.08, 93, 3, 12, 0.34$. The prediction (expected) response from the model is:

```
> xstar <- c(1,0.08,93,3,12,0.34) # add 1 for intercept.
> ystar <- sum(xstar*betahat)
> ystar
```

```
[1] 33.89224
```

Now consider the two types of confidence intervals: we must decide whether we are predicting the number of species on one new island or the mean response for all islands with the same predictors $\boldsymbol{x}_\star$.

For a mean response, we would use the second type of interval. First, we require the critical value from the t-distribution

```
> qcrit <- qt(0.975,24)
> qcrit
```

```
[1] 2.063899
```

The width of the confidence interval is:

```
> ciw <- qcrit*sqrt(sig2hat)*sqrt(xstar%*%XTX%*%xstar)
> ciw
```

```
         [,1]
[1,] 32.41583
```

and so the interval is

```
> c(ystar-ciw,ystar+ciw)
```

```
[1]  1.476401 66.308070
```

The prediction interval for a single future response is:

```
> ciw <- qcrit*sqrt(sig2hat)*sqrt(1+xstar%*%XTX%*%xstar)
> c(ystar-ciw,ystar+ciw)
```

```
[1] -96.06219 163.84667
```

2.5 Diagnostics

Coefficient of Determination

A way of measuring the goodness of fit of a linear model is by inspecting the coefficient of determination, which we now explain. In the simplest model with only an intercept term

$$\boldsymbol{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_1 + \boldsymbol{\epsilon}, \quad \mathrm{E}(\boldsymbol{e}) = 0$$

we have the RSS $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. Larger models with more parameters and large design matrices will have a smaller RSS.

For models that include an intercept term, a measure of the quality of a linear model is

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

which is called the coefficient of determination or R^2 statistic. Notice that $0 \leq R^2 \leq 1$ and $R^2 = 1$ corresponds to the "perfect" model.

Intuition

For the intercept only models, the $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is an estimator of σ^2 — call this the *total variance*. Then

$$\frac{\mathrm{RSS}}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \approx \frac{\text{var. of model}}{\text{variance}}.$$

Therefore, $R^2 \approx$ proportion of total variance of the data that is explained by the model.

There is also an alternate version of the R^2 statistic called the adjusted coefficient of determination defined as

$$\overline{R}^2 = 1 - \frac{\mathrm{RSS}/(n-p)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}.$$

Notice that the RSS terms are divided by the models' respective degrees of freedom - thus $\overline{R}^2$ attempts to account for the degrees of freedom for each model.

2.5.1 The Danger of Using R^2 : Anscombe Quartet

Using the R^2 and the adjusted R^2 summary statistics alone can be dangerous, as illustrated by Anscombe's quartet. Anscombe's quartet is 4 datasets — see table below:

	x1	x2	x3	x4	y1	y2	y3	y4
1	10.00	10.00	10.00	8.00	8.04	9.14	7.46	6.58
2	8.00	8.00	8.00	8.00	6.95	8.14	6.77	5.76
3	13.00	13.00	13.00	8.00	7.58	8.74	12.74	7.71
4	9.00	9.00	9.00	8.00	8.81	8.77	7.11	8.84
5	11.00	11.00	11.00	8.00	8.33	9.26	7.81	8.47
6	14.00	14.00	14.00	8.00	9.96	8.10	8.84	7.04
7	6.00	6.00	6.00	8.00	7.24	6.13	6.08	5.25
8	4.00	4.00	4.00	19.00	4.26	3.10	5.39	12.50
9	12.00	12.00	12.00	8.00	10.84	9.13	8.15	5.56
10	7.00	7.00	7.00	8.00	4.82	7.26	6.42	7.91
11	5.00	5.00	5.00	8.00	5.68	4.74	5.73	6.89

If we proceed to fit a linear model for the 4 datasets, we obtain the following parameter estimates and R^2 values:

	β_0	β_1	R^2	Adj. R^2
Model 1	3.000	0.500	0.667	0.629
Model 2	3.001	0.500	0.666	0.629
Model 3	3.002	0.500	0.666	0.629
Model 4	3.002	0.500	0.667	0.630

Table 2.2: Summary statistics

The parameter estimates and the R^2 statistic values are approximately the same for the 4 datasets. So these models are (more or less) equal, in terms of the fit and summary statistics. Lets do what we should have done at the start; look at the data:

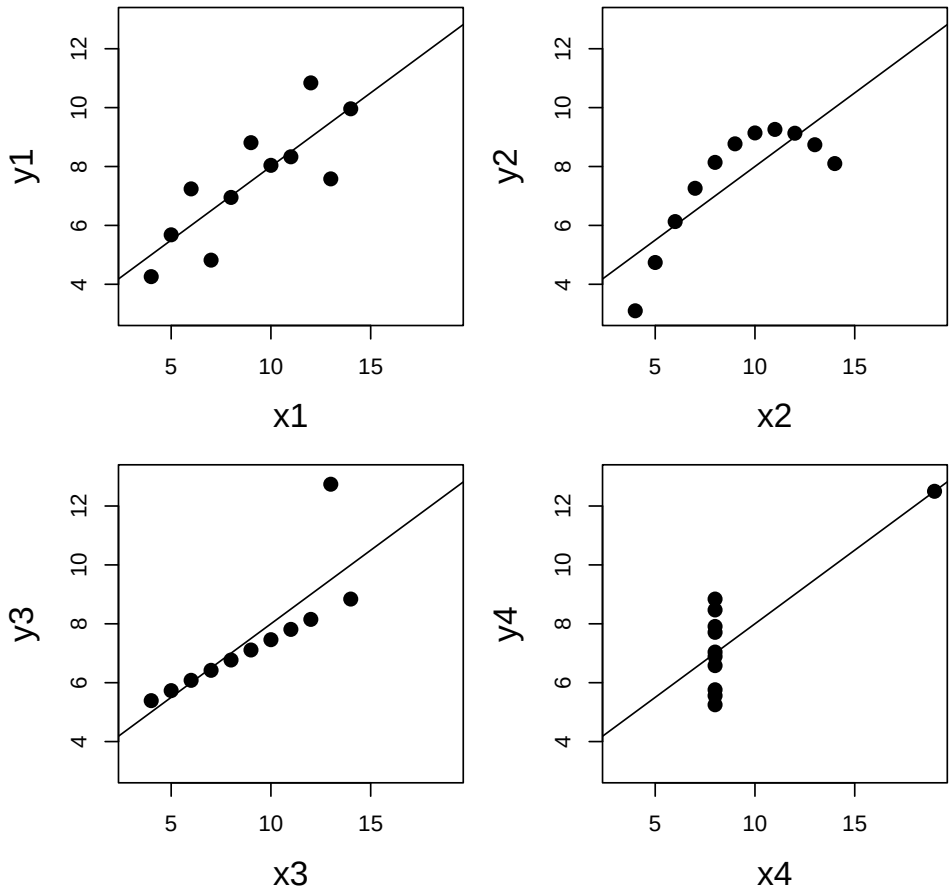


Figure 2.1: Anscombe's Quartet

Clearly, the data is different for each dataset and the fitted linear model may be inappropriate in some cases.

Note

- Looking at plots of the data *before* fitting is important.
- Using a summary statistic, such as R^2, on its own can be dangerous.

2.5.2 Outliers

An outlier is an observation that does not conform to the general pattern of the rest of the data. Outliers can occur due to, among other causes, error in the data recording, the data is a mixture of two or more populations and when the model requires improvement. These outlying cases may involve large residuals and often have dramatic effects on the fitted function. Therefore, it is important to study these outlying cases carefully.

A case may be outlying or extreme with respect to its Y value, its X value(s) or both.

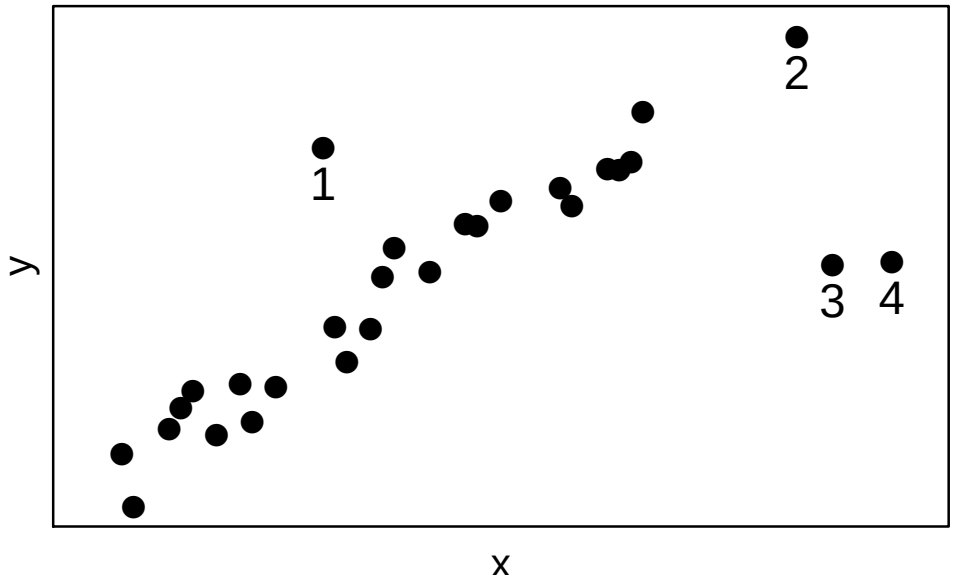


Figure 2.2: Illustration of different types of outliers.

In the scatter plot (Fig. 2.2) case 1 is outlying with respect to its Y value, given X. Cases 2, 3 and 4 are outlying with respect to their X values since they have much larger X values; cases 3 and 4 are also outlying with respect to their Y values given X.

Not all outlying cases have a strong influence on the fit. Case 1 may not be too influential because a number of other cases have similar X values. This will keep the fitted function from being displaced too far by the outlying case. Likewise, case 2 may not be too influential because its Y value is consistent with the regression relation displayed by the nonextreme cases. Cases 3 and 4, on the other hand, are likely to be very influential in affecting the fit of the regression function.

Notes

For models with one or two predictor variables, it is relatively easy to identify outlying cases, with respect to their X or Y values by means of box-plots, scatter plots, residuals etc and to study whether they are influential in affecting the fitted regression function. However, when more than two predictor variables are included in the model, identification of outliers by simple plots becomes difficult.

Residuals

A practical method for detecting outliers is to look at observations with residuals that are "large". We have

$$\begin{aligned} \boldsymbol{e} &= \\ \mathrm{E}(\boldsymbol{e}) &= \\ \mathrm{cov}(\boldsymbol{e}) &= \end{aligned}$$

We now discuss some refined residuals for identifying observations with outlying Y values. In the following, we are interested in identifying cases that are multivariable outliers with respect to their X values. We shall assume a design matrix of full rank.

Standardized Residuals

To use residuals for detecting outlying X observations, it is important to consider their variance — note that the variance of each residual may be substantially different. Therefore it is appropriate to consider the magnitude of each e_i relative to its estimated standard deviation. An estimator of the standard deviation of e_i is

$$\sqrt{\widehat{\sigma}^2(1-h_{ii})},$$

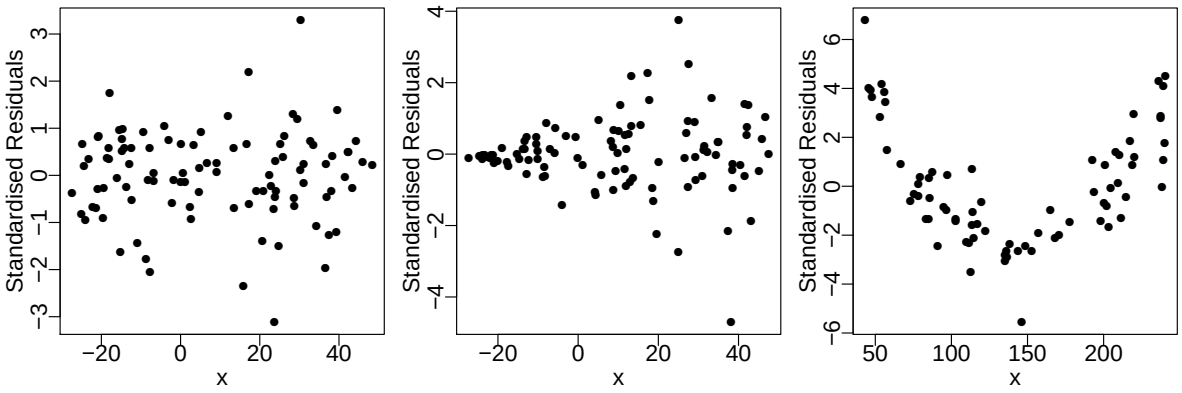
where $h_{ii} \equiv P_{ii}$. Recall that in practice, σ^2, is typically unknown. Then the **standardized residual** is

$$r_i := \frac{e_i}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}.$$

The studentized residuals r_i have constant variance when the model is appropriate. Studentized residuals often are called **internally studentized residuals**.

Using the plugin estimate for σ^2 means that we lose the normality, however the standardised residuals should be approximately normal. A plot of r_i against some other variable should not reveal any trends or patterns.

Residual Plots



Fine | Heteroscedastic error (non-constant variance) | Non-linear

Leverages

We may be interested in how much each observation influences the model fit. For instance, consider the residuals $\boldsymbol{e}$ where

$$\text{var}(e_i) = \sigma^2(1 - h_{ii}).$$

An observation's leverage is related to the variance of its residual.

Definition. The ith observation in a linear model has leverage equal to h_{ii}.

If an observation has leverage close to 1, its residual has a small variance. Notice that the leverage only depends on the design matrix X.

In practice, the leverages are compared to the "average" r/n and looking for $h_{ii} > 2r/n$, as $\sum_{i=1}^{n} h_{ii} = \text{trace}(P) = \text{rank}(X) = r$

Deleted Residuals

Intuition:

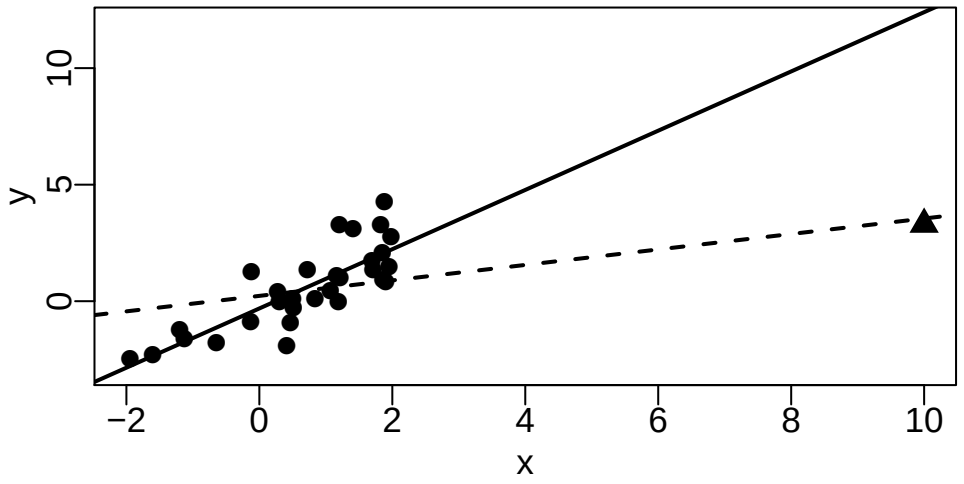


Figure 2.3: Illustration of observations with high leverage. Dotted line is a linear model fit including ▲; Solid line is a linear model without ▲

Fit the model without the ith observation. Estimate (or predict) the ith observation using x_i; denote the expected value by $\widehat{Y}_{(i)}$. The difference between the actual observed value Y_i and the estimated expected value $\widehat{Y}_{(i)}$ is the **deleted residual** denoted by d_i. Specifically,

$$d_i = Y_i - \widehat{Y}_{(i)}$$

An algebraically equivalent expression for d_i that does not require recomputation of the fit omitting the ith case is

$$d_i = \frac{e_i}{1 - h_{ii}}.$$

The estimated variance of the deleted residual for the ith case, d_i, is

$$\widehat{\sigma}^2_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)$$

where x_i is the predictor vector for the ith case, $\widehat{\sigma}^2_{(i)}$ is the mean square error when the ith case is omitted, and $X_{(i)}$ is the design matrix with the ith row deleted. An equivalent expression for this variance is

$$\frac{\widehat{\sigma}^2_{(i)}}{1 - h_{ii}}$$

Note: The estimated variance of the deleted residual can be obtained by "predicting" the ith observation and using section 2.4.

Studentized Deleted Residuals

As before, we can "studentized" the deleted residuals, by dividing it by its standard deviation. The **studentized deleted residual**, denoted by t_i, is

$$t_i = \frac{d_i}{\sqrt{\frac{\widehat{\sigma}^2_{(i)}}{1-h_{ii}}}} = \frac{e_i}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}$$

The studentized deleted residual t_i is also called an **externally studentized residual**.

2.5.3 Cook's Distance

Another way to measure the influence of the observation is to consider the change or influence it has on the estimator $\boldsymbol{\beta}$. One such measure is Cook's distance

$$C_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T X^T X (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\text{RSS}/(n-p)}, \quad \text{(leave-one-out estimator)}$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimator calculated without using the ith observation. The rule of thumb is to look at observations with C_i close to 1.

2.5.4 Distributional Checks

If $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma^2)$ distributed random variables, then

$$P(Y_i \leq y) = P\left(\frac{Y_i - \mu}{\sigma} \leq \frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

Then the empirical cdf

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) \to \Phi\left(\frac{x-\mu}{\sigma}\right) \quad n \to \infty.$$

Therefore,

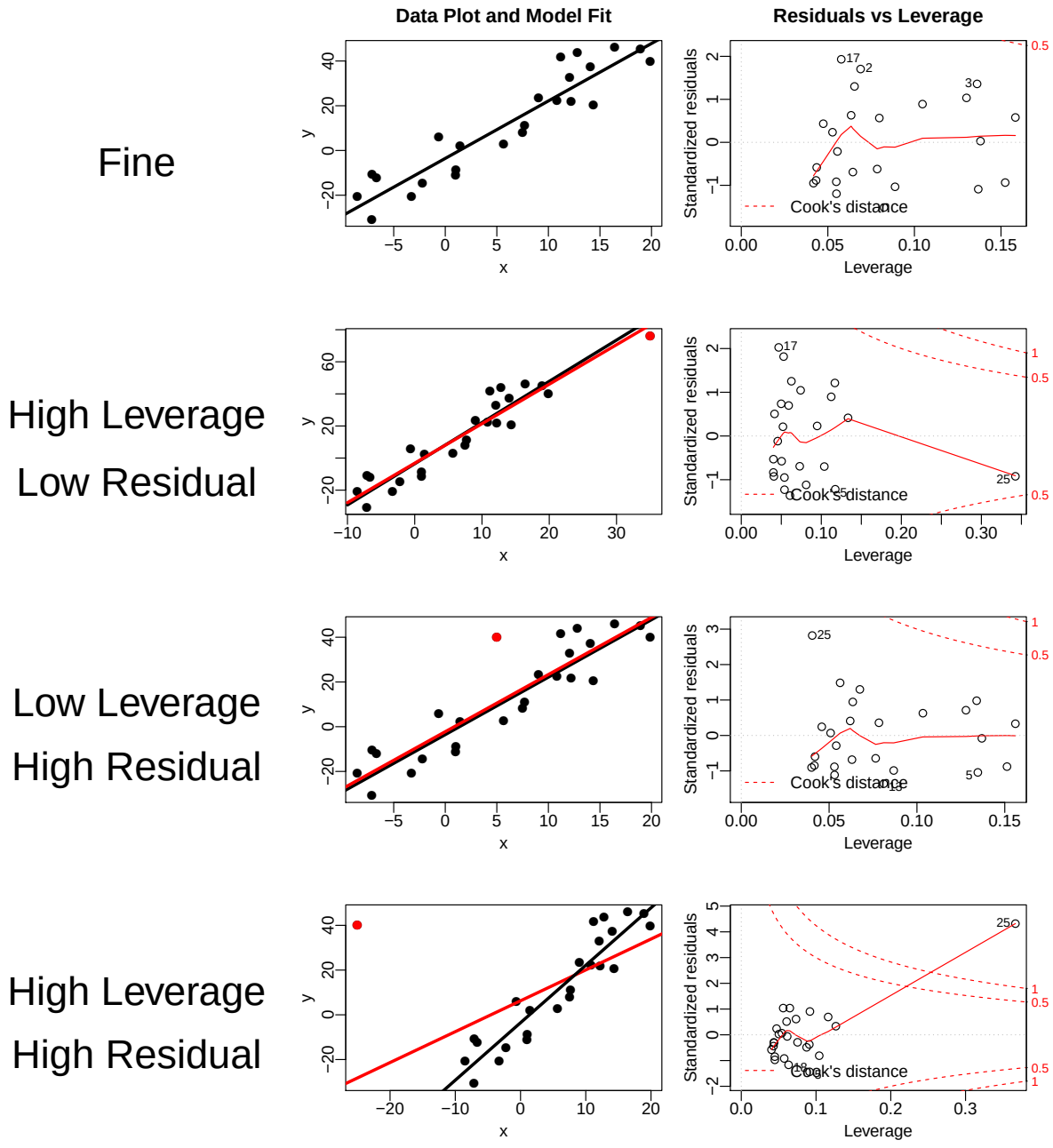
$$\Phi^{-1}(F_n(y_i)) \approx \frac{y_i - \mu}{\sigma}.$$

A plot of $\Phi^{-1}(F_n(y_i))$ against y_i should look like a straight line. This is called a Q-Q plot.

2.5.5 Interpretation of Cook's Distance

`cookdistance-data.RData`

We now investigate residuals, leverage and Cook's distance in more detail. Consider 4 artificial data sets which are presented below each with a normal linear model fitted. Further, a plot of the residuals against the leverages is presented.



The red data point is a suspicious data point i.e. either high leverage, high residual (in absolute value) or both. The red line is the fit including the red data point, and the black line is the fit excluding the red data point.

Note

A data point with a high leverage has the potential to change the model fit significantly. We also need to inspect the residual to see how much the point actually affects the fit. This is why we look at Cook's distance — it combines the leverage and standardised residual as it can be written equivalently as

$$C_i = r_i^2 \frac{h_{ii}}{(1 - h_{ii})r}$$

where r_i is the standardised residual and $r = \text{rank}(X)$.

The Cook's distance is presented in the plots on the previous page, in the form of contours — for instance those points lying between the 0.5 red dotted lines have a Cook's distance less than or equal to 0.5. The red dashed lines are contour lines of Cook's distance.

2.6 Worked Example

`carbo.R carbo-data.RData`

The data in Table 2.3 shows the percentage of total calories obtained from complex carbohydrates, for twenty male diabetics who have been on a high-carbohydrate diet, along with their age, weight and percentage of calories as protein.

#	Carbohydrate	Age	Weight	Protein
1	33	33	100	14
2	40	47	92	15
3	37	49	135	18
4	27	35	144	12
5	30	46	140	15
6	43	52	101	15
7	34	62	95	14
8	48	23	101	17
9	30	32	98	15
10	38	42	105	14
11	50	31	108	17
12	51	61	85	19
13	30	63	130	19
14	36	40	127	20
15	41	50	109	15
16	42	64	107	16
17	46	56	117	18
18	24	61	100	13
19	35	48	118	18
20	37	28	102	14

Table 2.3: Example Dataset

We take the Carbohydrate value as our responses Y_i with age, weight and protein as the covariates. Then we fit normal linear model with $Y_i \sim N(\mu_i, \sigma^2)$ where

$$\mu_i = \mathrm{E}(Y_i) = \beta_1 + \beta_2 \mathrm{age}_i + \beta_3 \mathrm{weight}_i + \beta_4 \mathrm{protein}_i \quad (i = 1, \ldots, 20).$$

```
> y <- dat$Carb #response Y is the Carbs.
> X <- cbind(1,dat$Age,dat$Weight,dat$Protein) #design matrix
```

Then, to find the maximum likelihood estimator of $\boldsymbol{\beta}$ we need to solve $X^T X\widehat{\boldsymbol{\beta}} = X^T \boldsymbol{y}$:

```
> beta.hat <- solve(t(X)%*%X, t(X)%*%y)
> t(beta.hat)
```

```
         [,1]       [,2]       [,3]     [,4]
[1,] 36.96006 -0.1136764 -0.2280174 1.957713
```

The unbiased estimator of the variance, $\text{RSS}/(n-p)$, can be computed and used to compute the standard deviation for each component of $\widehat{\boldsymbol{\beta}}$.

```
> n <- length(y)
> p <- length(beta.hat)
> sig.sq.hat <- sum((y-X%*%beta.hat)^2)/(n-p)
> sqrt(diag(sig.sq.hat*solve(t(X)%*%X)))
```

```
[1] 13.07128293  0.10932548  0.08328895  0.63489286
```

The residuals are

```
> ehat <- y-X%*%beta.hat
> summary(ehat)
```

```
       V1
 Min.   :-10.3424
 1st Qu.: -4.8203
 Median :  0.9897
 Mean   :  0.0000
 3rd Qu.:  3.8553
 Max.   :  7.9087
```

The R^2 and it the adjusted version $\overline{R}^2$ coefficients are

```
> RSS <- t(ehat)%*%ehat
> RSS0 <- sum((y-mean(y))^2)
> R2 <- 1-(RSS/RSS0)
> R2
```

```
          [,1]
[1,] 0.4805428
```

```
> 1-(RSS/(20-4))/(RSS0/(20-1))
```

```
          [,1]
[1,] 0.3831445
```

We can check these results using the `lm` function in `R` as follows:

```
> mylm1 <- lm(Carbohydrate~Age+Weight+Protein,data=dat)
> summary(mylm1)
```

```
Call:
lm(formula = Carbohydrate ~ Age + Weight + Protein, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
Age         -0.11368    0.10933  -1.040  0.31389
Weight      -0.22802    0.08329  -2.738  0.01460 *
Protein      1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.956 on 16 degrees of freedom
Multiple R-squared:  0.4805,        Adjusted R-squared:  0.3831
F-statistic: 4.934 on 3 and 16 DF,  p-value: 0.01297
```

To compare two models using the F-test (Theorem 4), we can use the `anova` command (not to be confused on the ANOVA test)

```
> mylm2 <- lm(Carbohydrate~Age,data=dat)
> anova(mylm2,mylm1)
```

```
Analysis of Variance Table

Model 1: Carbohydrate ~ Age
Model 2: Carbohydrate ~ Age + Weight + Protein
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     18 1088.98
2     16  567.66  2    521.32 7.3469 0.005452 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Double-check the F-statistic agrees

```
> RSS0 <- sum(mylm2$residuals^2)
> F <- ((RSS0-RSS)*16)/(RSS*2)
> F
```

```
         [,1]
[1,] 7.346886
```

and also check the corresponding p-values

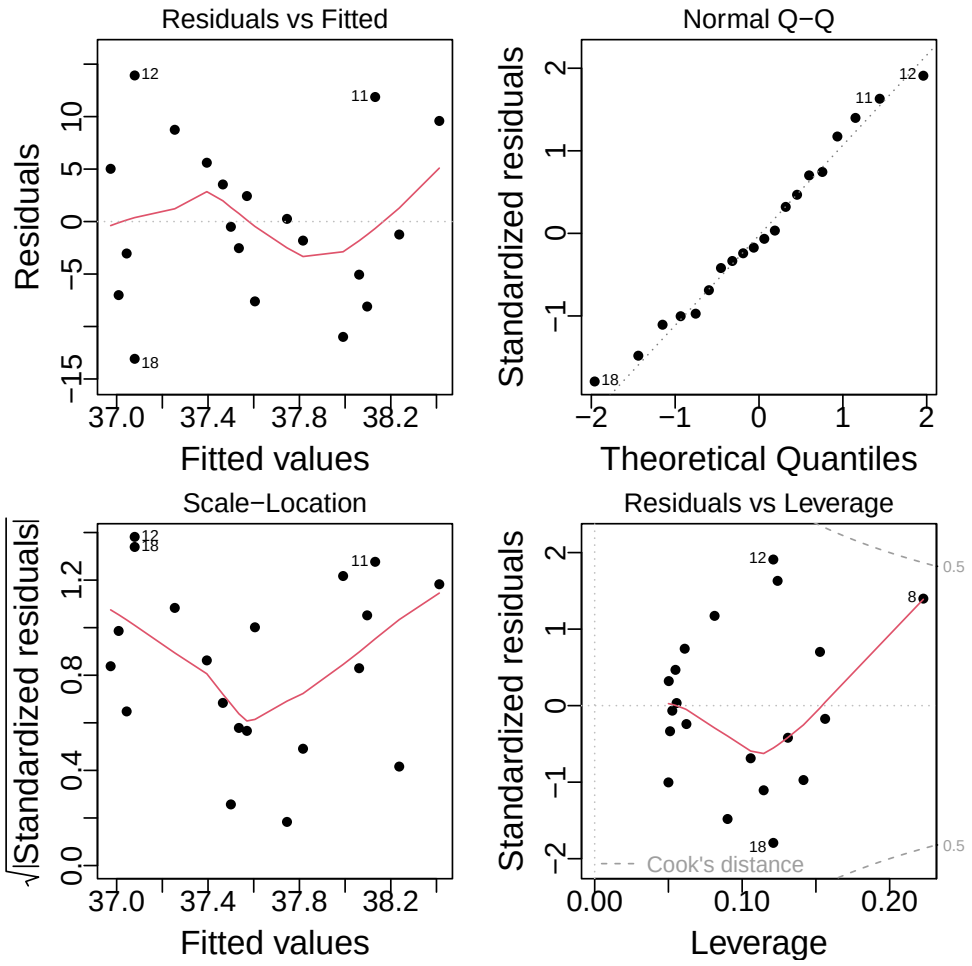
```
> pf(F,df1=2,df2=16,lower.tail=FALSE)
```

```
            [,1]
[1,] 0.005452024
```

Conclusion: Since the corresponding p-values is less than 0.01 we make the following conclusion: There is **sufficient** evidence to reject the null hypothesis at the 1% level (corresponding to the smaller model), and therefore accept the alternative. This means we should proceed to use the larger model.

Lastly, R can also produce some residual plots for us as follows:

```
> plot(mylm2)
```



2.7 Transformations

The way that the data are presented may not necessarily reflect the way we should use the data in our linear models. Recall the prestige dataset — I used the logarithm of income, rather than income directly. Learning to use transformations effectively is part of data analysis. Fortunately, computational tools exists to help us transform the data.

Transformations of the response and predictors can improve the fit and correct violations of model assumptions such as non-constant error variance.

2.7.1 Transforming the Response

Suppose that you are contemplating a logged response in a simple regression situation:

$$\log y = \beta_0 + \beta_1 x + \epsilon.$$

On the original scale of the response, this model becomes:

$$y = \exp(\beta_0 + \beta_1 x) \exp(\epsilon)$$

In this model, the errors enter **multiplicatively** and **not additively**. Therefore, for linear models the logged response model requires that we believe the errors enter multiplicatively on the original scale.

After transforming the response, any interpretations will need to be expressed on the original scale. For example, in the logged model above, your prediction would be $\exp(\widehat{y}_\star)$. If your prediction confidence interval in the logged scale was $[l, u]$, then you would use $[\exp l, \exp u]$. When you use a log transformation on the response, the regression coefficients have a particular interpretation:

$$\log \widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$$
$$\widehat{y} = e^{\widehat{\beta}_0} e^{\widehat{\beta}_1 x_1} \ldots e^{\widehat{\beta}_p x_p}$$

An increase of one unit in x_1 would multiply the predicted response (in the original scale) by $\exp(\widehat{\beta}_1)$.

Box-Cox Transformation

boxcox.R

The Box-Cox method is a popular way to determine a transformation on the response (and also the predictors). The idea is to work with the transformation:

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

For fixed $y > 0$, $y^{(\lambda)}$ is continuous in λ. For $y \leq 0$ the transformation is undefined. The idea is to choose λ that maximises the likelihood.

By a change of variables, we obtain the distribution of y; the resulting log-likelihood being:

$$\ell(\lambda, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}^{(\lambda)} - X\boldsymbol{\beta})^T(\boldsymbol{y}^{(\lambda)} - X\boldsymbol{\beta}) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i).$$

Substituting the MLEs of $\boldsymbol{\beta}$ and σ^2, for a fixed λ, we obtain the (profile) log-likelihood:

$$\ell(\lambda) = c - \frac{n}{2}\log(\text{RSS}_\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i) \tag{2.6}$$

where c is a constant not involving λ and RSS_λ is the residual sum of squares using $y^{(\lambda)}$ as the response.

The procedure is to evaluate the profile log-likelihood (2.6) for a range of possible values of λ. Rather than selecting the maximum, one often rounds to a value such as $-1, 0, 1/2, 1$ or 2, particularly if the profile log-likelihood is relatively flat around the maximum.

Formally, let $\widehat{\lambda}$ denote the value that maximises the profile log-likelihood. We can test the hypothesis $H_0 : \lambda = \lambda_0$ for any fixed value λ_0 by calculating the likelihood ratio

$$2(\ell(\widehat{\lambda}) - \ell(\lambda_0))$$

which is approximating χ^2_1. This can also be used to construct confidence intervals.

Fortunately, we can maximise the log-likelihood and construct confidence intervals numerically in R. We will need the boxcox function from the MASS package:

```
> library("MASS")
```

We shall demonstrate the boxcox function on the savings dataset

```
> library("faraway")
> data(savings)
> mylm <- lm(sr~pop15+pop75+dpi+ddpi,data=savings)
> boxcox(mylm,plotit=TRUE)
> boxcox(mylm,plotit=TRUE,lambda=seq(0.5,1.5,by=0.1))
```

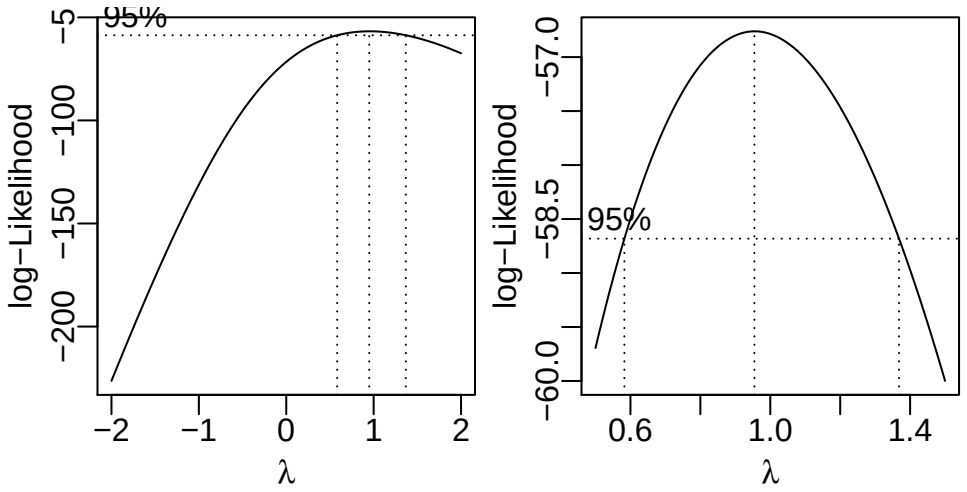


Figure 2.4: Output plots from `boxcox` command: Log-likelihood for Box-Cox transformations.

Not much can be seen from the first plot, so the range of λ is narrowed (second plot). These plots also include the confidence interval for λ. The confidence interval for λ is roughly $(0.6, 1.4)$. We can see that there is no good reason to transform.

Some general considerations concerning the Box-Cox method are

1. The Box-Cox method is affected by outliers.
2. If $\max_i y_i / \min_i y_i$ is small, then the Box-Cox will not have much effect.
3. There is some doubt whether the estimation of λ, counts as an extra parameter to be considered in the degrees of freedom.

Notes

- The Box-Cox method is not the only way of transforming — e.g. Yeo-Johnson transformations.
- You can take a Box-Cox style approach for each of the **predictors**, choosing the transformation to minimize the RSS.

2.8 Contrasts for Categorical Variables

We have encountered different types of variables so far — continuous and categorical. The way we handle categorical variables within a model needs some careful attention which we now discuss. As a running example in this section, let's return to the Prestige data set. Suppose we want to fit the model:

$$\text{[drop subscript i for now]} \quad \text{prestige} = \beta_1 + \beta_2 \text{education} + \beta_3 \text{type} + \epsilon \quad \epsilon \sim N(0, \sigma^2).$$

The variable education is assumed to be continuous, whereas the occupation type is categorical taking the the levels: blue collar (bc), professional (prof), and white collar (wc).

Let's first discuss the wrong way to include the occupation type into the linear model.

[START OF THE WRONG WAY]

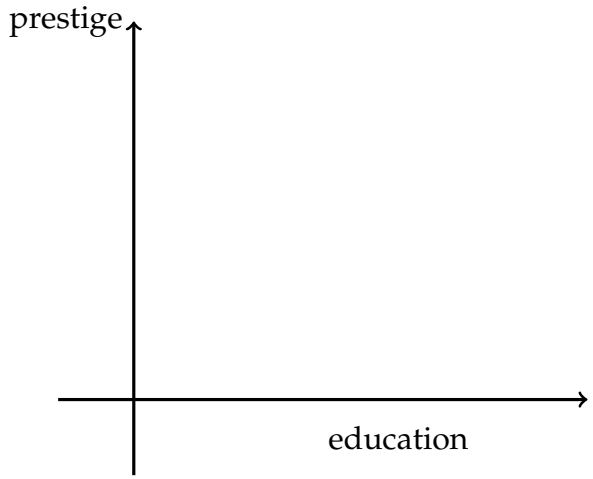
Suppose we enter the occupation type into the linear model by encoding occupation type numerically as shown below.

Label (used in dataset)	Encoding
blue collar	1
professional	2
white collar	3

In this model we assume that all three occupations exhibit the same relationship between education and prestige but that they start from different baselines. Using the encoding scheme described in the table above yields the following equations for the occupations.

$$\begin{aligned}
\text{blue collar: } & \text{prestige} = \beta_1 + \beta_2 \text{education} + \beta_3 \\
\text{professional: } & \text{prestige} = \beta_1 + \beta_2 \text{education} + 2\beta_3 \\
\text{white collar: } & \text{prestige} = \beta_1 + \beta_2 \text{education} + 3\beta_3
\end{aligned}$$

This translates into the following illustrative diagram:



The numeric coding scheme we have used for occupation has imposed an unwanted assumption/constraint on the location of the intercepts:

- While the location of the intercepts of blue collar and professional workers are arbitrary, the location of white collar workers is not. The numeric coding scheme has constrained, for a fixed level of education, the difference in prestige between blue collar and professional workers to be exactly the same as the difference between white collar and professional workers, β_3.
- Furthermore the prestige difference between white collar and blue collar workers, for fixed education level, is twice the difference between either one and professional workers.

These assumptions may or may not be correct — in general we would prefer to test this instead of assuming them through the postulated model. In order to correct this problem, we require a more general encoding of the categories.

Note

Just changing the number assignments, i.e. instead of using 1, 2 and 3 use other numbers, will lead to the same problem – just with a different set of constraints on the intercepts. Therefore, we require a way of encoding that does not enforce any constraints on the estimates of the model.

[END OF THE WRONG WAY]

The standard approach when handling a categorical variable with n distinct levels is to create $n - 1$ different regressors. These are often called dummy regressors.

If the Prestige data set is read into `R` using the `read.table` function, the character variable occupational type is automatically converted to a variable of class type

"factor". When a factor is used in a linear model it automatically employs the correct number of dummy regressors. The coding used is termed a contrast and in R the default contrast type for a factor is "treatment" and denoted `contr.treatment`. This produces the two dummy regressors X_1 and X_2 shown below.

```
      X1 X2
bc     0  0
prof   1  0
wc     0  1
```

Note

By default R orders the levels of the character variable and defines the dummy coding levels in alphabetical order. As a result the first level alphabetically becomes the baseline level.

If we fit a normal linear model in R with prestige as the response and education and occupation type as predictors by `lm(prestige~education+type,data=Prestige)` the following model is fitted:

$$\text{prestige} = \beta_1 + \beta_2 \text{education} + \beta_3 X_1 + \beta_4 X_2 + \epsilon.$$

- In the case $X_1 = 0$ and $X_2 = 0$ for blue collar workers, the model is

$$\text{prestige} = \beta_1 + \beta_2 \text{education} + \epsilon.$$

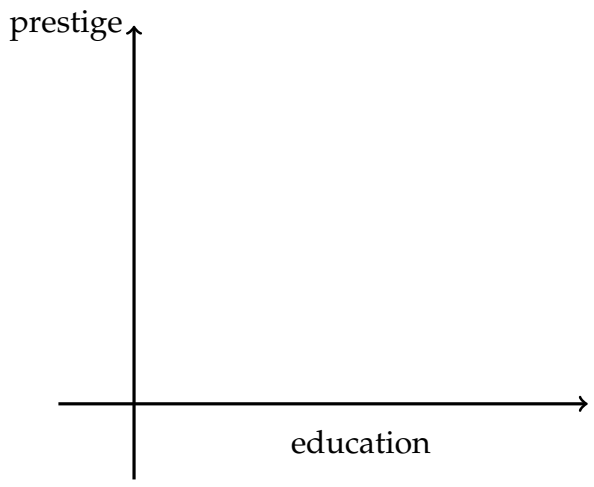
- In the case $X_1 = 1$ and $X_2 = 0$ for professional workers, the model is

$$\text{prestige} = (\beta_1 + \beta_3) + \beta_2 \text{education} + \epsilon.$$

- In the case $X_1 = 0$ and $X_2 = 1$ for white collar workers, the model is

$$\text{prestige} = (\beta_1 + \beta_4) + \beta_2 \text{education} + \epsilon.$$

These fitted models are illustrated in the following diagram:



Since β_1, β_3, and β_4 are estimated separately it is clear from the diagram above there are no constraints placed on the location of the intercepts.

It is also clear why we only need $n-1$ regressors to independently describe n levels of a categorical variable. In the additive model, the categorical variable levels just serve to change the intercept of the regression line. Since there is already an intercept in the model, it is used for one of levels of the categorical variable.

2.8.1 Other Codings

There are many possible coding schemes for categorical variables other than the `treatment` contrast. To simplify matters we shall consider a model in which the only predictor is the variable occupation. Thus in `R` notation we fit the model: `lm(prestige~type,data=Prestige)`. For the following coding schemes, we represent the occupation type using two regressors.

Dummy Coding

As explained above with dummy coding the occupation variable is converted into dummy variables X_1 and X_2 whose values for the various levels of occupation were listed above. By default the level that is first alphabetically becomes the baseline level. This is the `contr.treatment` coding scheme of `R`. If we specify the regression model `lm(prestige~type,data=Prestige),R` fits the model:

$$\text{prestige} = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \epsilon$$

The model equation can be used to predict the mean as a function of X_1 and X_2. When we choose values for X_1 and X_2 that correspond to the various occupations, we obtain the means of those professions. More precisely, with the dummy coding scheme we obtain the following equations for the means:

$$\begin{aligned}
\mu_{\text{bc}} &= \beta_1 \\
\mu_{\text{prof}} &= \beta_1 + \beta_2 \\
\mu_{\text{wc}} &= \beta_1 + \beta_3
\end{aligned}$$

By subtracting the mean for blue collar workers from each of the other equations we obtain expressions for β_2 and β_3.

$$\begin{aligned}
\mu_{\text{bc}} &= \beta_1 \\
\mu_{\text{prof}} - \mu_{\text{bc}} &= \beta_2 \\
\mu_{\text{wc}} - \mu_{\text{bc}} &= \beta_3
\end{aligned}$$

Thus in the dummy coding scheme each coefficient measures a difference in (conditional) mean between one classification level and the classification level that was

chosen as baseline. The intercept corresponds to the mean of the baseline classification level.

Deviation Coding

In R deviation coding is denoted `contr.sum`. To assign this contrast to the variable occupation type we would use the following statement:

```
contrasts(Prestige$type) <- "contr.sum"
```

The deviation coding scheme used in R yields the two dummy regressors shown below:

```
     X1 X2
bc    1  0
prof  0  1
wc   -1 -1
```

If we run `lm(prestige~type)`, R fits the model:

$$\text{prestige} = \beta_1 + \beta_2 X_1 + \beta_3 X_2$$

As before, the regression model estimates the conditional mean of y, conditional on the values that are specified for the predictors. Thus when we choose values for X_1 and X_2 that correspond to the various professions, we obtain the means of those professions. The deviation coding scheme yields the following equations for the means.

$$\begin{aligned}\mu_{\text{bc}} &= \beta_1 + \beta_2 \\ \mu_{\text{prof}} &= \beta_1 + \beta_3 \\ \mu_{\text{wc}} &= \beta_1 - \beta_2 - \beta_3\end{aligned}$$

Further, summing these equations give

$$\mu_{\text{bc}} + \mu_{\text{prof}} + \mu_{\text{wc}} = 3\beta_1 \therefore \beta_1 = \frac{\mu_{\text{bc}} + \mu_{\text{prof}} + \mu_{\text{wc}}}{3}$$

Thus the intercept in deviation coding corresponds to the mean of all three levels.

From the equations for blue collar and professional and using the formula for β_1 above we immediately obtain interpretations for β_2 and β_3.
$\beta_2 =$ the difference between the mean for blue collar and the overall mean
$\beta_3 =$ the difference between the mean for professional and the overall mean

Thus in deviation coding the coefficients measure the distance between individual levels and the mean of all the levels.

Helmert Coding

In R Helmert coding is denoted `contr.helmert`. Assigning this contrast to the variable occupation type is done with the following statement.

```
contrasts(Prestige$type) <- "contr.helmert"
```

The Helmert coding scheme yields the two dummy regressors shown below

```
     X1 X2
bc   -1 -1
prof  1 -1
wc    0  2
```

As before when we specify the regression model `lm(prestige~type)`, R actually fits the model: prestige $= \beta_1 + \beta_2 X_1 + \beta_3 X_2$. The Helmert coding scheme yields the following equations for the means.

$$\mu_{\text{bc}} = \beta_1 - \beta_2 - \beta_3$$
$$\mu_{\text{prof}} = \beta_1 + \beta_2 - \beta_3$$
$$\mu_{\text{wc}} = \beta_1 + 2\beta_3$$

As with deviation coding, if we add all three equations together we isolate β_1 and find that the intercept represents the mean of all the levels.

$$\beta_1 = \frac{\mu_{\text{bc}} + \mu_{\text{prof}} + \mu_{\text{wc}}}{3}$$

We can isolate β_2 by subtracting the blue collar mean from the professional mean.

$$\mu_{\text{prof}} - \mu_{\text{bc}} = 2\beta_2 \therefore \beta_2 = \frac{\mu_{\text{prof}} - \mu_{\text{bc}}}{2}$$

The best way to interpret this coefficient is to observe that if β_2 is not significantly different from zero we would conclude that the mean prestige for professionals and the mean prestige for blue collar workers are not significantly different from each other.

We can isolate β_3 by taking the average of the blue collar and professional means and then subtracting the result from the white collar mean.

$$\mu_{\text{wc}} - \frac{\mu_{\text{bc}} + \mu_{\text{prof}}}{2} \therefore \beta_3 = \frac{1}{3}\left(\mu_{\text{wc}} - \frac{\mu_{\text{bc}} + \mu_{\text{prof}}}{2}\right)$$

The best way to interpret this coefficient is to observe that if β_3 is not significantly different from zero we would conclude that the mean prestige of white collar workers is not significantly different from the mean prestige of blue collar and professional workers together.

Thus Helmert coding compares the current level with the average of the all the levels that preceded it. Thus Helmert contrast coding is especially appropriate if there is a natural order to the categories because then sequential comparisons of this sort make sense.

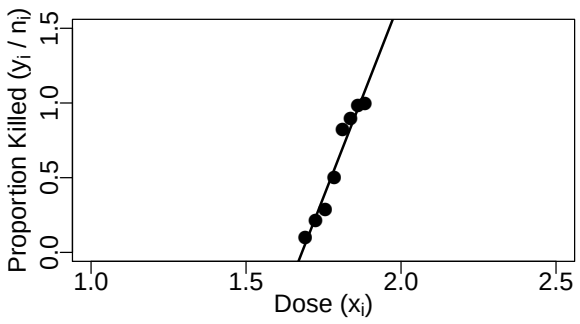
Chapter 3. Generalized Linear Models

Motivating Example

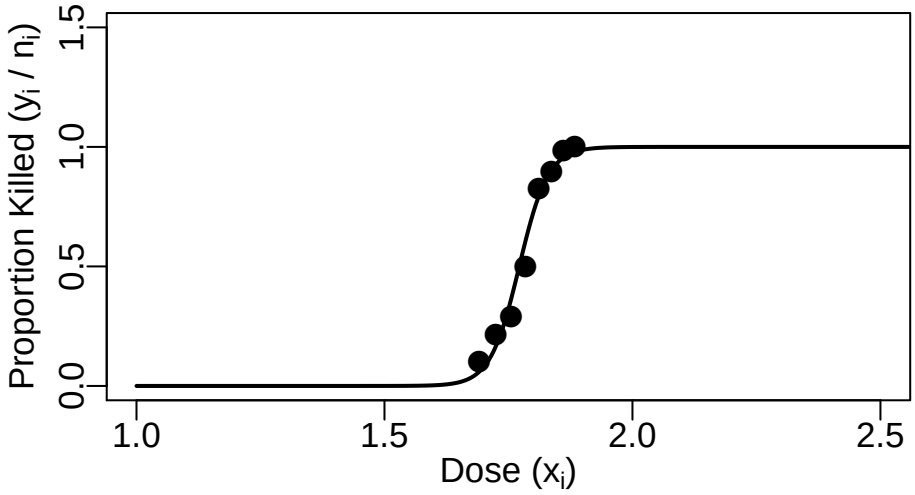
`beetle.R beetle-data.RData`

Let us consider the following dataset regarding numbers of beetles that died after five hours of exposure to gaseous carbon disulphide at various dose levels. Can the dosage be used to measure the proportion of deaths?

Dose x_i	Number n_i	Number Killed y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.861	62	61
1.8839	60	60



Here the response of interest is the proportion of deaths, which must lie between 0 and 1. Thus, a linear model will not explain the data adequately. Therefore, we need a different class of model - in this case, one where the fitted values remain between 0 and 1. In this chapter, we will discuss the class of generalized linear models (GLMs), that is able to give a suitable fit. For this example, a model fitted using a generalized linear model is presented below.



3.1 Specification of Generalized Linear Models

3.1.1 The components of GLMs

NORMAL LINEAR MODELS

1. **Random Component**
   The components of $\boldsymbol{Y}$ have independent normal distributions with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ and individual variance σ^2.

2. **Systematic Component**
   Using the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ form the linear predictor

$$\boldsymbol{\eta} = X\boldsymbol{\beta} = \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j$$

3. **Link**
   The link between the random and systematic components is

$$\boldsymbol{\eta} = \boldsymbol{\mu}$$

GENERALIZED LINEAR MODELS

1. **Random Component**
   The components of $\boldsymbol{Y}$ are independent and have the same distribution. This distribution is a member of an exponential family with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$.

2. **Systematic Component**
   Using the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ form the linear predictor

$$\boldsymbol{\eta} = X\boldsymbol{\beta} = \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j$$

3. **Link**
   The link between the random and systematic components is

$$\eta_i = g(\mu_i) \quad \text{for} \quad i = 1, \ldots, n.$$

In GLMs

- The **random component** specifies the probability distribution of the response variables. Specifically, the components of $\boldsymbol{y}$ have pdf or pmf from an exponential family of distributions (see section 3.1.2), with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$.
- The **systematic component** specifies a linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ as a function of the covariates and the unknown parameters.
- The **link function** g may be any monotonic differentiable function. The link function provides a functional relationship between the systematic component and the expectation of the response in the random component; namely $\boldsymbol{\eta} = g(\boldsymbol{\mu})$

3.1.2 Exponential Families

Definition. Consider a random variable Y whose probability density/mass function depends on the two parameters θ and ϕ. The distribution of Y is said to be a member of an exponential family if its probability density function can be written in the form

$$\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.1}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. If ϕ is known, this is an exponential family model with canonical parameter θ.

The function $a(\phi)$ commonly takes the form

$$a(\phi) = \phi/w,$$

where ϕ is called the dispersion parameter that is constant over all observations and w is a known prior weight which can vary across the observations. However, the practical cases of interest dealt with in this course will use the form $a(\phi) = \phi/w$, where in most cases $w = 1$.

Example Consider $Y \sim N(\mu, \sigma^2)$. The density function for Y can be written in exponential family form

■

Example Consider $Y \sim \text{Poisson}(\lambda)$. The mass function for Y can be written in exponential family form

■

Example Consider $Y \sim$ Inverse Gaussian(μ, λ). Its probability density function can be written in exponential family form as follows

$$\begin{aligned}\sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right\} &= \exp\left\{\frac{-\lambda(y-\mu)^2}{2\mu^2 y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right\} \\ &= \exp\left\{\frac{y(-1/2\mu^2) + (1/\mu)}{1/\lambda} - \frac{\lambda}{2y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right\}\end{aligned}$$

where we identify $\theta = (-1/2\mu^2)$, $\phi = 1/\lambda$ and

$$a(\phi) = \phi, \quad b(\theta) = -(-2\theta)^{1/2}, \quad c(y, \phi) = -\frac{1}{2}\left(\log(2\pi\phi y^3) + \frac{1}{\phi y}\right).$$

■

Note that if ϕ is unknown, it is called a nuisance parameter of the distribution as it will interfere with inference of μ or θ. We need to be wary about this when performing hypothesis tests or constructing confidence intervals (see later in section 3.3).

Properties of exponential family distributions

The mean and variance of an exponential family distribution are easily derived as follows. From the definition of any probability density function

$$\int f(y; \theta, \phi) dy = 1.$$

Under regularity conditions, we can differentiate with respect to θ to obtain

$$\frac{\partial}{\partial\theta}\int f(y; \theta, \phi) dy = \int \frac{\partial}{\partial\theta} f(y; \theta, \phi) dy = \frac{\partial}{\partial\theta} 1 = 0. \tag{3.2}$$

If we differentiate again with respect to θ, we get

$$\int \frac{\partial^2}{\partial\theta^2} f(y; \theta, \phi) dy = 0. \tag{3.3}$$

These general results can be applied to distributions from an exponential family. From (3.1) recall that general form of the pdf from an exponential family is:

so that

$$\frac{\partial}{\partial\theta} f(y; \theta, \phi) =$$

Then by (3.2) we find

by definition of the expected value. Thus

$$\mu \equiv \mathrm{E}(Y) = b'(\theta). \tag{3.4}$$

Next, we have

$$\frac{\partial^2}{\partial \theta^2} f(y;\theta,\phi) = -\frac{b''(\theta)}{a(\phi)} f(y;\theta,\phi) + \left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 f(y;\theta,\phi). \tag{3.5}$$

Plugging this into (3.3) gives

Rearranging yields

$$\mathrm{var}(Y) = b''(\theta)a(\phi).$$

The function $b''(\theta)$ is referred to as the variance function. The variance function considered as a function of μ will be denoted as $V(\mu)$. To summarise

$$\begin{aligned} \mathrm{E}(Y) &\equiv \mu = b'(\theta) \\ \mathrm{var}(Y) &= b''(\theta)a(\phi) \equiv V(\mu)a(\phi) \end{aligned}$$

Warning! In general, the variance function is not the same as the variance of Y.

Example Suppose $Y \sim N(\mu, \sigma^2)$. Recall that the exponential family form for the pdf of Y is defined as $\theta = \mu$, $\phi = \sigma^2$ with $a(\phi) = \phi$ and $b(\theta) = \theta^2/2$. Then $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$ are easily obtained:

■

Example Suppose $Y \sim \text{Poisson}(\lambda)$. Then for any $\lambda > 0$ and $y \in \mathbb{N}$

$$f(y;\lambda) = \exp(y \log \lambda - \lambda - \log(y!)),$$

where we identify $\theta = \log \lambda$, $a(\phi) = 1$ and

$$b(\theta) = \exp(\theta) \quad \text{and} \quad c(y,\phi) = -\log(y!).$$

We can check the mean and variance:

■

Therefore, we are able to compute the mean and variance of any exponential family distribution just by looking at the form of its probability density function and differentiating b.

3.1.3 Link Functions

Recall that in the specification of GLMs, the function g describes the link between the expected response $\boldsymbol{\mu} \equiv \mathrm{E}(\boldsymbol{Y})$ and the linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$. We introduced the link function g as

$$\eta_i = g(\mu_i) \quad \text{for} \quad i = 1, \ldots, n$$

as any monotonic differentiable function. Recall a function, g, is monotonically increasing (or decreasing), if for all z_1 and z_2 with $z_1 \leq z_2$, we have $g(z_1) \leq g(z_2)$ (or $g(z_1) \geq g(z_2)$)

Why link functions are necessary

Example Suppose that $Y_1, \ldots, Y_n$ are independent random variables such that $Y_i \sim \text{Poisson}(\lambda_i)$ for some $\lambda_i > 0$. Then $\mu_i \equiv \mathrm{E}(Y_i) = \lambda_i > 0$. It is plausible that the linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ can take any value on $\mathbb{R}$. Thus using the identity link function, i.e. $g(u) = u$, is not recommended as η_i may be negative while λ_i is always positive. Therefore, using another link function g is required. ■

Canonical Links

For exponential family distributions, there is a natural or canonical choice of link function. These canonical links occur when

$$\theta = \eta$$

where θ is the canonical parameter (see Definition 3.1.2). Recall that in the GLM specification, the link function can be any monotonic differentiable function. Therefore, using the canonical link is not necessary. However, as we shall see later, using the canonical link leads to some desirable results (see later in section 3.2).

Example of GLMs

Example Suppose $Y_1, \ldots, Y_n$ are independent random variables each with a Poisson distribution and

$$\mathrm{E}(Y_i) = \exp(\beta_1 + \beta_2 x_i)$$

Then this is a GLM as the Poisson distribution can be written in exponential family form and the link function $g(z) = \log(z)$ is monotonic and differentiable with linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$. ■

Example Suppose $Y_1, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$ random variables where

$$\mu_i = \beta_1 + \beta_2 x_i$$

Then this is a GLM as the Normal distribution can be written in exponential family form and the link function $g(z) = z$ (the identity) is monotonic and differentiable. ■

The generalization: from normal linear models to GLMs

Notice the two generalizations:

- The distribution of the response can be any member of the exponential family — not just a normal distribution.
- The link function, connecting the mean of the response and the linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$, may be any monotonic differentiable function — not only the identity function.

These generalizations lead to more complicated estimation and inference procedures, as we shall see.

3.2 Estimation

So far we have just introduced the class of models called the GLMs. Now we discuss how to estimate the unknown parameter $\boldsymbol{\beta}$ in a GLM. Recall the estimation procedure for normal linear models; we want to find $\widehat{\boldsymbol{\beta}}$ by maximising the log-likelihood. The log-likelihood for GLMs is

$$\ell(\boldsymbol{\beta}; \phi, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

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

$$\mu_i = b'(\theta_i), \quad g(\mu_i) = \eta_i, \quad \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad \text{for } i = 1, \ldots, n.$$

The approach used to find the estimator of $\boldsymbol{\beta}$ for GLMs is the same we used before; namely to maximise the log-likelihood by solving

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} \stackrel{!}{=} 0.$$

However, for GLMs, this (generally) involves solving a non-linear set of equations. Therefore, we use an iterative method to obtain a numerical solution.

Before we present the maximum likelihood estimation procedure for GLMs, we require a brief detour to introduce the numerical method and some maximum likelihood theory.

3.2.1 Maximum Likelihood Theory

Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ has a pdf or pmf $f(\boldsymbol{y}; \boldsymbol{\psi})$ for $\boldsymbol{y} \in \mathbb{R}^n$ and parameter vector $\boldsymbol{\psi} \in \mathbb{R}^p$. Then likelihood function is a function of $\boldsymbol{\psi}$ for each fixed $\boldsymbol{y}$ given by

$$L(\boldsymbol{\psi}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\psi}).$$

Similarly, the log-likelihood function is given by

$$\ell(\boldsymbol{\psi}; \boldsymbol{y}) = \log L(\boldsymbol{\psi}; \boldsymbol{y}) = \log f(\boldsymbol{y}; \boldsymbol{\psi}).$$

We then have the following:

$$\text{Score Vector:} \quad \boldsymbol{U}(\boldsymbol{\psi};\boldsymbol{y}) = \nabla \ell(\boldsymbol{\psi};\boldsymbol{y}) \quad \text{where } \nabla = \left(\frac{\partial}{\partial \psi_1}, \ldots, \frac{\partial}{\partial \psi_p}\right)$$

$$\text{so that } U_j(\boldsymbol{\psi};\boldsymbol{y}) = \frac{\partial}{\partial \psi_j}\ell(\boldsymbol{\psi};\boldsymbol{y}), \quad j = 1, \ldots, p$$

$$\text{Observed Information Matrix:} \quad \mathcal{I} = -\nabla\nabla^T \ell(\boldsymbol{\psi};\boldsymbol{y})$$

$$\text{so that } \mathcal{I}_{jk}(\boldsymbol{\psi};\boldsymbol{y}) = -\frac{\partial^2}{\partial \psi_j \partial \psi_k}\ell(\boldsymbol{\psi};\boldsymbol{y})$$

$$\text{Fisher's Information Matrix:} \quad \mathcal{J}(\boldsymbol{\psi}) = \mathrm{E}(\mathcal{I}(\boldsymbol{\psi};\boldsymbol{Y}))$$

If $Y_1, \ldots, Y_n$ are iid with pdf or pmf $f(y_i;\boldsymbol{\psi})$, $y \in \mathbb{R}$ then

$$L(\boldsymbol{\psi}; y_1, \ldots, y_n) = \prod_{i=1}^{n} L(\boldsymbol{\psi}, y_i) \quad \text{and} \quad \ell(\boldsymbol{\psi}; y_1, \ldots, y_n) = \sum_{i=1}^{n} \ell(\boldsymbol{\psi}; y_i)$$

Example Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Exponential}(1/\mu)$ for some $\mu > 0$. We have

$$f(y_i;\mu) = \frac{1}{\mu}\exp\left(-\frac{y_i}{\mu}\right) \quad y_i \geq 0$$

$$\ell(\mu; y_1, \ldots y_n) =$$

$$U = \frac{\partial}{\partial \mu}\ell(\mu; y) =$$

$$\mathcal{I} = -\frac{\partial^2}{\partial \mu^2}\ell(\mu; y_1, \ldots, y_n) =$$

$$\mathcal{J} =$$

■

Example Suppose $Y_1, \ldots, Y_n \sim N(\mu_i, \sigma^2)$ independently with $\sigma^2 > 0$ known and some for $\mu_i \in \mathbb{R}$. We have

$$f(y_i; \mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$$

$$f(y_1, \ldots, y_1; \boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$$

$$\ell(\boldsymbol{\mu}; y_1, \ldots, y_n) =$$

$$U_j = \frac{\partial}{\partial \mu_j} \ell(\boldsymbol{\mu}; \boldsymbol{y}) =$$

$$\mathcal{I}_{jk} =$$

$$\mathcal{J}_{jk} =$$

■

Theorem 5. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ have a pdf or pmf $f(\boldsymbol{y}; \boldsymbol{\psi})$ for $\boldsymbol{y} \in \mathbb{R}^n$ and parameter vector $\boldsymbol{\psi} \in \mathbb{R}^p$. Then, under certain regularity conditions,

$$\mathrm{E}(\boldsymbol{U}) = \boldsymbol{0} \quad \text{and} \quad \mathrm{cov}(\boldsymbol{U}) = \mathrm{E}(\boldsymbol{U}\boldsymbol{U}^T) = \mathcal{J}.$$

Proof. For $j = 1, \ldots, p$ we have

$$\mathrm{E}(U_j) =$$

It then follows that $\mathrm{cov}(\boldsymbol{U}) = \mathrm{E}(\boldsymbol{U}\boldsymbol{U}^T)$ as $\mathrm{E}(\boldsymbol{U}) = \boldsymbol{0}$. Thus we only need to show that $\mathrm{E}(\boldsymbol{U}\boldsymbol{U}^T) = \mathcal{J} = \mathrm{E}(\mathcal{I})$. See problem sheet for remainder of proof □

We now return to exponential families. We shall write $x_{ij} = (X)_{ij}$; also recall that the parameter of interest is $\boldsymbol{\beta} \in \mathbb{R}^p$.

Theorem 6. Suppose we have $Y_1, \ldots, Y_n$ independent random variables from an exponential family with means $\mu_1, \ldots, \mu_n$ and variance functions $V(\mu_1), \ldots, V(\mu_n)$ with common dispersion parameter ϕ. Then for $j = 1, \ldots, p$ and $k = 1, \ldots, p$,

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right] \quad \text{and} \quad \mathcal{J}_{jk} = \sum_{i=1}^{n} \left[\frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right].$$

Proof. The likelihood function for $Y_1, \ldots, Y_n$ is

$$L(\boldsymbol{\beta}; y_1, \ldots, y_n) =$$

where $\theta(\mu_i) = \theta(g^{-1}(\eta_i))$, $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. Therefore, the loglikelihood is

$$\ell(\boldsymbol{\beta}; y) =$$

For score function

$$U_j = \frac{\partial \ell(\boldsymbol{\beta}; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} \quad \text{for } j = 1, \ldots, p,$$

we decompose the partial derivative as

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \quad \text{where} \quad \theta_i \equiv \theta(\mu_i).$$

Now consider each partial derivative in turn. First,

$$\frac{\partial \ell_i}{\partial \theta_i} =$$

Second,

$$\frac{\partial \theta_i}{\partial \mu_i} =$$

Third,

$$\frac{\partial \mu_i}{\partial \beta_j} =$$

Inserting these into the score U_j yields

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right], \tag{3.6}$$

as required.

For the Fisher information matrix, $\mathcal{J} = \mathrm{E}(\boldsymbol{U}\boldsymbol{U}^T)$, we use (3.6) to get

$$\mathcal{J}_{jk} =$$

where we used that the observations are independent. □

These results can be succintly be written as

$$\boldsymbol{U} = X^T W \boldsymbol{c}, \tag{3.7}$$

where the components of $\boldsymbol{c}$ are given by

$$c_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i},$$

and the Fisher information matrix can be represented as

$$\mathcal{J} = X^T W X, \tag{3.8}$$

with $W \in \mathbb{R}^{n \times n}$ being a diagonal matrix with entries

$$w_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Recall that to find the MLE, $\widehat{\boldsymbol{\beta}}$, the procedure is to differentiate the log-likelihood with respect to $\boldsymbol{\beta}$ and solve equal to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} \stackrel{!}{=} \mathbf{0}.$$

This is equivalent to solving $U_j = 0$ for $j = 1, \ldots, p$. Therefore, the maximum likelihood equations are

$$\sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) \right] = 0, \quad \text{for } j = 1, \ldots, p. \tag{3.9}$$

In general, this is a non-linear system of equations in $\boldsymbol{\beta}$ and a numerical solution is required. We proceed to solve (3.9) using a variant of the Newton-Raphson algorithm.

Algorithm. ***Newton-Raphson:*** *For* $f(\boldsymbol{x}) : \mathcal{X} \subseteq \mathbb{R}^p \to \mathbb{R}; \boldsymbol{x} \mapsto f(\boldsymbol{x})$*, the algorithm below:*

$$\boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(m)} - H^{-1}\mathbf{g}, \quad \textit{where} \quad H = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^{(m)}}, \ \mathbf{g} = \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}^{(m)}}$$

will converge to $\boldsymbol{x}^\star = \underset{\boldsymbol{x}}{\textit{argmax}}\, f(\boldsymbol{x})$ *(under regularity conditions).*

Using the Newton-Rapshon algorithm for $\ell(\boldsymbol{\beta}; \boldsymbol{y})$ yields:

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} + \left(\mathcal{I}^{(m)}\right)^{-1} \boldsymbol{U}^{(m)},$$

where $\boldsymbol{U}^{(m)}$ is the score vector evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(m)}$, and similarly for the observed information matrix $\mathcal{I}^{(m)}$. In maximum likelihood estimation it is common to estimate $\mathcal{I}$ by $\mathrm{E}(\mathcal{I}) = \mathcal{J}$ i.e. Fishers information matrix*. The result is the Fisher scoring algorithm:

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} + \left(\mathcal{J}^{(m)}\right)^{-1} \boldsymbol{U}^{(m)}. \tag{3.10}$$

Continuing to find simpler representations for the iterations of the scoring algorithm:

Plugging in (3.7) and (3.8) into (3.10) gives

$\widehat{\boldsymbol{\beta}}^{(m+1)} =$

(3.11)

These iterated weighted least squares (IWLS) equations (3.11) may be viewed as a weighted least squares equations using the linear model

$$\boldsymbol{z} = X\boldsymbol{\beta} + \boldsymbol{c}, \quad \text{where } \boldsymbol{c} = (\boldsymbol{y} - \boldsymbol{\mu})(\partial\eta/\partial\mu) \sim N(\mathbf{0}, W^{-1}). \tag{3.12}$$

To see this, start by rewriting (3.12) as

$$\boldsymbol{z} = X\boldsymbol{\beta} + W^{-1/2}\boldsymbol{\epsilon}.$$

Then multiplying on the left by $W^{1/2}$

$$\underbrace{W^{1/2}\boldsymbol{z}}_{\widetilde{\boldsymbol{y}}} = \underbrace{W^{1/2}X}_{\widetilde{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, I_n),$$

*see problem sheet

which is of the form of a normal linear model. We can write the estimate down directly

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{y} \\ &= (X^TWX)^{-1}X^TW\boldsymbol{z},\end{aligned}$$

which is similar to (3.11). This motivates the iterative weighted least squares (IWLS) algorithm.

Algorithm 3.1 Iterative weighted least squares (IWLS) algorithm

1: Given a current estimate $\widehat{\boldsymbol{\beta}}$, form the linear predictor $\widehat{\boldsymbol{\eta}}$ and the fitted values $\widehat{\boldsymbol{\mu}}$.
2: Form the adjusted dependent variable

$$z_i = \widehat{\eta}_i + (y_i - \widehat{\mu}_i)\frac{\partial \eta}{\partial \mu}\bigg|_{\mu=\widehat{\mu}_i}$$

3: Form the estimated weights $\widetilde{w}_{ii}$ by

$$\widetilde{w}_{ii}^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 V(\mu)\bigg|_{\mu=\widehat{\mu}_i}$$

4: Regress z_i on $\boldsymbol{x}_i$ with weights $\widetilde{w}_{ii}$ and obtain the new estimate $\widehat{\boldsymbol{\beta}}$.
5: Repeat steps 1 to 4 until convergence.

Note the relationship between the two weights: $\widetilde{w}_{ii} = \phi w_{ii}$ ¶(see below)

Lastly, we have that

$$\begin{aligned}\operatorname{cov}(\widehat{\boldsymbol{\beta}}) &= \operatorname{cov}\left\{\left(X^TWX\right)^{-1}X^TW\boldsymbol{z}\right\} \\ &= \left[\left(X^TWX\right)^{-1}X^TW\right]\underbrace{\operatorname{cov}(\boldsymbol{z})}_{W^{-1}}\left[\left(X^TWX\right)^{-1}X^TW\right]^T \\ &= \left(X^TWX\right)^{-1} = \phi(X^T\widetilde{W}X)^{-1} \\ &= \mathcal{J}^{-1}. \end{aligned} \tag{3.13}$$

Notice, that $\mathcal{J}$ involves the dispersion parameter ϕ — this needs to be known in order to compute $\mathcal{J}$.

We estimate $\operatorname{cov}(\widehat{\boldsymbol{\beta}})$ by "plugging" in the last $W^{(m)}$ from the IWLS algorithm to get

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) \approx (X^TW^{(m)}X)^{-1}$$

To be clear[¶] why we are using two weights: Recall that we want to solve the maximum likelihood equations :

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] = 0, \quad \text{for } j = 1, \ldots, p. \qquad \text{(3.9 revisited)}$$

Since we have that $\text{var}(Y_i) = a(\phi)V(\mu_i) = \phi V(\mu_i)$, solving (3.9) is equivalent to solving

$$\sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{V(\mu_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] = 0, \quad \text{for } j = 1, \ldots, p.$$

Therefore, within the IWLS algorithm (Algorithm 3.1) we use

$$\widetilde{w}_{ii} = \frac{1}{V(\mu)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \Big|_{\mu = \widehat{\mu}_i}$$

rather than

$$w_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \Big|_{\mu = \widehat{\mu}_i} = \frac{1}{\phi V(\mu)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \Big|_{\mu = \widehat{\mu}_i}$$

Example calculations

Example For independent $Y_1, \ldots, Y_n$ with $Y_i \sim \text{Binomial}(n_i, \pi_i)$, for known n_i,

$$f(y_i; \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i},$$

Then, working with the canonical link:

$\mu_i \equiv \mathrm{E}(Y_i) =$

$\theta_i =$

$b(\theta_i) =$

$b'(\theta_i) =$

$b''(\theta_i) =$

$V(\mu_i) =$

$\eta_i =$

$\frac{\partial \eta_i}{\partial \mu_i} =$

$z_i = \widehat{\eta}_i +$

$w_{ii}^{-1} =$

■

Example For independent $Y_1, \ldots, Y_n$ with $Y_i \sim \text{Normal}(\zeta_i, \sigma^2)$ for known $\sigma^2 > 0$,

$$f(y_i; \lambda_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \zeta_i)^2\right\}.$$

Then, working with the canonical link:

$\mu_i \equiv \text{E}(Y_i) = \zeta_i$

$\theta_i =$

$a(\phi) =$

$b(\theta_i) =$

$b'(\theta_i) =$

$b''(\theta_i) =$

$V(\mu_i) =$

$\eta_i =$

$\frac{\partial \eta}{\partial \mu} =$

$z_i = \widehat{\eta}_i +$

$w_{ii}^{-1} =$

■

Example For independent $Y_1, \ldots, Y_n$ where $Y_i \sim \text{Poisson}(\lambda_i)$

$$f(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}.$$

Then, working with the identity link:

$\mu_i \equiv E(Y_i) =$

$\theta_i =$

$a(\phi) =$

$b(\theta_i) =$

$b'(\theta_i) =$

$b''(\theta_i) =$

$V(\mu_i) =$

$\eta_i \overset{\text{identity link}}{=}$

$\frac{\partial \eta_i}{\partial \mu_i} =$

$z_i = \widehat{\eta}_i +$

$w_{ii}^{-1} =$

■

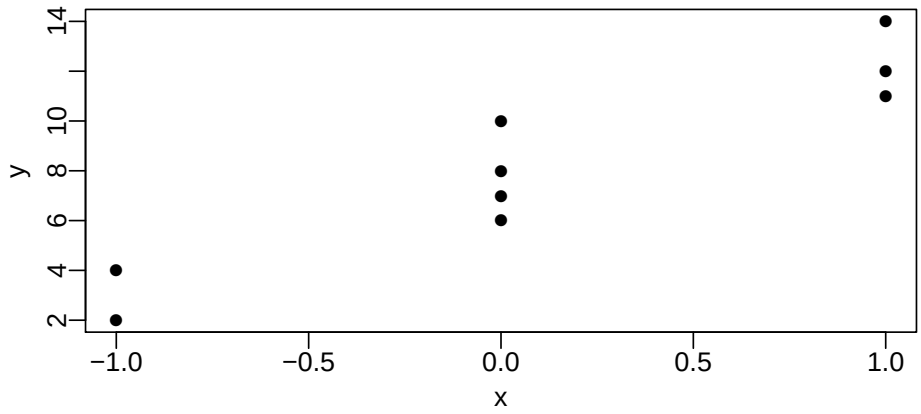
Example in R

`poisson.R poisson-data.RData`

Consider the following artificial dataset

	1	2	3	4	5	6	7	8	9
y	2	4	6	7	8	10	11	12	14
x	-1	-1	0	0	0	0	1	1	1

Table 3.1: Example Dataset



A possible model may be one using a Poisson distribution as the response distribution. Lets model the relationship between Y_i and x_i as

$$\mu_i \equiv \mathrm{E}(Y_i) = \beta_1 + \beta_2 x_i, \quad \text{for } i = 1, \ldots, 9,$$

and $Y_i \sim \text{Poisson}(\mu_i)$. Thus we have taken the link function as the identity; then

$$\frac{\partial \mu_i}{\partial \eta_i} = 1 \quad \text{and} \quad w_{ii} = \frac{1}{\text{var}(Y_i)} = \frac{1}{\beta_1 + \beta_2 x_i}.$$

The design matrix is

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_9 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$$

Start by forming the response vector

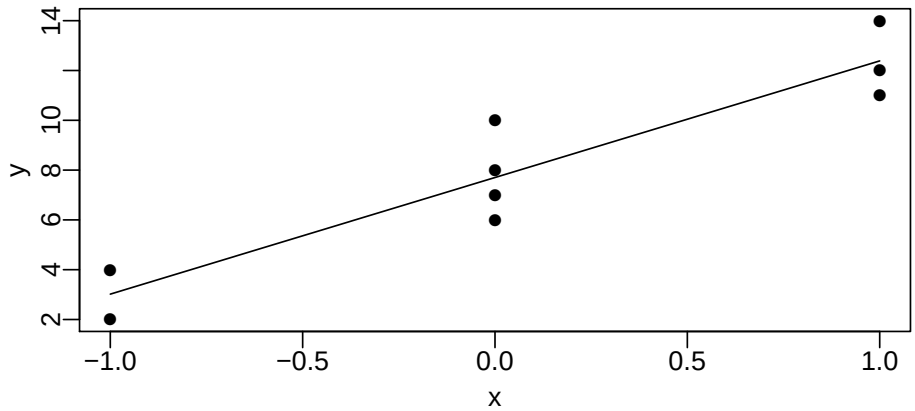
```
> y <- dat$y
```

We can then fit the model using the iterative weighted least squares algorithm (Algorithm 3.1).

```
> beta <- c(20,4) #initial guess
> for (i in 1:25){
+     eta <- cbind(1,x)%*%beta #estimated linear predictor
+     mu <- eta  #estimated mean response
+     z <- y #form the adjusted variate
+     w <- 1/mu #weights
+     lmod <- lm(z~x, weights=w) #regress z on x with weights w
+     beta <- as.numeric(lmod$coeff) #new beta
+     print(beta) #print out the beta estimate every iteration
+ }
```

```
[1] 7.703704 4.666667
[1] 7.701896 4.682932
[1] 7.701886 4.683026
[1] 7.701886 4.683027
...
```

The algorithm seems to converge at $\boldsymbol{\beta}$ =(7.702, 4.683).



Using (3.13) we can work out the estimated standard error of these parameters:

```
> X <- cbind(1,x)
> J <- t(X)%*%diag(1/as.vector(mu))%*%X
> invJ <- solve(J)
> sqrt(as.vector(diag(invJ)))
```

```
[1] 0.9020884 1.1317672
```

In R there is a function that automatically fits the model for us:

```
> myglm <- glm(y~x,family=poisson(link="identity"))
> summary(myglm)
```

```
Call:
glm(formula = y ~ x, family = poisson(link = "identity"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6382  -0.4012  -0.1100   0.4495   0.7913

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.7019     0.9021   8.538  < 2e-16 ***
x             4.6830     1.1318   4.138 3.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16.4022  on 8  degrees of freedom
Residual deviance:  2.1658  on 7  degrees of freedom
AIC: 40.682

Number of Fisher Scoring iterations: 4
```

The estimates of the parameter and its standard deviation seem to agree, which should be no surprise as the estimation procedures are exactly the same.

Notes

- The estimation procedure may fail (e.g. see later in Section 3.8.3) – in `R` the maximum number of iterations is 25 by default.
- The convergence criterion used by the `glm` command in `R` will be discussed later in Section 3.6.1.
- A difference between the results may be due to a different initial starting point.
- Notice that the dispersion parameter, ϕ, is recognised by `R` as 1.

■

3.3 Inference

As in the inference section for linear models, section 2.3, we discuss how to construct confidence intervals and perform hypothesis tests. In particular, for hypothesis testing we shall discuss how to compare two related models. For GLMs, we say that two models are related if

1. the distribution of the response $\boldsymbol{Y}$ is the same; and
2. the same link functions is used.

The models differ in the number of parameters i.e. the dimensionality of $\boldsymbol{\beta}$. To compare models we require a measure of their **goodness of fit**. We shall present goodness of fit statistics based on the log-likelihood function.

3.3.1 Sampling Distributions for GLMs

In this section, we discuss the sampling distributions[†] relevant to GLMs. We shall use the following notion: under appropriate regularity conditions, which are satisfied for generalized linear models, if S is a statistic of interest, then

$$\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{var}(S)}} \dot{\sim} N(0,1)$$

or equivalently

$$\frac{(S - \mathrm{E}(S))^2}{\mathrm{var}(S)} \dot{\sim} \chi^2_1,$$

where we shall use $\dot{\sim}$ to denote "approximately distributed as".

If there is a vector of statistics of interest

$$\boldsymbol{S} = \begin{pmatrix} S_1 \\ \vdots \\ S_p \end{pmatrix}$$

with asymptotic expectation $\mathrm{E}(\boldsymbol{S})$ and asymptotic variance-covariance matrix Q, then asymptotically

$$\boldsymbol{S} - \mathrm{E}(\boldsymbol{S}) \dot{\sim} N(\boldsymbol{0}, Q) \tag{3.14}$$

and further, asymptotically

$$(\boldsymbol{S} - \mathrm{E}(\boldsymbol{S}))^T Q^{-1} (\boldsymbol{S} - \mathrm{E}(\boldsymbol{S})) \dot{\sim} \chi^2_p \tag{3.15}$$

provided Q is non-singular so that Q^{-1} exists and is unique.

[†]The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample

To be clear, these approximations follow from the central limit theorems:

Central Limit Theorem

Suppose $A_1, \ldots, A_n$ are iid random variables with $\mathrm{E}(A_i) = \mu$ and $\mathrm{var}(A_i) = \sigma^2$ (both finite). Then

$$\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} A_i - \mu}{\sqrt{\sigma^2}}\right) \xrightarrow{\mathrm{d}} N(0,1) \quad \text{as } n \to \infty$$

where $\xrightarrow{\mathrm{d}}$ means convergence in distribution.

Multivariate Central Limit Theorem

If $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are iid random vectors with $\mathrm{E}(\boldsymbol{A}_i) = \boldsymbol{\mu} \in \mathbb{R}^p$ and finite, positive definite, symmetric covariance matrix Q then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i - \boldsymbol{\mu}\right) \xrightarrow{\mathrm{d}} N(\boldsymbol{0}, Q) \quad \text{as } n \to \infty$$

The asymptotic chi-squared result (3.15) follows from the fact: If $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, Q)$ where $\mu \in \mathbb{R}^p$ and $Q \in \mathbb{R}^{p \times p}$ is a positive definite, symmetric matrix, then

$$(\boldsymbol{Z} - \boldsymbol{\mu})^T Q^{-1} (\boldsymbol{Z} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Sampling distribution for the score

We can apply the ideas above to the score vector $\boldsymbol{U}$. Recall from (3.6) that

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right], \quad j = 1, \ldots, p.$$

Since $E(\boldsymbol{U}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{U}) = \mathcal{J}$ we have, from (3.14) asymptotically

$$\boldsymbol{U} \dot{\sim} N(\boldsymbol{0}, \mathcal{J}) \tag{3.16}$$

and, from (3.15), asymptotically,

$$\boldsymbol{U}^T \mathcal{J}^{-1} \boldsymbol{U} \dot{\sim} \chi^2_p \tag{3.17}$$

Example Suppose $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu, \sigma^2)$ where μ is the parameter of interest and $\sigma^2 > 0$ is a known constant. The log-likelihood function is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) =$$

Then the score is

$$U =$$

and the Fisher's information matrix is

$$\mathcal{J} =$$

Then by rearranging (3.16), we get (asymptotically)

$$\overline{Y} \sim N(\mu, \sigma^2/n).$$

We can use this result to construct a confidence interval for μ. For example, a 95% confidence interval for μ is $\overline{y} \pm 1.96\sigma/\sqrt{n}$ approximately due to asymptotic normality[‡]. ■

Sampling distribution for MLEs

Let $\widehat{\boldsymbol{\beta}}$ be the MLE for $\boldsymbol{\beta}$ in a GLM i.e.

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\text{argmax}}\, \ell(\boldsymbol{\beta})$$

and so $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \frac{\partial \ell(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \mathbf{0}$.

Now consider the 1st order Taylor expansion of $\boldsymbol{U}(\boldsymbol{\beta})$ about the MLE $\widehat{\boldsymbol{\beta}}$:

$$\boldsymbol{U}(\boldsymbol{\beta}) \approx$$

where we approximated $\boldsymbol{U}'$ by $\mathrm{E}(\boldsymbol{U}') = -\mathcal{J}$. Consequently, we find

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left\{\mathcal{J}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \boldsymbol{U}(\boldsymbol{\beta}),$$

assuming that $\mathcal{J}$ is invertible. If $\mathcal{J}$ is regarded as a constant then $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. It turns out that this is at least asymptotically true.

The variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right] =$$

[‡]Actually, this is exact and not approximate as the observations are Normal in this example. This coincides with the results presented in Chapter 2

where we have used that $\mathcal{J} = \mathcal{J}^T$ and treated $\mathcal{J}$ as a constant.

Thus, by (3.14), we have

$$\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \mathcal{J}^{-1}). \tag{3.18}$$

and also, using (3.15), we have

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathcal{J}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \dot{\sim} \chi^2_p. \tag{3.19}$$

This is called the **Wald statistic**.

We can use (3.18) to construct confidence intervals and conduct hypothesis tests involving the components of $\boldsymbol{\beta}$. For instance, from (3.18), we have

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\text{var}(\widehat{\beta}_1)}} \dot{\sim} N(0,1), \quad \text{where} \quad \text{var}(\widehat{\beta}_1) = \underbrace{(X^T W X)^{-1}_{11}}_{\text{the 1,1 entry of } (X^T W X)^{-1}}$$

which we can use to construct confidence intervals for β_1. Further, under the null hypothesis $H_0 : \beta_1 = 0$ we have $\frac{\widehat{\beta}_1}{\sqrt{\text{var}(\widehat{\beta}_1)}} \dot{\sim} N(0,1)$.

3.4 Prediction

For given covariates $\boldsymbol{x}_\star$, we have[§] $\widehat{\eta}_\star = \boldsymbol{x}_\star^T \widehat{\boldsymbol{\beta}}$ with variance $\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star$. Therefore, an approximate confidence interval can be constructed using the normal distribution. To obtain an approximate confidence interval in terms of μ we can use the inverse of the link function to transform the end points. As with the linear models, we can use the `predict` function in `R` for GLMs. More precisely, an approximate 95% confidence interval for the mean response of a prediction with covariates $\boldsymbol{x}_\star$ is

$$\left(g^{-1}\left(\widehat{\eta}_\star - 1.96\sqrt{\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star} \right), g^{-1}\left(\widehat{\eta}_\star + 1.96\sqrt{\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star} \right) \right),$$

where g^{-1} is the inverse of the link function g.

[§] $\boldsymbol{x}_\star$ is a column vector containing covariates and typically an intercept. Also, note that $\widehat{\eta}_\star$ is just a number.

3.4.1 Measuring the Goodness of Fit

One measure of the goodness of fit of a GLM is to compare it with a more general model with the maximum number of parameters that can be estimated. This model is called the saturated model. It is a GLM with exactly the same response distribution and link function as the model of interest.

For n observations, $y_1, \ldots, y_n$, all with potentially different values for the linear predictors $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, a saturated model can be specified with n parameters. It turns out that the fitted saturated model forces $\mu_i \equiv \mathrm{E}(Y_i) = y_i$ for $i = 1, \ldots, n$ which achieves the maximum attainable log-likelihood.

For n observations from a GLM with n parameters, the log-likelihood is

$$\ell(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

Then, for $i = 1, \ldots, n$,

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

For this it is clear that log-likelihood is maximised if we force $\mu_i = y_i$ for $i = 1, \ldots, n$.

Now reconsider the log-likelihood in terms of the mean response $\boldsymbol{\mu}$:

$$\ell(\boldsymbol{\mu}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{a(\phi)} + c(y_i, \phi) \right).$$

Thus, the maximum achievable log-likelihood is $\ell(\boldsymbol{y}, \phi; \boldsymbol{y})$.

For the model of interest, a GLM with typically less parameters than observations ($p < n$), we can maximise the log-likelihood to get $\widehat{\boldsymbol{\beta}}$, then $\widehat{\boldsymbol{\eta}} = X\widehat{\boldsymbol{\beta}}$ and so the estimated mean response $\widehat{\boldsymbol{\mu}} = g^{-1}(\widehat{\boldsymbol{\eta}})$ — leading the maximised log-likelihood $\ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$.

Deviance

A measure of the goodness of fit of a GLM is the deviance.

Definition. The deviance for a model with estimated mean response $\widehat{\boldsymbol{\mu}}$ is defined as

$$D = 2\phi \left\{ \ell(\boldsymbol{y}, \phi; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y}) \right\},$$

and the scaled deviance is $D^* = D/\phi$.

To make things concrete: Suppose we have a GLM of interest and consider the following extreme models with the same response distribution and link function:

Saturated Model: The GLM with number of parameters, p, equal to the number of (distinct) observations. In this case, $\mu_i \equiv \mathrm{E}(Y_i)$ is equal to the observed response y_i, so there is no variation in the random component.

Null Model: The GLM with only one parameter ($p = 1$) representing a common mean response μ for all ys.

In practice, the null model will be too simple and the saturated model is uninformative as it just repeats the observed response. We aim for a model with a likelihood close to the likelihood of the saturated model, but with fewer parameters.

Deviance — constrained, unconstrained likelihood?

Consider a saturated model where $\boldsymbol{\beta} \in \mathbb{R}^n$ and (as usual) $\boldsymbol{\eta} = X\boldsymbol{\beta}$, with $X \in \mathbb{R}^{n \times n}$. Assuming that X is invertible, the model imposes no constraints on the linear predictor $\boldsymbol{\eta}$ — it can take any value on $\mathbb{R}^n$. In turn, this means $\boldsymbol{\mu}$ and also θ are unconstrained. Thus, for the saturated model ; the maximised log-likelihood is

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n \text{ s.t. } \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})} \; \max_{\boldsymbol{\eta} \in \mathbb{R}^n} \ell(\boldsymbol{\mu}; \boldsymbol{y}) = \ell(\boldsymbol{y}; \boldsymbol{y}).$$

For models with $p < n$ parameters, the maximised log-likelihood is

$$\max_{\substack{\boldsymbol{\mu} \in \mathbb{R}^n \text{ s.t. } \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta}) \\ \text{for some } \boldsymbol{\beta} \in \mathbb{R}^p}} \ell(\boldsymbol{\mu}; \boldsymbol{y}) = \ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y})$$

For instance, suppose we have a GLM with $n = 3$, $p = 1$ with $X = (1\ 1\ 1)^T$ and identity link function $g(z) = z$. Then

$$\eta_i = \beta \in \mathbb{R} \text{ for } i = 1, \ldots, n,$$

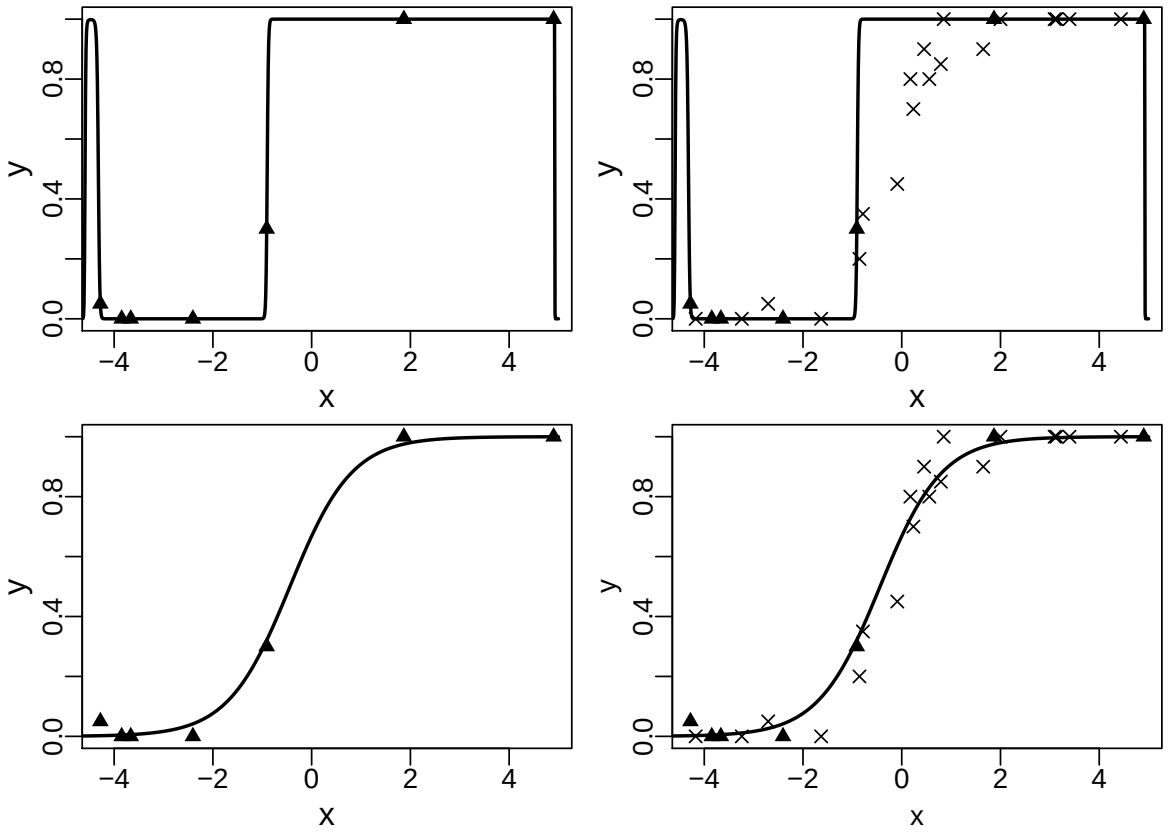
and

$$\boldsymbol{\mu} = \begin{pmatrix} g^{-1}(\beta) \\ g^{-1}(\beta) \\ g^{-1}(\beta) \end{pmatrix} = \beta \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Thus $\boldsymbol{\mu}$ is constrained to lie on a line in $\mathbb{R}^3$.

Why we do not use the saturated model.

We can make the fit as close as possible, in terms of minimising the deviance, by including sufficiently many parameters. Below are some fits for a GLM based on the 7 data points represented as triangles. The first row corresponds to a saturated model — note how the fit goes through every data point. The second row is another fit using just 2 parameters. The second column is a repeat of the first, but with additional data points — note how the saturated model would give poor predictions for the additional data points. Thus simplicity, represented by parsimony of parameters, is a desirable feature for models; we do not include parameters that are unnecessary.



Let us now switch back, considering the log-likelihood in terms of $\boldsymbol{\beta}$ — write the scaled deviance as

$$D^* = 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y})\right\},$$

where $\widehat{\boldsymbol{\beta}}_{sat}$ is the MLE of $\boldsymbol{\beta}_{sat}$ for the saturated model and $\widehat{\boldsymbol{\beta}}$ is still the MLE of $\boldsymbol{\beta}$ in the model of interest.

Example Consider the with $Y_1, \ldots, Y_n$ independent where $Y_i \sim N(\mu_i, \sigma^2)$ and

$$\mathrm{E}(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

The log-likelihood function is

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^T(\boldsymbol{y} - \boldsymbol{\mu}).$$

Setting $\boldsymbol{\mu} = \boldsymbol{y}$ yields the maximum achievable log-likelihood:

$$\ell(\boldsymbol{y}; \boldsymbol{y}) =$$

Then for any other model with $p < n$ (not a saturated model) we find the maximum likelihood estimate $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ — which in turn gives the estimated mean response $\widehat{\boldsymbol{\mu}} = g^{-1}(\widehat{\boldsymbol{\eta}}) = X\widehat{\boldsymbol{\beta}}$. So

$$\ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) =$$

Therefore the deviance is

$$D = 2\phi\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y})\right\}$$

■

3.4.2 Pearson's X^2 statistic

Another important measure of the discrepancy of a GLM is Pearson's X^2 statistics.

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Pearson's X^2 statistic shares similar asymptotic distribution properties with the deviance. However, in this course we shall focus on the deviance as a measure of goodness-of-fit as

- the deviance has a general advantage as a discrepancy measure in that it is additive for nested sets of models (see next section).
- the deviance leads to better normalised residuals.

3.4.3 Hypothesis Testing

We now discuss how to compare GLMs by hypothesis testing. These tests will involve the deviance and scaled deviance, therefore we first need their sampling distributions.

Model Comparison with known ϕ

First, consider the 2nd order Taylor expansion of the log-likelihood around the MLE $\widehat{\boldsymbol{\beta}}$:

$$\ell(\boldsymbol{\beta}) \approx$$

where we approximated $\boldsymbol{U}'$ by $\mathrm{E}(\boldsymbol{U}') = -\mathcal{J}$. Consequently, we find

$$2\left\{\ell(\widehat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta})\right\} = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \mathcal{J}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \tag{3.20}$$

which is approximately χ^2_p distributed by (3.19). Rewriting the deviance and using (3.20) gives

$$D^* =$$

If the model with p parameters is correct, then

$$D^* \dot{\sim} \chi^2_{n-p}.$$

Now consider comparing two nested models as follows: Consider the null hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix},$$

corresponding to model M_0 and a more general hypothesis, the alternative,

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

corresponding to model M_1 with $q < p < n$. We can test H_0 against H_1 using the difference of the scaled deviances

$$\begin{aligned} D_0^* - D_1^* &= 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y})\right\} - 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y})\right\} \\ &= 2\left\{\ell(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y})\right\}. \end{aligned}$$

The statistic $D_0^* - D_1^*$ is approximately χ^2_{p-q} distributed under the null hypothesis.

Model Comparison with unknown ϕ

Under the null hypothesis we have

$$D_1^* \dot{\sim} \chi^2_{n-p} \quad \text{and} \quad D_0^* - D_1^* \dot{\sim} \chi^2_{p-q}.$$

If we consider D_1^* and $D_0^* - D_1^*$ as asymptotically independent, then

$$\frac{(D_0^* - D_1^*)/(p-q)}{D_1^*/(n-p)} \dot{\sim} F_{(p-q),(n-p)}.$$

The advantage of this test statistic is that we can multiply top and bottom by ϕ to get a test statistic based on the deviance:

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} \dot{\sim} F_{(p-q),(n-p)}.$$

The advantage of this hypothesis test for model comparison is that it does not depend on ϕ.

Warning

There are certain assumptions made when proving that the deviance, D, is approximately or asymptotically χ^2_{n-p} distributed which we should be wary of

- The observations are independent and distributed according to some member of an exponential family.
- The approximation relies on the number of parameters in the model staying fixed, while the sample size tends to infinity. But the saturated model has as many parameters as number of data. Also, in the Binomial case the χ^2_{n-p} approximation (or asymptotics) for the deviance are based upon the following assumptions (McCullagh & Nelder, 1989, p. 118)
  - the observations are truly distributed independently according to the binomial distribution;
  - The approximation is based on a limiting operation in which n is fixed and $n_i \to \infty$ for each i (and in fact $n_i\pi_i(1-\pi_i) \to \infty$).

However, the χ^2 approximation is sound when comparing two nested models, as the deviance for the saturated model cancels out.

3.4.4 Estimating the Dispersion

Recall that the MLE for $\boldsymbol{\beta}$ does not depend on the dispersion ϕ. However, in cases where the dispersion is unknown, it must be estimated. There are two commonly used estimators:

First is based on the deviance:

$$\widehat{\phi}_D = \frac{D}{n-p},$$

where D is the deviance of the model. This follows from the expected value of $D/\phi = D^* \dot{\sim} \chi^2_{n-p}$.

Second is

$$\widehat{\phi}_P = \frac{X^2}{n-p}$$

where X^2 is Pearson's statistic — this is based on the approximation $X^2/\phi \dot{\sim} \chi^2_{n-p}$.

Note

If $Z \sim \chi^2_d$ then $\mathrm{E}(Z) = d$; that is the expected value of a χ^2 distribution is equal to its degrees of freedom.

We can then plug in an estimator for the dispersion parameter ϕ to estimate $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$:

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) \approx \widehat{\phi}(X^T \widetilde{W} X)^{-1} \quad \text{(see page 59)}.$$

3.4.5 Akaike's Information Criteria (AIC)

An alternative statistic to compare two models was suggested by Akaike: pick whichever model minimises

$$\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\beta}}) + 2p,$$

for a given set of data. Notice, that it is similar to using the log-likelihood of the data, but with a penalisation term for the number of parameters i.e. the more parameters included, the higher the AIC. Therefore the AIC involves a trade-off between goodness-of-fit of the model and its complexity (in terms of its number of parameters).

3.5 Diagnostics

As with the normal linear models, we can use residuals to explore the fit of GLMs. For GLMs we require extenstions of residual definitions to accommodate for all non-Normal distributions.

3.5.1 Residuals

Pearson's Residuals

Definition. For a single observation y, Pearson's residual is defined as

$$r_p = \frac{y - \mu}{\sqrt{V(\mu)}}.$$

Deviance Residuals

Suppose the deviance, D, is used as a measure of discrepancy of a GLM. Then each observation contributes a quantity d_i to D so that "$D = \sum_{i=1}^{n} d_i$". Thus, it makes sense to define a deviance based residual.

Definition. For a single observation, y_i, the deviance residual is defined as

$$r_D = \text{sign}(y_i - \mu_i)\sqrt{d_i},$$

thus the deviance is $D = \sum_i r_D^2$.[¶]

Example For the Poisson distribution, we have

Pearson's residual: $r_p =$

Deviance residual: $r_D =$

■

Similar with the residuals for linear models, we can standardise to account for each observation's leverage. To this end, we require the corresponding hat matrix for GLMs:

$$P = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}.$$

The standardised Pearson's and deviance residuals are obtained by dividing by $\sqrt{(1 - P_{ii})\widehat{\phi}}$.

[¶] $\text{sign}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$

Note

These residuals are approximately normal in general. However, the deviance residuals are the closest, for n large and n_i (for Binomial) large. In practice, one should not expect them to lie on a straight line in a QQ plot, but rather on a smooth curve — departure from this curve may indicate an outlier.

3.5.2 Cook's Distance

Similar to the normal linear model case, we can examine the Cook's distance for the observations.

$$C_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T (X^T W X) \left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{p\widehat{\phi}}$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimator calculated without using the ith observation. Again we look for large C_i close to 1.

3.6 Worked Examples

3.6.1 Worked Example 1

`seeds-data.RData` `seeds.R`

The data presented in Table 3.2 shows the number of Orobanche seeds germinated for two genotypes and two treatments.

#	Germinated y	Total tested n	Genotype x_1	Treatment x_2
1	10	39	0	0
2	23	62	0	0
3	23	81	0	0
4	26	51	0	0
5	17	39	0	0
6	5	6	0	1
7	53	74	0	1
8	55	72	0	1
9	32	51	0	1
10	46	79	0	1
11	10	13	0	1
12	8	16	1	0
13	10	30	1	0
14	8	28	1	0
15	23	45	1	0
16	0	4	1	0
17	3	12	1	1
18	22	41	1	1
19	15	30	1	1
20	32	51	1	1
21	3	7	1	1

Table 3.2: Orobanche seeds Dataset

We are interested in, y_i/n_i, the proportion, so we want the fit to remain between 0 and 1. Suppose $Y_1, \ldots, Y_n$ are independent Binomial(n_i, π_i) random variables and take the logit link function. Take the design matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{21,1} & x_{21,2} & x_{21,1}x_{21,2} \end{pmatrix}$$

Using the binomial distribution we would have $\mu_i \equiv \mathrm{E}(Y_i) = n_i\pi_i$ with pdf

$$\prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} \left(1 - \pi_i\right)^{n_i - y_i}.$$

Now we fit the model using the IWLS algorithm (Algorithm 3.1)

```
> beta <- c(0.5,0.5,0,0) #initial guess
> #inverse logit function
> inv.link <- function(u)
+     n*(1/(1+exp(-u)))
> #deviance function
> D <- function(p){
+     a <- y*log(y/p)
+     b <- (n-y)*log((n-y)/(n-p))
+     a[y==0] <- 0
+     2*sum(a+b)
+ }
> oldD <- D(inv.link(as.numeric(X%*%beta)))
> jj <- 0
> while(jj==0){
+     eta <- X%*%beta #estimated linear predictor
+     mu <- inv.link(eta)  #estimated mean response
+     detadmu <- n/(mu*(n-mu))
+     z <- eta+ (y-mu)*detadmu #adjusted dependent variable
+     w <- mu*(n-mu)/n #weights
+     lmod <- lm(z~x, weights=w) #weighted regression of z on x
+     beta <- as.vector(lmod$coeff) #new beta
+     newD <- D(inv.link(X%*%beta))
+     control <- abs(newD-oldD)/(abs(newD)+0.1)
+     if(control<1e-8)
+         jj <- 1
+     oldD <- newD
+ }
> beta #final estimate
```

```
[1] -0.5581717  0.1459269  1.3181819 -0.7781037
```

```
> newD #last deviance calculated
```

```
[1] 33.27779
```

Notice that this time we have used the change in the deviance as the convergence criterion; this is what `R` uses: If

$$\frac{\left|D^{\text{new}} - D^{\text{old}}\right|}{\left|D^{\text{new}}\right| + 0.1}$$

is less than 1×10^{-8} then the algorithm is deemed to have converged.

By (3.13), the standard errors for these estimates are

```
> J <- t(X)%*%diag(as.vector(w))%*%X
> invJ <- solve(J)
> beta.sd <- sqrt(as.vector(diag(invJ)))
> beta.sd
```

```
[1] 0.1260213 0.2231657 0.1774677 0.3064330
```

The deviance residuals (3.5.1) are

```
> p <- as.vector(inv.link(X%*%beta))
> a <- y*log(y/p)
> b <- (n-y)*log((n-y)/(n-p))
> a[y==0] <- 0
> d <- sign(y-mu)*sqrt(2*(a+b))
> summary(d)
```

```
       V1
 Min.   :-2.01617
 1st Qu.:-1.24398
 Median : 0.05995
 Mean   :-0.08655
 3rd Qu.: 0.84695
 Max.   : 2.12122
```

We can test individual parameters using (3.18). The corresponding p-values are:

```
> z <- beta/beta.sd
> z
```

```
[1] -4.429187  0.653895  7.427729 -2.539229
```

```
> 2*(1-pnorm(abs(z),lower.tail=TRUE))
```

```
[1] 9.458885e-06 5.131795e-01 1.105782e-13 1.110970e-02
```

The AIC (section 3.4.5) is

```
> -2*sum(dbinom(y,n,as.vector(mu/n),log=TRUE)) + 2*length(beta)
```

```
[1] 117.874
```

Of course, R can do this for us:

```
> dat2 <- seeds
> y <- cbind(dat2$r, dat2$n - dat2$r)
> my.bin.glm <- glm(y ~ seed * extract, family = binomial,
+     data = dat2)
> summary(my.bin.glm)
```

```
Call:
glm(formula = y ~ seed * extract, family = binomial, data = dat2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.5582     0.1260  -4.429 9.46e-06 ***
seed           0.1459     0.2232   0.654   0.5132
extract        1.3182     0.1775   7.428 1.10e-13 ***
seed:extract  -0.7781     0.3064  -2.539   0.0111 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

Number of Fisher Scoring iterations: 4
```

3.6.2 Worked Example 2

The data used in this example can be directly loaded in `R` as `infert`.

The data are from a case-control study of secondary infertility - this is where someone has had one or more pregnancies in the past, but is having difficulty conceiving again. The command `?infert` provides a summary of what each variable represents, and a link to the original study. Table 3.3 presents the first 6 data only — in total there are 248 cases.

	education	age	parity	induced	case	spontaneous	stratum	pooled.stratum
1	0-5yrs	26	6	1	1	2	1	3
2	0-5yrs	42	1	1	1	0	2	1
3	0-5yrs	39	6	2	1	0	3	4
4	0-5yrs	34	4	2	1	0	4	2
5	6-11yrs	35	3	1	1	1	5	32
6	6-11yrs	36	4	2	1	1	6	36

Table 3.3: Infertility Dataset

Let us fit a binomial model to see if `case`, a binary observation, can be predicted using the other measures as covariates. Note that ~. means that we include all other data columns singly (useful for large datasets).

```
> myglm0 <- glm(case ~.,data=infert,family=binomial)
> summary(myglm0)
```

```
Call:
glm(formula = case ~ ., family = binomial, data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q
-1.7975  -0.7836  -0.4599   0.8556
    Max
 2.8998

Coefficients:
                  Estimate Std. Error
(Intercept)      -4.039297   2.135797
education6-11yrs  1.320471   1.565614
education12+ yrs  3.489701   2.965837
age               0.078590   0.038060
parity           -0.451423   0.276877
```

```
induced           1.435629   0.320870
spontaneous       2.191282   0.329069
stratum          -0.002842   0.014621
pooled.stratum   -0.078768   0.043800
                 z value Pr(>|z|)
(Intercept)       -1.891   0.0586 .
education6-11yrs   0.843   0.3990
education12+ yrs   1.177   0.2393
age                2.065   0.0389 *
parity            -1.630   0.1030
induced            4.474 7.67e-06 ***
spontaneous        6.659 2.76e-11 ***
stratum           -0.194   0.8459
pooled.stratum    -1.798   0.0721 .
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 254.53  on 239  degrees of freedom
AIC: 272.53

Number of Fisher Scoring iterations: 4
```

At first glance, the residual deviance looks reasonable for the degrees of freedom.

We can look at sequentially adding terms using (somewhat confusingly the `anova` function again):

```
> anova(myglm0,test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: case

Terms added sequentially (first to last)
```

```
              Df Deviance Resid. Df
NULL                             247
education      2    0.002       245
age            1    0.006       244
parity         1    0.026       243
induced        1    0.056       242
spontaneous    1   58.284       241
stratum        1    0.003       240
pooled.stratum 1    3.263       239
              Resid. Dev  Pr(>Chi)
NULL              316.17
education         316.17   0.99886
age               316.16   0.94012
parity            316.14   0.87088
induced           316.08   0.81372
spontaneous       257.80 2.269e-14 ***
stratum           257.79   0.95346
pooled.stratum    254.53   0.07085 .
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

The `anova` function recognises that a GLM was fitted and produces an analysis of deviance table. The `test="Chisq"` option reports the *p*-values on the right.

Model comparisons and selection can be done automatically in `R`. First, consider similar models where just one parameter is dropped. This is achieved using the `drop1` function:

```
> drop1(myglm0,test="Chisq")
```

```
Single term deletions

Model:
case ~ education + age + parity + induced + spontaneous +
         stratum + pooled.stratum
               Df Deviance    AIC
<none>              254.53 272.53
education       2   256.76 270.76
age             1   258.85 274.85
```

```
parity          1   257.26 273.26
induced         1   277.42 293.42
spontaneous     1   316.03 332.03
stratum         1   254.57 270.57
pooled.stratum  1   257.79 273.79
                  LRT  Pr(>Chi)
<none>
education       2.230   0.32791
age             4.315   0.03778 *
parity          2.728   0.09861 .
induced        22.890 1.716e-06 ***
spontaneous    61.504 4.419e-15 ***
stratum         0.038   0.84591
pooled.stratum  3.263   0.07085 .
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

This suggests dropping the `education` and `stratum` parameters (and perhaps more) from the model.

The `drop1` function has a brother function — the `add1` function, which adds individual terms to the given model:

```
> add1(myglm0,~.^2,test="Chisq")
```

```
Single term additions

Model:
case ~ education + age + parity + induced + spontaneous +
         stratum +  pooled.stratum

                          Df Deviance
<none>                         254.53
education:age              2   254.51
education:parity           2   254.33
education:induced          2   247.58
education:spontaneous      2   252.28
```

```
education:stratum           2   254.41
education:pooled.stratum    2   254.01
age:parity                  1   254.19
age:induced                 1   254.20
age:spontaneous             1   251.39
age:stratum                 1   254.36
age:pooled.stratum          1   254.51
parity:induced              1   252.90
parity:spontaneous          1   252.57
parity:stratum              1   254.48
parity:pooled.stratum       1   254.42
induced:spontaneous         1   254.50
induced:stratum             1   248.77
induced:pooled.stratum      1   250.36
spontaneous:stratum         1   251.27
spontaneous:pooled.stratum  1   253.63
stratum:pooled.stratum      1   254.42
                              AIC
<none>                      272.53
education:age               276.51
education:parity            276.33
education:induced           269.58
education:spontaneous       274.28
education:stratum           276.41
education:pooled.stratum    276.01
age:parity                  274.19
age:induced                 274.20
age:spontaneous             271.39
age:stratum                 274.36
age:pooled.stratum          274.51
parity:induced              272.90
parity:spontaneous          272.57
parity:stratum              274.48
parity:pooled.stratum       274.42
induced:spontaneous         274.50
induced:stratum             268.77
induced:pooled.stratum      270.36
spontaneous:stratum         271.27
spontaneous:pooled.stratum  273.63
stratum:pooled.stratum      274.42
```

```
                            LRT
<none>
education:age               0.0198
education:parity            0.1977
education:induced           6.9513
education:spontaneous       2.2478
education:stratum           0.1219
education:pooled.stratum    0.5227
age:parity                  0.3444
age:induced                 0.3282
age:spontaneous             3.1409
age:stratum                 0.1723
age:pooled.stratum          0.0205
parity:induced              1.6332
parity:spontaneous          1.9616
parity:stratum              0.0521
parity:pooled.stratum       0.1089
induced:spontaneous         0.0323
induced:stratum             5.7651
induced:pooled.stratum      4.1678
spontaneous:stratum         3.2615
spontaneous:pooled.stratum  0.9052
stratum:pooled.stratum      0.1120
                           Pr(>Chi)
<none>
education:age               0.99013
education:parity            0.90589
education:induced           0.03094 *
education:spontaneous       0.32500
education:stratum           0.94089
education:pooled.stratum    0.77003
age:parity                  0.55728
age:induced                 0.56670
age:spontaneous             0.07635 .
age:stratum                 0.67805
age:pooled.stratum          0.88626
parity:induced              0.20126
parity:spontaneous          0.16134
parity:stratum              0.81942
parity:pooled.stratum       0.74145
```

```
induced:spontaneous         0.85736
induced:stratum             0.01635 *
induced:pooled.stratum      0.04120 *
spontaneous:stratum         0.07093 .
spontaneous:pooled.stratum  0.34140
stratum:pooled.stratum      0.73790
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

The ~.^2 informs R to consider all possible two-factor interactions. The result above suggests including an education:induced, induced:stratum and induced:pooled.stratum terms.

The drop1 and add1 perform many individual χ^2 tests between models — it does not select any particular model.

The step function will automatically search through the models for us.

```
> stepsearch <- step(myglm0,~.^2,test="Chisq")
```

... this is produce a lot of output! To summarise the step search, we can use the anova component:

```
> stepsearch$anova
```

```
               Step Df Deviance
1                   NA       NA
2 + induced:stratum -1 5.765135
3       - education  2 1.749457
4 + age:spontaneous -1 4.072592
5  - pooled.stratum  1 1.888343
  Resid. Df Resid. Dev      AIC
1       239   254.5310 272.5310
2       238   248.7658 268.7658
3       240   250.5153 266.5153
4       239   246.4427 264.4427
5       240   248.3310 264.3310
```

Reading from top to bottom: a `induced:stratum` parameter was added, then the `education` parameter removed, etc. The criteria to add or remove parameters is based on the AIC. The step search continues until the AIC cannot be reduced any further.

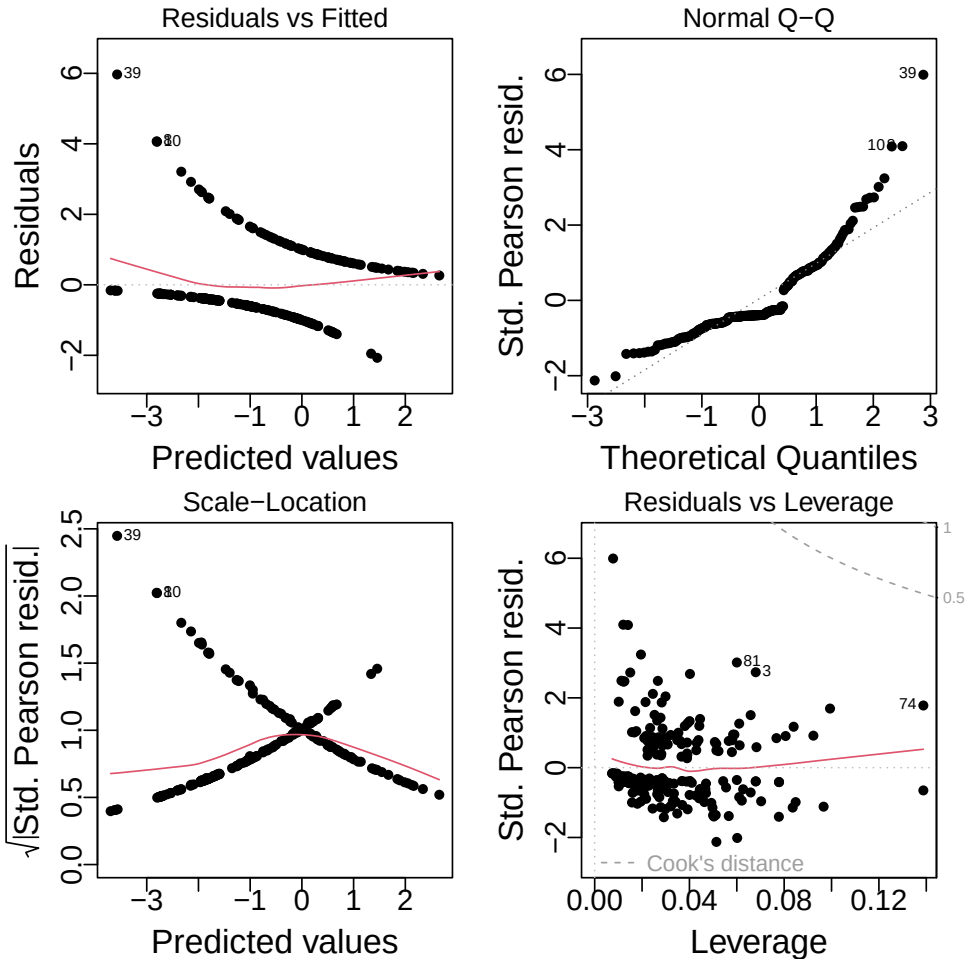
Finally, a summary oxf the final chosen model can be called (no need to fit again):

```
> summary(stepsearch)
```

```
Call:
glm(formula = case ~ age + parity + induced + spontaneous +
        stratum +  induced:stratum + age:spontaneous,
        family = binomial, data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q
-1.8253  -0.7699  -0.5257   0.8536
    Max
 2.6835

Coefficients:
                 Estimate Std. Error
(Intercept)     -1.086896   1.463758
age              0.000298   0.041023
parity          -0.880102   0.193553
induced          2.278841   0.487873
spontaneous     -0.503841   1.395070
stratum          0.003730   0.008352
induced:stratum -0.025723   0.009161
age:spontaneous  0.080038   0.044768
                z value Pr(>|z|)
(Intercept)      -0.743  0.45776
age               0.007  0.99420
parity           -4.547 5.44e-06 ***
induced           4.671 3.00e-06 ***
spontaneous      -0.361  0.71798
stratum           0.447  0.65519
induced:stratum  -2.808  0.00498 **
age:spontaneous   1.788  0.07380 .
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
```

```
  '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 248.33  on 240  degrees of freedom
AIC: 264.33

Number of Fisher Scoring iterations: 4
```

Finally, let us inspect the diagnostic plots.



The residual plots are less informative than they were for linear models. The response contains less information than a continuous one. Nevertheless, the issue of outliers and influential observations are just as prevalent in logistic regression as for linear models — so look at the Cook's distance plot and investigate any highly influential observations.[11]

[11] but this is not the end of the investigation...

Finally, for convenience, the residual values as follows:

```
> residuals(stepsearch,type="pearson")
> residuals(stepsearch,type="deviance")
```

and the Cook's distances:

```
> cooks.distance(stepsearch)
```

and the standardised residuals:

```
> rstandard(stepsearch,type="pearson")
> rstandard(stepsearch,type="deviance")
```

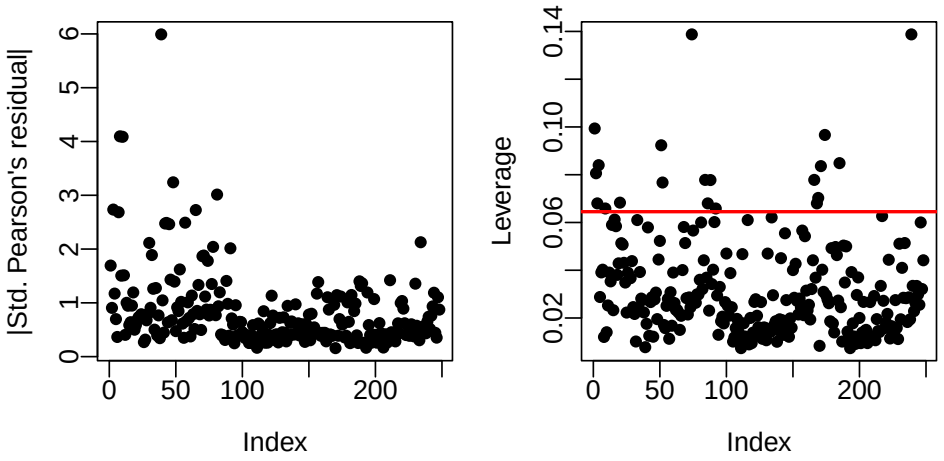
The commands above can be executed upon any `glm` model fitted in `R`.

We can proceed to check for large (in magnitude) residuals and observations with high leverages:

```
> plot(abs(rstandard(stepsearch,type="pearson")))
> plot(influence(stepsearch)$hat) #extract hat matrix values
> l.threshold <- 2*8/248 #rule of thumb
> l.threshold
```

```
[1] 0.06451613
```

```
> abline(h=l.threshold) #adds a horizontal line
```



We can mark any suspicious points and investigate if removing them significantly influences the fit.

A handy way of looking for these: First find the largest, say 5, standardised residuals:

```
> order(abs(rstandard(stepsearch,type="pearson")),
        decreasing=TRUE)[1:5]
```

```
[1] 39  8 10 48 81
```

These are the corresponding index numbers.

Next, find the observations 5 with the largest leverages:

```
> order(influence(stepsearch)$hat,decreasing=TRUE)[1:5]
```

```
[1]  74 239   1 174  51
```

If we see any reoccuring indices, we may want to investigate further, as they will have both high leverage and high magnitude residual.

3.7 Binomial Regression

Suppose the independent response variable $Y_1, \ldots, Y_n$ where $Y_i \sim \text{Binomial}(n_i, \pi_i)$, so that

$$\text{P}(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

Suppose that for the ith response we also observe covariates $x_{i,1}, x_{i,2} \ldots$. Following the linear model approach, we construct a linear predictor:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j.$$

We shall discuss the three common choices of link functions are used for binomial regression:

1) logit: $\eta = \log\left(\dfrac{\pi}{1-\pi}\right)$

2) probit: $\eta = \Phi^{-1}(\pi)$, where Φ^{-1} is the inverse cdf of a standard normal distribution.

3) complementary log log: $\eta = \log(-\log(1-\pi))$

Note

- A logistic binomial regression models is a GLM with binomial response and logit link function and a binary logistic binomial regression model is one with $n_i = 1$.
- The link above is defined through π and not $\mu \equiv n\pi$; this is annoying, but sometimes used in practice. You can use either π or μ to get the same results.

Worked Example

`challenger.R challenger-data.RData`

In January 1989, the space shuttle Challenger exploded shortly after launch. An investigation was conducted into the cause of the crash, paying particular attention to the O-ring seals in the rocket boosters. Could the failure of the O-rings have been predicted? Table 3.4 contains information about the damage of O-rings from 22 previous shuttle missions.

Temperature	No. Failure O-rings	Total No. O-rings
66	0	6
70	1	6
69	0	6
68	0	6
67	0	6
72	0	6
73	0	6
70	0	6
57	1	6
63	1	6
70	1	6
78	0	6
67	0	6
53	2	6
67	0	6
75	0	6
70	0	6
81	0	6
76	0	6
79	0	6
76	0	6
58	1	6

Table 3.4: Challenger Dataset

Let us proceed to fit a logistic regression model in R:

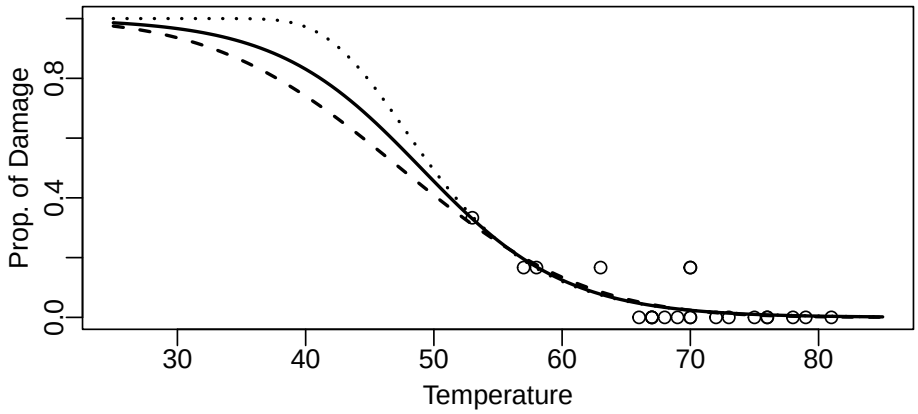
```
> mat <- cbind(dat$Fail,6-dat$Fail)
> logitmod <- glm(mat~Temp,
+                 family=binomial(link=logit),data=dat)
> probitmod <- glm(mat~Temp,
+                  family=binomial(link=probit),data=dat)
> cloglogmod <- glm(mat~Temp,
+                   family=binomial(link=cloglog),data=dat)
```

The estimated parameter values using the different link functions look very different:

```
                 logit     probit    cloglog
(Intercept)  8.6615667  4.1452794  7.9384946
Temp        -0.1768048 -0.0875188 -0.1665137
```

However, the actual fit given by each model are similar, especially around the observations:

```
> plot(x=dat$Temp,y=dat$Fail/6)
> x <- seq(25,85,by=0.5) # dummy x values
> logiteta <- 8.6615667-0.1768048*x
> probiteta <- 4.1452794-0.0875188*x
> cloglogeta <-  7.9384946-0.1665137*x
> lines(x,1/(1+exp(-logiteta)),lwd=2)
> lines(x,pnorm(probiteta),lwd=2,lty=2)
> lines(x,inv.clog(cloglogeta),lwd=2,lty=3)
```



We can predict the response given by each model at 31F:

```
> xstar <- 31
> 1/(1+exp(-(8.6615667-0.1768048*xstar)))
```

```
[1] 0.9600983
```

```
> pnorm(4.1452794-0.0875188*xstar)
```

```
[1] 0.9239562
```

```
> inv.clog(7.9384946-0.1665137*xstar)
```

```
[1] 0.9999999
```

The models suggest that there is a very high probability of damage at this temperature.

Let us check the deviance for the logistic model using the χ^2 test:

```
> pchisq(deviance(logitmod),df.residual(logitmod),lower=FALSE)
```

```
[1] 0.9776587
```

This p-value based on the χ^2 test of the deviance in well in excess of e.g. the 5% level. Thus we may conclude by saying model fits the data well. We may *not* say that the model is correct.

To construct a confidence interval for the prediction for the model using the logit link at 31F. Then

```
> ustar <- c(1,31)
> etastar <- ustar%*%logitmod$coefficients
> etastar
```

```
         [,1]
[1,] 3.180617
```

```
> 1/(1+exp(-etastar))
```

```
          [,1]
[1,] 0.9600983
```

```
> J <- vcov(logitmod)
> se <- sqrt(t(ustar)%*%J%*%ustar)
```

Then for an approximate 95% confidence interval (see section 3.4) on the probability scale:

```
> 1/(1+exp(-c(etastar - 1.96*se,etastar + 1.96*se)))
```

```
[1] 0.3957676 0.9988700
```

We can obtain the predictions directly using the `predict` command:

```
> predict(logitmod,newdata=list(Temp=31))
```

```
       1
3.180617
```

```
> predict(logitmod,newdata=list(Temp=31),type="response")
```

```
        1
0.9600983
```

The `predict` function can also be used to extract the fitted mean response values, $\widehat{\boldsymbol{\mu}}$:

```
> predict(logitmod,type="response")
```

The same procedure above can be applied for the other link functions.

Odds

Odds are sometimes a better scale than probability to represent chance.

- A 4-1 against bet would pay £4 for every £1.
- A 4-1 on bet would pay £1 for every £4.

Let p be the probability and o be the odds, where we represent e.g. 4-1 against as $1/4$ and 4 - 1 on as 4. In general, we have the relationship

$$o = \frac{p}{1-p} \quad \text{and} \quad p = \frac{o}{1+o}.$$

Odds Ratio

Suppose we have two groups where the probability of an event being in the first group is p_1 and the probability for the second group is p_2. Then the odds ratio is:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

An odds ratio

- equal to 1 indicates the event is equally likely to occur in both groups.
- greater than 1 indicates the event is more likely to occur in the first group.
- less than 1 indicates the event is less likely to occur in the first group.

Odds Interpretation

Suppose we have a logistic Binomial regression model using two covariates x_1 and x_2. The linear predictor is given by

$$\eta = \log\left(\frac{p}{1-p}\right) = \log(o) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

We can interpret β_1 as follows: a unit increase in x_1 keeping x_2 held fixed increases the log-odds of success by β_1, or equivalently increases the odds of success by a factor of $\exp(\beta_1)$.

Note

This interpretation follows from using the logit link — no simple interpretation exists for other links.

3.7.1 Worked Example

`breastfeeding.R  breastfeeding-data.RData`

Consider the data in Table 3.5 containing information from a study on infant respiratory disease by type of breast feeding and gender. The ratios presented are of the form "# with disease / total #".

	Bottle only	Some breast with supplement	Breast only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

Table 3.5: Breast Feeding data

Can gender and feeding type be used to measure whether or not infants contract a respiratory disease?

Let us proceed to fit a logistic regression model in R:

```
> mat <- cbind(babyfood$disease, babyfood$nondisease)
> lrmod <- glm(mat~ sex + food, family=binomial, data=babyfood)
> summary(lrmod)
```

```
Call:
glm(formula = mat ~ sex + food, family = binomial, data = babyfood)

Deviance Residuals:
      1        2        3        4        5        6
 0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4
```

We expect the χ^2 approximation for the deviance to be accurate for this data (why?).

Consider the interpretation of the coefficient for breast feeding. We have

```
> exp(-0.6693)
```

```
[1] 0.5120669
```

Following the log-odds interpretation: breast feeding reduces the odd of respiratory disease by 51% of that for bottle feeding. We could compute the confidence interval on the odds scale. However, to get better coverage properties we compute the interval on the log-odds scale and then transforming the endpoints as follows:

```
> z <- qnorm(0.025,lower.tail=FALSE) # critical value
> z
```

```
[1] 1.959964
```

```
> exp(c(-0.669-z*0.153,-0.669+z*0.153))
```

```
[1] 0.3795099 0.6913386
```

3.7.2 Prospective and Retrospective Sampling

Prospective Sampling : In prospective sampling, the covariates are fixed and then the outcome is observed.

Retrospective Sampling : In retrospective sampling, the outcome is fixed and then the covariates are observed.

Consider a study in which the contraction of a disease is of interest. Let

- ϖ_0 be the probability that an individual is included in the study if they do not have the disease,
- ϖ_1 be the probability that an individual is included in the study if they do have the disease.

For a prospective study, $\varpi_0 = \varpi_1$ as we have no knowledge of the outcome. For a retrospective study, typically ϖ_1 is much greater than ϖ_0.

For a given covariate x, let

- $p^*(x)$ denote the conditional probability that an individual has the disease given inclusion in the study.
- $p(x)$ denote the unconditional probability that an individual has the disease, as we would obtain from a prospective study.

Then by Bayes Theorem:

$$p^*(x) = \frac{\varpi_1 p(x)}{\varpi_1 p(x) + \varpi_0(1 - p(x))}.$$

Rearranging yields

$$\text{logit}(p^*(x)) = \log\left(\frac{\varpi_1}{\varpi_0}\right) + \text{logit}(p(x)).$$

So the only difference between the retrospective and prospective study is the intercept term $\log(\varpi_1/\varpi_0)$.

- Generally, ϖ_1/ϖ_0 is unknown — meaning we cannot estimate β_0.
- However, knowledge of the other β can be used to assess the relative error of the covariates.

Now return to the respiratory disease example:

Prospective Sampling : In the infant respiratory example, we would select a sample of newborns whose parents had chosen a particular method of feeding and then monitor them for their first year.

Retrospective Sampling : In the infant respiratory example, typically, we would find infants visiting a doctor with a respiratory disease in the first year and then record their gender and method of feeding. We would also obtain a sample of respiratory disease-free infants. How these samples are collected is important — we require that the probability of inclusion in the study is independent of the predictor.

Suppose that the respiratory disease example had been a prospective study. Then, focussing on the boys only,

- Given the infant is breast fed, the log-odds of having a respiratory disease are $\log(47/447) = -2.25$.
- Given the infant is bottle fed, the log-odds of having a respiratory disease are $\log(77/381) = -1.60$.

The difference between these two log-odds, $\Delta = -1.60 - (-2.25) = 0.65$, represents the increased risk of respiratory disease incurred by bottle feeding relative to breast feeding. This is the log-odds ratio.

Now suppose that the respiratory disease example had been a retrospective study — we could compute the log-odds of feeding type given respiratory disease status and then find the difference. The log-odds ratio would be exactly the same:

$$\Delta = \log(77/47) - \log(381/447) = 0.65$$

This shows that a retrospective design is as effective as a prospective design for estimating Δ.**

Notes

- Retrospective designs are cheaper, faster and more efficient, so it is convenient that the same results may be obtained from a prospective study.
- Retrospective studies are typically less reliable than prospective studies - relies on historical records that may be incomplete or inaccurate.

**See Supplementary handout for further details.

Summary for using the logit link

Canonical choice for binomial response GLMs is the logit link. Using the logit link

- leads to simplier mathematics,
- easy interpretation using odds;
- allows for easy analysis of retrospective data.

3.8 Poisson Regression

So far we have examined the following cases for the response:

- Response is real $\rightarrow$ normal linear model.
- Response is a probability $\rightarrow$ Binomial regression.
- Response is a bounded integer $\rightarrow$ Binomial regression.

What if the response is an unbounded integer?

One possible approach is to use a Poisson distribution as the response. Let the response Y follow a Poisson distribution with mean $\lambda > 0$:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Examples where the response can be modelled as Poisson are:

- counting the occurrence of a rare event (i.e. a small probability of success).
- counting the number of events in a time interval.

Recall that the Poisson distribution is a member of the exponential family as its pmf can be written as

$$\exp\left\{y\log(\lambda) - \lambda - \log(y!)\right\},$$

where $\theta = \log(\lambda)$, $b(\theta) = \lambda = \exp(\theta)$, $a(\phi) = \phi = 1$ and $c(y, \phi) = -\log(y_i!)$. Therefore the canonical link is

$$\eta = \theta = \log(\lambda)$$

Using the canonical link, the likelihood, for independent $Y_1, \ldots, Y_n$ where $Y_i \sim$ Poisson(λ_i), is:

$$L(\boldsymbol{\beta}; \boldsymbol{y}) =$$

and the log-likelihood is:

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) =$$

The discrepancy of the model can be measured using the deviance, which is

$$D = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\widehat{\lambda}_i) - (y_i - \widehat{\lambda}_i)\right)$$

We can use the deviance to

- judge the goodness of fit of the model using the χ^2_{n-p} approximation.
- compare two nested models.

Alternatively, we can use Pearson's X^2 statistic as a goodness of fit measure, which for a Poisson response distribution, takes the form:

$$X^2 = \sum_{i=1}^{n}\frac{\left(y_i - \widehat{\lambda}_i\right)^2}{\widehat{\lambda}_i}$$

Note

A key property of the Poisson distribution is that its mean is equal to its variance.

3.8.1 Worked Example

In `R` fit the linear model using the following command:

```
> mylm1 <- lm(interlocks~assets+sector+nation,data=Ornstein)
```

The normal linear model fitted has response vector and design matrix:

$$\boldsymbol{Y} = \qquad\qquad , \quad X =$$

The `Ornstein` dataset contains 248 rows of data. The observations are the 248 largest Canadian firms specifying their `assets` in millions of dollars, the `sector` the firm belongs to, the `nation` that controls the firm and the number of `interlock`ing director and executive positions shared with other firms.

Can we use a companys' `assets`, `sector` and `nation` values to measure the number of interlocking positions?

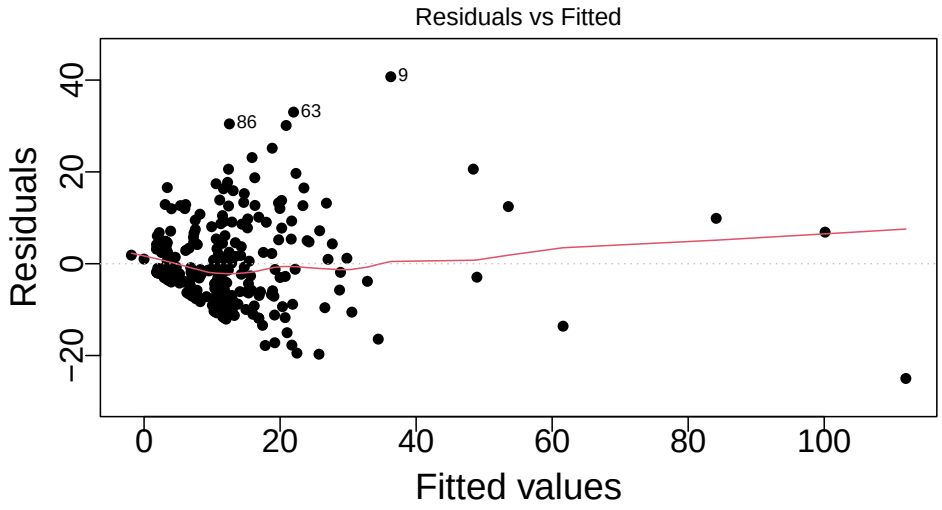
Let us take a look at the summary of linear model fitted in R:

```
> summary(mylm1)
```

```
Call:
lm(formula = interlocks ~ assets + sector + nation, data = Ornstein)

Residuals:
    Min      1Q  Median      3Q     Max
-25.001  -6.602  -1.629   4.780  40.728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.027e+01  1.561e+00   6.575 3.14e-10 ***
assets       8.096e-04  6.119e-05  13.231  < 2e-16 ***
sectorBNK   -1.781e+01  5.906e+00  -3.016  0.00284 **
sectorCON   -4.709e+00  4.728e+00  -0.996  0.32034
sectorFIN    5.153e+00  2.646e+00   1.948  0.05266 .
sectorHLD    8.777e-01  4.004e+00   0.219  0.82669
sectorMAN    1.149e+00  2.065e+00   0.556  0.57849
sectorMER    1.491e+00  2.636e+00   0.566  0.57206
sectorMIN    4.880e+00  2.067e+00   2.361  0.01905 *
sectorTRN    6.171e+00  2.760e+00   2.236  0.02629 *
sectorWOD    8.228e+00  2.679e+00   3.072  0.00238 **
nationOTH   -1.241e+00  2.695e+00  -0.461  0.64555
nationUK    -5.775e+00  2.674e+00  -2.159  0.03184 *
nationUS    -8.618e+00  1.496e+00  -5.760 2.64e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.827 on 234 degrees of freedom
Multiple R-squared:  0.6463,        Adjusted R-squared:  0.6267
F-statistic: 32.89 on 13 and 234 DF,  p-value: < 2.2e-16
```

What can we say about the model fit from this summary?

Lets take a look at a residual plot:



This plot shows signs of heteroscedasity. Perhaps using a GLM with a Poisson response distribution could explain the data better. We now proceed to fit the model with independent $Y_1, \ldots, Y_n$ with $Y_i \sim \text{Poisson}(\lambda_i)$ and

$$\text{E}(Y_i) = \exp(\beta_1 + \beta_2 \text{assets}_i + \beta_3 \text{sector}_i + \beta_4 \text{nation}_i), \quad i = 1, \ldots, n.$$

Therefore, the link function is log link (canonical link).

```
> myglm <- glm(interlocks~assets+sector+nation,
+              data=Ornstein,family=poisson)
> summary(myglm)
```

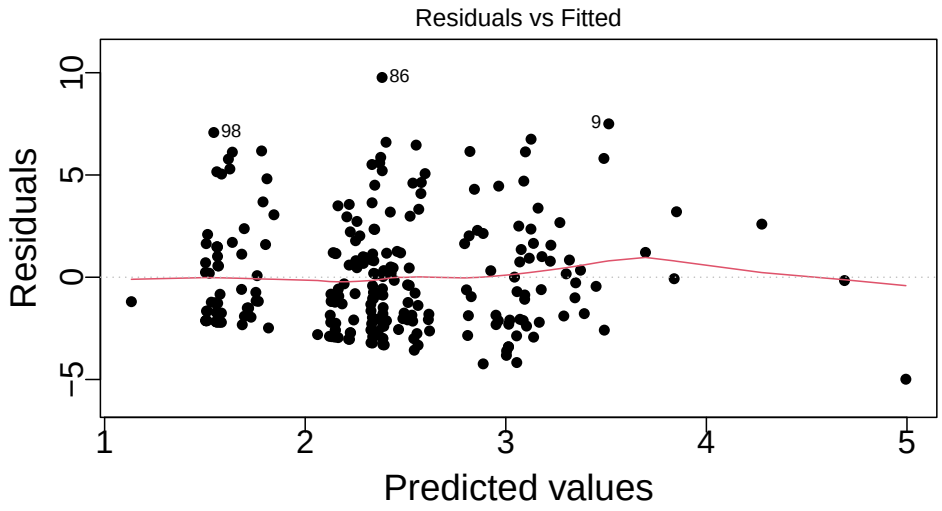
```
Call:
glm(formula = interlocks ~ assets + sector + nation,
        family = poisson, data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9908  -2.4767  -0.8582   1.3472   7.3610

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.325e+00  5.193e-02  44.762  < 2e-16 ***
assets       2.085e-05  1.202e-06  17.340  < 2e-16 ***
sectorBNK   -4.092e-01  1.560e-01  -2.623  0.00872 **
sectorCON   -6.196e-01  2.120e-01  -2.923  0.00347 **
sectorFIN    6.770e-01  6.879e-02   9.841  < 2e-16 ***
sectorHLD    2.085e-01  1.189e-01   1.754  0.07948 .
sectorMAN    5.260e-02  7.553e-02   0.696  0.48621
sectorMER    1.777e-01  8.654e-02   2.053  0.04006 *
sectorMIN    6.211e-01  6.690e-02   9.283  < 2e-16 ***
sectorTRN    6.778e-01  7.483e-02   9.059  < 2e-16 ***
sectorWOD    7.116e-01  7.532e-02   9.447  < 2e-16 ***
nationOTH   -1.632e-01  7.362e-02  -2.217  0.02663 *
nationUK    -5.771e-01  8.903e-02  -6.482 9.05e-11 ***
nationUS    -8.259e-01  4.897e-02 -16.867  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1887.4  on 234  degrees of freedom
AIC: 2813.4

Number of Fisher Scoring iterations: 5
```

Based on the (residual) deviance, we can perform a χ^2_{n-p} significance test:

```
> pchisq(deviance(myglm), df.residual(myglm), lower=FALSE)
```

```
[1] 5.793237e-256
```

This is an extremely small p-value indicating an ill-fitting model if the Poisson response distribution is correct. Why is this a poor fit? Let us look at the diagnostic plots.



This residual plot looks better than that for the linear model — there is hardly any biased. However, the residuals are quite large — larger than the Poisson distribution suggest. This is likely caused by taking a too simple structure for the covariates. Note that the problem lies with the standard errors not the estimates — the model can be used to make predictions, but not to make inference.

The Poisson distribution is restrictive in the sense that it has only one parameter, forcing the mean to equal the variance of the observations,which is not very flexible for fitting purposes. This problem can be alleviated by estimating ϕ, which then leads to better representative standard errors. To this end, we compute Pearson's dispersion estimate $\widehat{\phi}_p$:

```
> res_df <- df.residual(myglm)
> phihat <- sum(residuals(myglm,type="pearson")^2)/res_df
> phihat
```

```
[1] 7.943697
```

We can then "plug-in" this estimate into the model:

```
> summary(myglm, dispersion=phihat)
```

```
Call:
glm(formula = interlocks ~ assets + sector + nation,
        family = poisson, data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9908  -2.4767  -0.8582   1.3472   7.3610

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.325e+00  1.464e-01  15.882  < 2e-16 ***
assets       2.085e-05  3.389e-06   6.152 7.64e-10 ***
sectorBNK   -4.092e-01  4.397e-01  -0.931 0.352038
sectorCON   -6.196e-01  5.974e-01  -1.037 0.299703
sectorFIN    6.770e-01  1.939e-01   3.492 0.000480 ***
sectorHLD    2.085e-01  3.350e-01   0.622 0.533800
sectorMAN    5.260e-02  2.129e-01   0.247 0.804857
sectorMER    1.777e-01  2.439e-01   0.728 0.466323
sectorMIN    6.211e-01  1.886e-01   3.294 0.000989 ***
sectorTRN    6.778e-01  2.109e-01   3.214 0.001309 **
sectorWOD    7.116e-01  2.123e-01   3.352 0.000803 ***
nationOTH   -1.632e-01  2.075e-01  -0.787 0.431534
nationUK    -5.771e-01  2.509e-01  -2.300 0.021456 *
nationUS    -8.259e-01  1.380e-01  -5.984 2.17e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 7.943697)

    Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1887.4  on 234  degrees of freedom
AIC: 2813.4

Number of Fisher Scoring iterations: 5
```

Note that the estimation of β is independent of the dispersion parameter, ϕ, therefore modifying ϕ does not change the estimated coefficients. Further, notice, in this example, that now some of the coefficients have become non-significant.

3.8.2 Overdispersion

The term overdispersion means that the observed variance of the response is larger than the variation implied by the distribution used to fit the model. Overdispersion can be caused by several different problems — we state some below:

- Observations for different individuals with the same covariates do not have exactly the same distribution; that is, there are unaccounted for individual differences not included in the model.
- Observations may be correlated or clustered, while the specified variance function wrongly assumes uncorrelated data

One approach to mitigate the problem of overdispersion is to estimate the dispersion parameter, ϕ rather than assume $\phi = 1$ for the binomial and Poisson distributions. The procedure is to "plug-in" the estimated dispersion parameter $\widehat{\phi}$ into the analysis, as done in the worked example above. As mentioned above, estimating the dispersion parameter has no effect on the estimate of β but it inflates all their standard errors.

3.8.3 Estimation Problems

It can be the case that the `glm` function in `R` fails to convergence. Problems may arise due to problems with the Fisher scoring method or a "bad" initial starting point, however sometimes it is a problem with the data themselves, which is exhibited in the following example.

The following data set contains the values of androgen and estrogen (types of hormones) from 26 healthy males with their sexual orientation.

	androgen	estrogen	orientation		androgen	estrogen	orientation
1	3.9	1.8	s	14	3.9	3.9	g
2	4.0	2.3	s	15	3.4	3.6	g
3	3.8	2.3	s	16	2.3	2.5	g
4	3.9	2.5	s	17	1.6	1.7	g
5	2.9	1.3	s	18	2.5	2.9	g
6	3.2	1.7	s	19	3.4	4.0	g
7	4.6	3.4	s	20	1.6	1.9	g
8	4.3	3.1	s	21	4.3	5.3	g
9	3.1	1.8	s	22	2.0	2.7	g
10	2.7	1.5	s	23	1.8	3.6	g
11	2.3	1.4	s	24	2.2	4.1	g
12	2.5	2.1	g	25	3.1	5.2	g
13	1.6	1.1	g	26	1.3	4.0	g

Table 3.6: Hormone Dataset

Suppose we fit a binomial model to see if the orientation can be predicted from the hormone values.

```
> myglm <- glm(orientation~estrogen+androgen,
+              data=hormone,family=binomial)
```

Executing the above command gives the following warnings

```
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Lets take a look at the summary

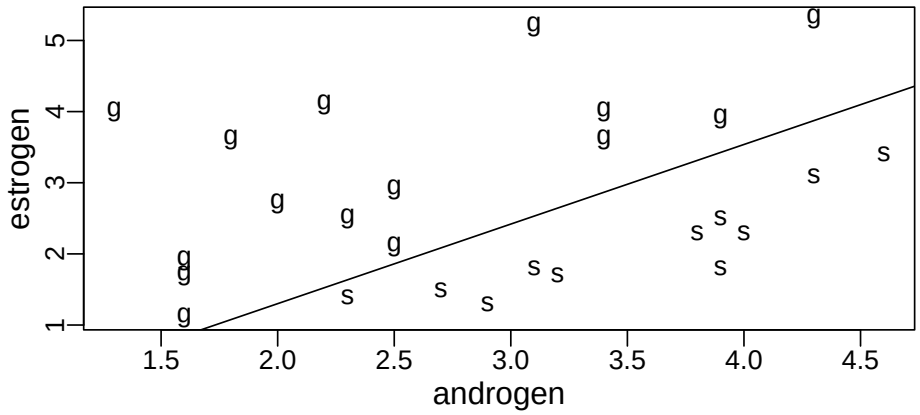
```
> summary(myglm)
```

```
Call:
glm(formula = orientation ~ estrogen + androgen,
        family = binomial, data = hormone)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.759e-05  -2.100e-08  -2.100e-08   2.100e-08   3.380e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -84.49  136095.03  -0.001    1.000
estrogen       -90.22   75910.98  -0.001    0.999
androgen       100.91   92755.62   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.5426e+01  on 25  degrees of freedom
Residual deviance: 2.3229e-09  on 23  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```

Notice

- the residual deviance is extraordinarily small indicating a good fit, yet none of the parameters are significant as they have very high standard errors.
- the default maximum number of iterations (25) has been reached.

The explanation for these results can be seen from a plot of the data.



The plot of the data reveals that the two classes of orientation are separable so that a perfect fit is possible.

The model fits the data perfectly - it is completely overfit. This results in unstable estimates of the parameters and their standard errors. Different data would likely lead to very different estimates, and predictions for unseen observations may well be poor.

Possible approaches to deal with these types of problems:

- Exact logistic regression.
- Bias reduction method — `R` package available `brlr`.

However, these methods are outside the scope of the course.

3.9 Quasi-Likelihood

Recall that a GLM is determined by the

-
-

We now discuss an approach that only requires specification of the link and variance functions of the model, but not the distribution of the response.

Motivation

Suppose we have the independent random variables $Y_1, \cdots, Y_n$. Let Y_i have mean μ_i and variance $\phi V(\mu_i)$. Define the score as

$$U_i = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

It follows that

$$\begin{aligned} \mathrm{E}(U_i) &= 0 \\ \mathrm{var}(U_i) &= \frac{1}{\phi V(\mu_i)} \\ \text{and } -\mathrm{E}\left(\frac{\partial U_i}{\partial \mu_i}\right) &= \frac{1}{\phi V(\mu_i)}. \end{aligned}$$

These properties are shared by the score function of members of the exponential family. This suggests that we may use U_i as a score. Hence

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt,$$

should behave like a log-likelihood function for μ_i (if the integral exists). We shall refer to Q_i as the log quasi-likelihood (or confusingly as just the quasi-likelihood) for μ_i. As we assume the observations are independent, the log quasi-likelihood for the complete data is just the sum of the components: $Q = \sum_{i=1}^{n} Q_i$.

Example Take $V(\mu) = 1$ and $\phi = \sigma^2$. Then

$$U =$$

$$\text{and } Q =$$

which is the same as the log-likelihood of a normal distribution up to constants. ■

In general, using variance functions associated with members of the exponential family recovers the log-likelihood. Further, other choices of $V(\mu)$ may not correspond to any known distribution or may even lead to something that is not a distribution.

Estimation

The estimation of $\boldsymbol{\beta}$ in the model is obtained by maximising the log quasi-likelihood, Q. We can again use the IWLS algorithm (Algorithm 3.1) . The only exception is now the dispersion parameter ϕ is estimated by

$$\widehat{\phi}_P = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)},$$

which is based on Pearson's X^2 statistic. We do not use the deviance estimator for ϕ as it is based on the likelihood — not reliable here.

Inference

By analogy, the quasi-deviance function for a single observation is [††]

$$D_i = -2\phi Q_i = 2 \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt,$$

and the total quasi-deviance is the sum over the D_i. This total quasi-deviance can be used like the ordinary deviance to perform inference on the model.

[††] see problem sheet

Example in R

`quasi-seeds.R  seeds-data.RData`

We continue with the seeds example presented in section 3.6.1. We can fit a corresponding quasi-binomial model as follows:

```
> my.quasi.bin <- glm(prop ~ seed * extract,
        family = quasibinomial(link = "logit"),
+     data = dat)
> summary(my.quasi.bin)
```

```
Call:
glm(formula = prop ~ seed * extract,
 family = quasibinomial(link = "logit"),
    data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.88834  -0.18458   0.01555   0.13622   0.38167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.5262     0.2764  -1.904  0.07398 .
seed          -0.2000     0.3969  -0.504  0.62079
extract        1.4479     0.3865   3.747  0.00161 **
seed:extract  -0.8478     0.5496  -1.543  0.14135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.08915264)

    Null deviance: 3.9112  on 20  degrees of freedom
Residual deviance: 1.8151  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Notice that the estimated dispersion parameter is not 1 (as in the previous analysis), but $\phi = 0.0892$, far less than 1. Further, the models deviance has been significantly reduced.

Comparison of multiple quasi models can be done using the F-test e.g.

```
> my.quasi.bin.2 <- glm(prop ~ seed + extract,
        family = quasibinomial,
```

```
+     data = dat)
> anova(my.quasi.bin.2, my.quasi.bin, test = "F")
```

```
Analysis of Deviance Table

Model 1: prop ~ seed + extract
Model 2: prop ~ seed * extract
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        18     2.0280
2        17     1.8151  1  0.21282 2.3871 0.1407
```

This F-value is not significant, therefore there is insufficient evidence to reject the null hypothesis, i.e. use the smaller model.

The other options for use of quasi-likelihoods in R are

```
quasi(link = "identity", variance = "constant")

quasibinomial(link = "logit")

quasipoisson(link = "log")
```

■

Note

- The dispersion, ϕ, can be modelled as a free parameter which is useful in modelling overdispersion or underdispersion.
- Although using the quasi-likelihood approach is attractive as it uses fewer assumptions — the quasi based estimators are generally less efficient than corresponding regular likelihood-based estimator — so if information about the distribution is available, use it.

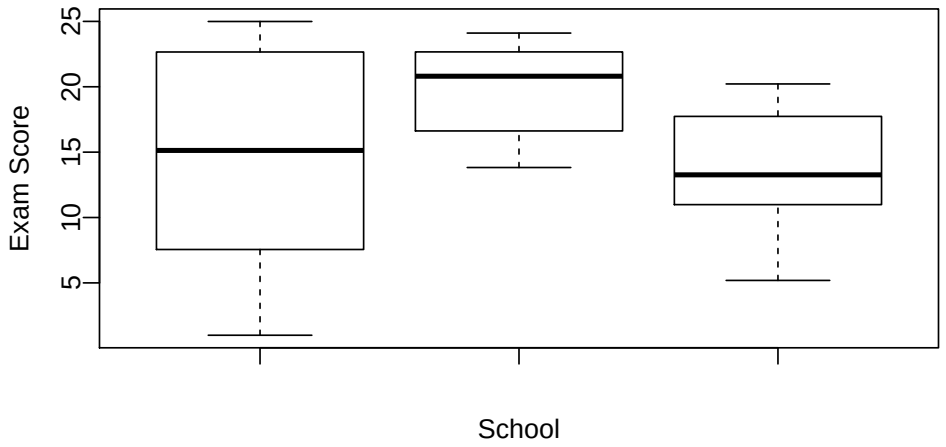
Chapter 4. Normal Linear Mixed Models

Models with mixed effects consist of both a fixed effect and a random effect. In the previous chapters we have only considered fixed effects; where the only source of randomness arises from considering the samples as independent and random samples. Random effects are used to model more complex correlation structures, such when there is more than one observation in a group. In this setting, we expect each group to vary independently, but measurements from the same group may be correlated.

We shall only consider random effects for normal linear models. However, random effects can be included in other, more complicated models e.g. non-linear, non-Normal (outside the scope of this course).

Motivating Example

Consider the dataset containing the math exam scores of students from 3 different schools. There are 10 scores from each school in this dataset. The data is presented in the boxplots below:



First, ignore the grouping structure of the data and assume the simple model:

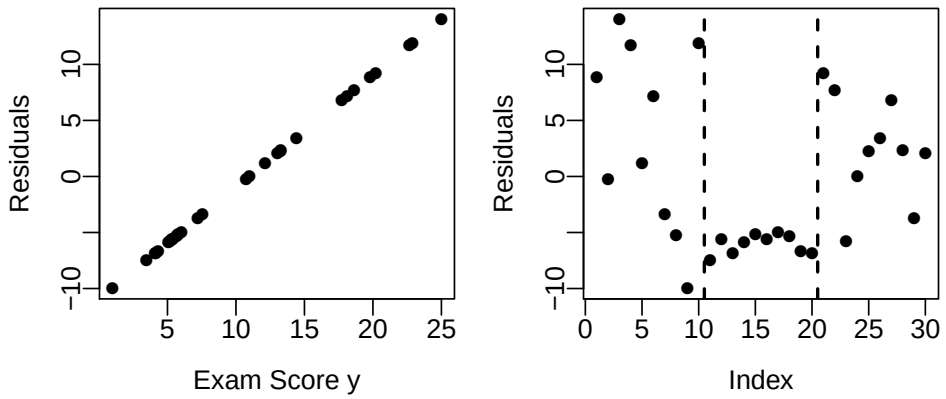
$$Y_{ij} = \mu + \epsilon_{ij}, \quad j = 1, \ldots m \quad i = 1, \ldots, K_j,$$

where Y_{ij} is the observed score for student i from school j, μ is the mean score across the population of students being sampled, and the ϵ are independent $N(0, \sigma^2)$ error terms.

```
> mylm <- lm(score~1,data=dat)
```

Let's look at the summary of this model and some plots of the residuals:

```
Call:
lm(formula = score ~ 1, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-9.966 -5.624 -1.810  5.941 14.031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.96       1.26   8.701  1.4e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.9 on 29 degrees of freedom
```



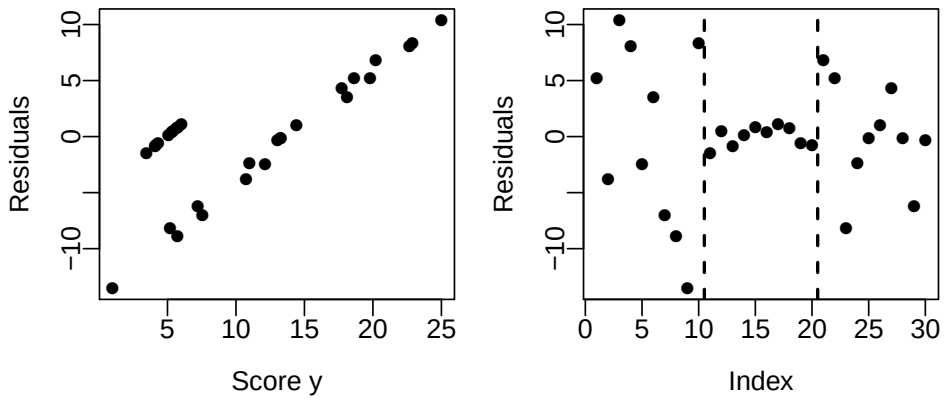
Notice that the group school effects are incoporated into the residuals — we see clear division in the residuals given by the different schools. This leads to an inflated estimate of variability i.e. the RSS.

One approach would be to incorporate the group effects, by introducing a seperate mean for each school:

```
> mylm2 <- lm(score~as.factor(school)+0,data=dat)
```

```
Call:
lm(formula = score ~ as.factor(school) + 0, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5675  -2.1747   0.0172   2.9415  10.4295

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
as.factor(school)140   14.564      1.751   8.318 6.30e-09 ***
as.factor(school)142    4.927      1.751   2.814  0.00902 **
as.factor(school)155   13.394      1.751   7.650 3.15e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.537 on 27 degrees of freedom
Multiple R-squared:  0.834,        Adjusted R-squared:  0.8155
F-statistic: 45.21 on 3 and 27 DF,  p-value: 1.162e-10
```

Notice that the residual standard error has been slightly reduced in comparison to the previous model. Let us inspect the residual plots again.



These residual plots still show a trend.

This model has several disadvantages:

- it only models the specific sample of students used in the experiment. We cannot say anything about the population of students from which the sample was drawn.
- it does not provide an estimate of the between-school variability.

We clearly need a class of model that incorporates the group structure of the data. Then we can pose interesting questions such as: what was the average maths exam score across the population, and how much variation were there between schools.

4.1 Specification of Normal Linear Mixed Models

Definition. The normal linear mixed model is defined as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon},$$

where

- $\boldsymbol{Y} \in \mathbb{R}^n$ is the response vector,
- $X \in \mathbb{R}^{n \times p}$ is the design matrix for the fixed effect,
- $Z \in \mathbb{R}^{n \times m}$ model matrix for the random effects,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter,
- $\boldsymbol{\nu}$ is an m-variate vector with $\boldsymbol{\nu} \sim N(\mathbf{0}, \sigma_\nu^2 I_m)$,
- $\boldsymbol{\epsilon}$ is an n-variate vector with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n)$.

and $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$ are independent.

The joint distribution of $\boldsymbol{Y}$ can be easily derived as follows: First,

$$\mathrm{E}(\boldsymbol{Y})$$

Next,

$$\mathrm{cov}(\boldsymbol{Y})$$

Since $\boldsymbol{Y}$ is a sum of normally distributed variables, it follows that

$$\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma_\epsilon^2(I_n + Z\Psi Z^T)), \quad \text{where} \quad \Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2} I_m$$

For compactness, we sometimes group the variances into one parameter $\boldsymbol{\tau} = (\sigma_\epsilon^2, \sigma_\nu^2)^T$, called the variance-components of the model.

Example A simple normal linear model is

$$Y_{ij} = \mu + \nu_j + \epsilon_{ij}, \quad j = 1, \dots m, \; i = 1, \dots, K_j,$$

where, as before, μ is the fixed effect, while $\nu_j \sim N(0, \sigma_\nu^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. Casting this model into the definition above:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{K_1,1} \\ Y_{1,2} \\ \vdots \\ Y_{K_2,2} \\ \vdots \end{pmatrix}, \quad X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

$$\boldsymbol{\beta} = \mu, \quad \boldsymbol{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_m \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{K_1,1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{K_2,2} \\ \vdots \end{pmatrix},$$

where $n = \sum_{j=1}^{m} K_j$, $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times 1}$, $Z \in \mathbb{R}^{n \times m}$, $\boldsymbol{\beta} \in \mathbb{R}$. This model is called the one-way random effects model. ■

4.2 Estimation

Estimation for models with random effects is not as straightforward as it was for fixed effects models. The standard methods of estimation for normal linear mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML). We shall discuss both methods in this section.

4.2.1 Estimation of $\boldsymbol{\beta}$

Under a normal linear mixed model, the distribution of $\boldsymbol{Y}$ is

$$\frac{1}{(2\pi)^{n/2}|\sigma_\epsilon^2 V|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y}-X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})\right\},$$

where $V := I_n + Z\Psi Z^T$. Thus, the log-likelihood function is given by

$$\ell(\boldsymbol{\beta},\boldsymbol{\tau};\boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 V\right| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y}-X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y}-X\boldsymbol{\beta}),$$

where the variance-covariance term $\boldsymbol{\tau}$ appears in V. By differentiating the log-likelihood with respect to $\boldsymbol{\beta}$, we obtain

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \tag{4.1}$$

The standard approach to find the MLEs is to solve $\frac{\partial \ell}{\partial \boldsymbol{\beta}} = 0$ (and $\frac{\partial \ell}{\partial \boldsymbol{\tau}} = 0$).

For simplicity, assume that X has full rank; so that $\text{rank}(X) = p$. Let $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}})$ be the MLE. From (4.1) we obtain

$$\widehat{\boldsymbol{\beta}} = \tag{4.2}$$

where $V_{\widehat{\boldsymbol{\tau}}} = V(\widehat{\boldsymbol{\tau}})$, that is, V with the MLE $\widehat{\boldsymbol{\tau}}$ plugged in. Thus once the MLE of $\boldsymbol{\tau}$ is found, the MLE of $\boldsymbol{\beta}$ can be calculated by expression (4.2). Thus one procedure is to first solve for $\widehat{\boldsymbol{\tau}}$, then compute $\widehat{\boldsymbol{\beta}}$ using (4.2). However, finding $\widehat{\boldsymbol{\tau}}$ is not so simple.

4.2.2 Estimation of Variance Components

There are several approaches to estimating the variance components, σ_ϵ^2 and σ_ν^2, of normal linear mixed models. The first approach we discuss is to use the analysis of variances:

4.2.2.1 ANOVA estimators

Let's start with the simplest possible random effects model to gain intuition:

$$Y_{ij} = \mu + \nu_j + \epsilon_{ij}, \quad j = 1, \ldots, m \quad i = 1, \ldots, K_j.$$

where ν_j is a random effect with $\mathrm{E}(\nu_j) = 0$, $\mathrm{var}(\nu_j) = \sigma_\nu^2$, and $\mathrm{E}(\epsilon_{ij}) = 0$, $\mathrm{var}(\epsilon_{ij}) = \sigma_\epsilon^2$. This induces a correlation between the observations in the same group

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2}.$$

Proof.

□

The correlation, ρ, is known as the intraclass correlation coefficient (ICC). Note

- if $\sigma_\nu^2 = 0$ then $\rho = 0$.
- if $\sigma_\nu^2 \gg \sigma_\epsilon^2$ then $\rho \approx 1$.

For simplicity, suppose there are an equal number of observations for each of the m groups i.e. set $K_j = K$ — referred to as a balanced design. We can decompose the variation as follows:

$$\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_{\bullet j})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m}(\overline{Y}_{\bullet j} - \overline{Y})^2$$

where

$$\overline{Y}_{\bullet j} = \frac{1}{K}\sum_{i=1}^{K} Y_{ij} \quad \text{and} \quad \overline{Y} = \frac{1}{mK}\sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij}.$$

This decomposition can be written as: $SST = SSE + SSA$ where SSE is the residual sum of squares, SST is the total sum of squares (corrected for the mean) and SSA is the sum of square due to $\boldsymbol{\nu}$. It turns out that (see Lemmas below):

$$\mathrm{E}(SSE) = m(K-1)\sigma_\epsilon^2, \quad \mathrm{E}(SSA) = (m-1)(K\sigma_\nu^2 + \sigma_\epsilon^2)$$

which suggests using the estimators:

$$\widehat{\sigma}_\epsilon^2 = \frac{SSE}{m(K-1)} =: MSE, \quad \widehat{\sigma}_\nu^2 = \frac{SSA/(m-1) - \widehat{\sigma}_\epsilon^2}{K} =: \frac{MSA - MSE}{K},$$

where $MSA = SSA/(m-1)$. These estimators, $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_\nu^2$, are known as the ANOVA estimators.

Lemma 7.

$$SST = SSE + SSA$$

Proof.

□

Lemma 8. The expectations of these sums of squares are:

i) $\mathrm{E}(SSE) = m(K-1)\sigma_\epsilon^2$

ii) $\mathrm{E}(SSA) = (m-1)(K\sigma_v^2 + \sigma_\epsilon^2)$

iii) $\mathrm{E}(SST) = (mK-1)\sigma_\epsilon^2 + K(m-1)\sigma_v^2$

Proof. i) For SSE, the expectation of the summand is

$$\mathrm{E}\left\{\left(Y_{ij} - \overline{Y}_{\bullet j}\right)^2\right\}$$

ii) For $\mathrm{E}(SSA)$, again start by looking at the expectation of the summand

$$\mathrm{E}\left\{\left(\overline{Y}_{\bullet j} - \overline{Y}\right)^2\right\}$$

iii) Plugging in the results for i) and ii) yields:

$$\begin{aligned}\mathrm{E}(SST) =& m(K-1)\sigma_\epsilon^2 + (m-1)(K\sigma_\nu^2 + \sigma_\epsilon^2) \\ =& (mK-1)\sigma_\epsilon^2 + K(m-1)\sigma_\nu^2\end{aligned}$$

□

These were the first estimators developed for the variance components. They have an explicit form suitable for hand calculations, which was important in precomputing days. However, they have a number of disadvantages:

1. The estimates can take negative values. For instance, if $MSA < MSE$ then $\sigma_\nu^2 < 0$.

2. A balanced design is assumed — where the number of observations in each group is equal. In such settings, the ANOVA decomposition into sum of squares is unique. For unbalanced data, this is not true and therefore we must choose which ANOVA decomposition and therefore the ANOVA estimator to use.

3. The need for complicated algebraic calculations. Formulae for the simpler models are easy to derive and code — but more complex models will require difficult constructions.

4.2.2.2 Maximum Likelihood Approach

Recall that the normal linear mixed model can be written as

$$\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma_\epsilon^2 \underbrace{(I_n + Z\Psi Z^T)}_{=:V}), \quad \text{where} \quad \Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2} I_m.$$

If Ψ is known (i.e. the variance components), we can estimate $\boldsymbol{\beta}$ using generalised least squares. However, the estimation of the variance components is often the purpose of the analysis. Standard maximum likelihood is one method of estimation that we can use. If we let $V = I_n + Z\Psi Z^T$, then the distribution of $\boldsymbol{Y}$ is

$$\frac{1}{(2\pi)^{(n/2)} \left|\sigma_\epsilon^2 V\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\}.$$

Thus the log-likelihood takes the form

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 V\right| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}), \tag{4.3}$$

where the variance-covariance term $\boldsymbol{\tau}$ appears in V.

We can then proceed to optimise to find the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. This is straightforward in principle, but difficult in practice. More complex models involving larger numbers of random effects parameters can be difficult to estimate. One particular problem is that the variance cannot be negative so the MLE for the variance might be zero. Nevertheless, we rely on numerical techniques to find the MLEs.* A problem with MLEs is that they can be biased — as illustrated in the following example.

*the optimisation techniques are not examinable.

Example (Neyman-Scott)

Consider the model

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

for $i = 1, 2$ and $j = 1, \ldots, m$. We are interested in estimating σ^2.

The joint pdf of $\boldsymbol{Y}$ is

$$\prod_{j=1}^{m} \left[\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp\left\{ -\frac{1}{2\sigma^2} \left(Y_{1j} - \mu_j \right)^2 \right\} \exp\left\{ -\frac{1}{2\sigma^2} \left(Y_{2j} - \mu_j \right)^2 \right\} \right].$$

Hence the log-likelihood is

The derivative with respect to μ_j is

for $j = 1, \ldots, m$. Consequently, their MLEs are

$$\widehat{\mu}_j =$$

Similarly we find

$$\frac{\partial \ell}{\partial \sigma^2} =$$

Thus the MLE of σ^2 is

$$\widehat{\sigma}^2 =$$

We then proceed to plug in the MLEs for μ_j into the formula for $\widehat{\sigma}^2$ to get

$$\widehat{\sigma}^2 =$$

Now consider the bias of this estimator for σ^2.

$$\mathrm{E}(\widehat{\sigma}^2) = \qquad .$$

Since $4s_j^2 = \left\{(Y_{2j} - \mu_j) - (Y_{1j} - \mu_j)\right\}^2$, it follows from independence that

$$4\mathrm{E}(s_j^2) =$$

and so

$$\mathrm{E}(s_j^2) =$$

Finally,

$$\mathrm{E}(\widehat{\sigma}^2) =$$

which shows that $\widehat{\sigma}^2$ is biased and does not even become unbiased as $n = 2m \to \infty$. Here, the maximum likelihood estimator is not consistent. ■

4.2.2.3 Restricted Maximum Likelihood

The Neyman-Scott example demonstrates that the MLE approach can lead to a biased estimator of the variance. Notice that the fixed effects, the μ_is, are considered to be nuisance parameters and the main interest lies in the variance term σ^2. We now discuss a method that can be used to estimate the parameters of interest without dealing with nuisance parameters.

Example (Neyman-Scott continued)

Consider the transformation $B_j = Y_{1j} - Y_{2j}$. Then $B_1, \ldots, B_m$ are independent and $N(0, 2\sigma^2)$ distributed. Notice that this transformation leads to a distribution that does not involves the nuisance parameters $\mu_1, \ldots, \mu_m$. Then the joint pdf of $B_1, \ldots, B_m$ is

$$f(b_1, \ldots, b_m) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(2\sigma^2)}} \exp\left\{-\frac{1}{2(2\sigma^2)} b_j^2\right\}.$$

Hence the log-likelihood is

and it follows that

$$\frac{\partial \ell}{\partial \sigma^2} =$$

Therefore, the MLE of σ^2 is

$$\widehat{\sigma}^2 =$$

Finally, we can compute the expectation of this MLE and find that is it indeed an unbiased estimator of σ^2:

$$\mathrm{E}(\widehat{\sigma}^2) =$$

■

The trick shown in this example is to apply a transformation to the data to eliminate the fixed effects; then use to transformed data to estimate the variance component. This idea can be applied in general.

Assume (w.l.o.g.) that X has full rank i.e. $\mathrm{rank}(X) = p$. Let $L \in \mathbb{R}^{n \times (n-p)}$ such that $\mathrm{rank}(L) = n - p$ and

$$L^T X = 0.$$

Then multiplying the mixed normal linear model equation by L^T on the left we get

$$\underbrace{L^T \boldsymbol{Y}}_{=:\boldsymbol{B}} =$$

After the transformation we have

$$\boldsymbol{B} \sim N(0, \sigma_\epsilon^2 L^T V L).$$

It follows that the joint pdf of $\boldsymbol{B}$ is

$$\frac{1}{(2\pi)^{(n-p)/2} \left|\sigma_\epsilon^2 L^T V L\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2} \boldsymbol{b}^T (L^T V L)^{-1} \boldsymbol{b}\right\}$$

so that the log-likelihood is

$$\ell_R(\boldsymbol{\tau}; \boldsymbol{b}) = -\frac{(n-p)}{2} \log(2\pi) - \frac{1}{2} \log\left|\sigma_\epsilon^2 L^T V L\right| - \frac{1}{2\sigma_\epsilon^2} \boldsymbol{b}^T (L^T V L)^{-1} \boldsymbol{b}.$$

To emphasis this is the restricted log-likelihood we use the subscript R. The REML (**re**stricted **m**aximum **l**ikelihood) estimator of $\boldsymbol{\tau}$ is defined as the maximiser of ℓ_R. As with ordinary MLE; such a maximiser satifies the REML equation

$$\frac{\partial \ell_R}{\partial \boldsymbol{\tau}} = \mathbf{0}.$$

Again, we shall rely on numerical methods to find the maximiser of the REML equation. After the random effects are estimated; the fixed effects are subsequently estimated.

Notice that we have already seen a candidate for L. Let $L^T = I_n - X(X^T X)^{-1} X^T$, then[†]

$$L^T \boldsymbol{Y} = L^T X \boldsymbol{\beta} + L^T Z \boldsymbol{\nu} + L^T \boldsymbol{\epsilon}$$

[†]see problem sheet

4.2.3 Worked Example

REM-pulp.R

We now illustrate the fitting methods used in R using some data from an experiment to test the paper brightness depending on a shift operator:

```
> library("faraway")
> data("pulp")
```

	bright	operator		bright	operator
1	59.8	a	11	60.7	c
2	60.0	a	12	60.7	c
3	60.8	a	13	60.5	c
4	60.8	a	14	60.9	c
5	59.8	a	15	60.3	c
6	59.8	b	16	61.0	d
7	60.2	b	17	60.8	d
8	60.4	b	18	60.6	d
9	59.9	b	19	60.5	d
10	60.0	b	20	60.5	d

Table 4.1: Pulp Dataset

We start by calculating the ANOVA estimators:

```
> myaov <- aov(bright~operator,data=pulp)
> summary(myaov)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
operator     3   1.34  0.4467   4.204 0.0226 *
Residuals   16   1.70  0.1062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> coef(myaov)
```

```
(Intercept)   operatorb   operatorc   operatord
      60.24       -0.18        0.38        0.44
```

The estimate of $\sigma_\epsilon^2 = 0.1062$ and of σ_ν^2 is

```
> (0.4467-0.1062)/5
```

```
[1] 0.0681
```

We now demonstrate the maximum likelihood estimators.

```
> library("lme4")
> mylme <- lmer(bright~1+(1|operator),data=pulp)
```

Take note of how the model is specified in the `lmer` function: the first part corresponds to the fixed effects; the second part corresponds to the random effect, where `(1|operator)` indicates that the data are grouped by `operator` and the `1` indicates that the random effect is constant within each group.

```
> summary(mylme)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: bright ~ 1 + (1 | operator)
   Data: pulp

REML criterion at convergence: 18.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4666 -0.7595 -0.1244  0.6281  1.6012

Random effects:
 Groups   Name        Variance Std.Dev.
 operator (Intercept) 0.06808  0.2609
 Residual             0.10625  0.3260
Number of obs: 20, groups:  operator, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.4000     0.1494   404.2
```

Also note that the default fitting method is REML. The reported standard deviations are just the square root of the reported variances — not estimates of the uncertainty in the variances.

The maximum likelihood estimates can also be computed.

```
> mylme2 <- lmer(bright~1+(1|operator),data=pulp,REML=FALSE)
> summary(mylme2)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: bright ~ 1 + (1 | operator)
   Data: pulp

     AIC      BIC   logLik deviance df.resid
    22.5     25.5     -8.3     16.5       17

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.50554 -0.78116 -0.06353  0.65850  1.56232

Random effects:
 Groups   Name        Variance Std.Dev.
 operator (Intercept) 0.04575  0.2139
 Residual             0.10625  0.3260
Number of obs: 20, groups:  operator, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.4000     0.1294   466.7
```

4.3 Inference

To compare two nested models M_0 (smaller) and M_1, we compute the difference of the log-likelihoods (or difference of deviances)

$$2\left\{\ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\tau}}_1; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\tau}}_0; \boldsymbol{y})\right\},$$

where $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\tau}}_0$ are the MLEs of the parameters for M_0 and $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\tau}}_1$ are the MLEs of the parameters for M_1.

We have two choices: either (wrongly) assume the variance parameters are equal to their estimated values, then perform a χ^2-test, or try to take into account their uncertainty, then use an F-test.

Testing whether to include the random effects or not involves the hypothesis:

$$H_0 : \ \sigma_\nu^2 = 0.$$

The standard derivation of the asymptotic χ^2 distribution depends on the null hypothesis lying on the interior of the parameter space. This assumption does not hold for this test. Consequently, one may resort to sampling procedures – see parametric bootstrap below (Section 4.3.1)

If the χ^2 approximation is used, then the test will tend to be conservative in the sense that the p-values will tend to be larger than they should be.

inference in the mixed model is intricate, and often controversial

Warning

We cannot use the REML estimation method if the deviance test is to be used. The reason is that; the REML estimates the random effects by eliminating the fixed effects (we transformed using L). If these fixed effects are changed, the likelihoods of the two models are not directly comparable. Therefore, we should use the maximum likelihood estimation method in this situation.

4.3.1 Worked Example (Continued) — Parametric Bootstrap

`REM-pulp-cont.R`

Returning to the `pulp` example. We fitted the mixed effects model using the maximum likelihood (ML) approach

```
> mylme2 <- lmer(bright~1+(1|operator),data=pulp,REML=FALSE)
```

We now consider a test concerning the random effects — therefore we must use ML. Say we are interested in comparing the `mylme2` and

```
> mylm <- lm(bright~1,data=pulp)
```

that is, comparing if the random effects should be included or not. Our null model is the smaller linear model `mylm`. Note that as there are no random effects in this model, we must use the `lm` function. For once, we cannot use `anova` and therefore must compute things directly:

```
> d <- as.numeric(2*(logLik(mylme2)-logLik(mylm)))
> d
```

```
[1] 2.568371
```

```
> pchisq(d,1,lower=FALSE)
```

```
[1] 0.1090199
```

This p-value is now well above the 5% significance level. We cannot say that this result is necessarily wrong, but the use of the χ^2 approximation does cause us to doubt the result.

We can use the parametric bootstrap approach to obtain a more accuate p-value. We need to estimate the probability, given that the null hypothesis is true, of observing a `d` value of 2.568371 or greater. Under the null hypothesis, $Y \sim N(\mu, \sigma^2)$. A simulation approach would proceed to generate data under this model, fit the null and alternative models and then compute the deviance. The process would be repeated a large number of times and the proportion of deviance values exceeding the observed value of 2.568371 would be used to estimate the p-value. In practice, we do not know the true values of μ and σ^2 but we can use the estimated values — this distinguishes the parametric approach from the simulation approach. We can simulate responses under the null as follows:

```
> y <- simulate(mylm)
```

Now taking the generated data, fit both the null and alternative model compute the deviance and repeat, say 1000 times:

```
> ds <- numeric(1000)
> for(i in 1:1000){
+     y <- unlist(simulate(mylm))
+     nullmod <- lm(y~1)
+     altmod <- lmer(y~1+(1|operator),data=pulp,REM=FALSE)
+     ds[i] <- as.numeric(2*(logLik(altmod)-logLik(nullmod)))
+ }
```

We may examine the distribution of the bootstrapped deviance values i.e. look at the proportion of values close to zero

```
> mean(ds<0.00001)
```

```
[1] 0.698
```

This clearly indicates that the deviance in this case does not have the χ^2 distribution under the null hypothesis.
Our estimated p-value is:

```
> phat <- mean(ds>2.568371)
> phat
```

```
[1] 0.02
```

We should compute the standard error for this estimate (where does this come from?)

```
> sqrt(0.02*0.98/1000)
```

```
[1] 0.004427189
```

So we can be fairly sure it is under 5%. If in doubt, do more replications.

4.4 Prediction

4.4.1 Best prediction when all the parameters are known

When the fixed effects, $\boldsymbol{\beta}$, and variance components, $\boldsymbol{\tau}$, are known, the best predictor for $\boldsymbol{\nu}$, in the sense of minimum mean squared error, is its conditional expectation given the data; that is,

$$\widetilde{\boldsymbol{\nu}} := \mathrm{E}(\boldsymbol{\nu}|\boldsymbol{y}) = \Psi Z^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}).$$

To show $\mathrm{E}(\boldsymbol{\nu}|\boldsymbol{y}) = \Psi Z^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})$, first consider the covariance between $\boldsymbol{Y}$ and $\boldsymbol{\nu}$:

.

The proof then follows by considering the joint over $\boldsymbol{Y}$ and $\boldsymbol{\nu}$ which is an $(n+m)$-dimensional multivariate Gaussian distribution:

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{\nu} \end{pmatrix} \sim N\left(\begin{pmatrix} X\boldsymbol{\beta} \\ 0 \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} I_n + Z\Psi Z^T & Z\Psi \\ \Psi Z^T & \Psi \end{pmatrix}\right)$$

where we also used that $\Psi^T = \Psi$. Now recall the result:

$$\begin{pmatrix} \boldsymbol{W}_a \\ \boldsymbol{W}_b \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right)$$

$$\Rightarrow \ \boldsymbol{W}_b \mid \boldsymbol{W}_a = \boldsymbol{w}_a \sim N\left(\boldsymbol{\mu}_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\boldsymbol{w}_a - \boldsymbol{\mu}_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right).$$

This allows us to obtain the conditional $\boldsymbol{\nu}|\boldsymbol{y}$ (i.e. the posterior) without performing an intergration:

$$\widetilde{\boldsymbol{\nu}} := \mathrm{E}(\boldsymbol{\nu} \mid \boldsymbol{Y} = \boldsymbol{y}) =$$

Once the best predictors $\widetilde{\boldsymbol{\nu}}$ are obtained, the best predictor of $\eta = x_\star^T\boldsymbol{\beta} + z_\star^T\boldsymbol{\nu}$ is

$$\widehat{\eta}_{\mathrm{B}} := x_\star^T\boldsymbol{\beta} + z_\star^T\widetilde{\boldsymbol{\nu}} = x_\star^T\boldsymbol{\beta} + z_\star^T\Psi Z^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}).$$

The subscript B refers to the best predictor.

4.4.2 Best linear unbiased prediction (BLUP)

If the fixed effects, $\boldsymbol{\beta}$, are unknown but the variance components, $\boldsymbol{\tau}$, are known, $\widehat{\eta}_{\mathrm{B}}$ is not a predictor. In this case, it is customary to replace $\boldsymbol{\beta}$ by is MLE $\widehat{\boldsymbol{\beta}}$, which is

$$\widehat{\boldsymbol{\beta}} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{Y}.$$

After replacing $\boldsymbol{\beta}$ by its MLE, the predictor for $\eta := x_\star^T \boldsymbol{\beta} + z_\star^T \boldsymbol{\nu}$ becomes

$$\widehat{\eta}_{\mathrm{BLUP}} = x_\star^T \widehat{\boldsymbol{\beta}} + z_\star^T \widetilde{\boldsymbol{\nu}}_{\widehat{\boldsymbol{\beta}}} = x_\star^T \widehat{\boldsymbol{\beta}} + z_\star^T \Psi Z^T V^{-1} (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}).$$

This is the best linear unbiased predictor (BLUP) of η, in the sense that (i) is it linear in $\boldsymbol{y}$, (ii) its expected value is equal to η and (iii) it minimises the MSE among all linear unbiased predictors. The vector

$$\widetilde{\boldsymbol{\nu}}_{\widehat{\boldsymbol{\beta}}} := \Psi Z^T V^{-1} (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})$$

is also called the BLUP of $\boldsymbol{\nu}$.

4.4.3 Empirical BLUP

In practice, the fixed effects and variance components are typically unknown. Therefore, in most cases neither the best predictor nor the BLUP is computable. In such cases, one replaces the vector of variance components, $\boldsymbol{\tau}$, which is involved in the expression of BLUP, by, $\widehat{\boldsymbol{\tau}}$. That is, instead of predicting the random effects by $\widetilde{\boldsymbol{\nu}}$ we use

$$\widehat{\boldsymbol{\nu}} := \mathrm{E}(\boldsymbol{\nu} | \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}) = \Psi Z^T V_{\widehat{\boldsymbol{\tau}}}^{-1} (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})$$

where $V_{\widehat{\boldsymbol{\tau}}}$ is V with an estimator of $\boldsymbol{\tau}$ plugged in. The resulting predictor, often called empirical BLUP or eBLUP, is

$$\widehat{\eta}_{\mathrm{eBLUP}} = x_\star^T \widehat{\boldsymbol{\beta}} + z_\star^T \widehat{\boldsymbol{\nu}} = x_\star^T \widehat{\boldsymbol{\beta}} + z_\star^T \Psi Z^T V_{\widehat{\boldsymbol{\tau}}}^{-1} (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}).$$

Note

Recall that the random effects, $\boldsymbol{\nu}$, are not parameters: they are **random variables**. So rather than estimating their values, we **predict** them. Therefore, the conditional expectations are referred to as predictors.

4.5 Worked Example

REM-sim.R

We begin by simulating some data as follows:

```
> n <- 60
> X <- rnorm(n,0,1)
> Z <- c(rep(1,20),rep(2,20),rep(3,20))
> nu <- rnorm(3,0,1)
> beta <- 1.5
> y <- nu[Z]+beta*X+rnorm(60,0,0.1)
> dat <- data.frame(y=y,Z=Z,X=X)
```

We have simulated from 3 groups with an associated random effect with standard deviation 0.1 and additional covariate X, representing a fixed effect with regression coefficient 1.5 with standard deviation 1.

Next , we fit the normal linear mixed model, $\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}$, as follows

```
> library("lme4")
> mylme <- lmer(y~X+(1|Z)+0,data=dat,REML=FALSE); summary(mylme)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: y ~ X + (1 | Z) + 0
   Data: dat

     AIC      BIC   logLik deviance df.resid
   -71.4    -65.1     38.7    -77.4       57

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.58247 -0.59496 -0.00208  0.49970  2.02786

Random effects:
 Groups   Name        Variance Std.Dev.
 Z        (Intercept) 0.77931  0.8828
 Residual             0.01123  0.1060
Number of obs: 60, groups:  Z, 3

Fixed effects:
  Estimate Std. Error t value
X  1.51026    0.01572   96.09
```

We see that the output includes the standard effects summary and an additional summary for the random effects; reporting the variance of the random effect (named as `Intercept`) and the residual variance. Notice that the estimated residual variance is accurate (true value is 0.01), but the estimated random effect variance is not as accurate (true value is 1). Why? The estimated coefficient, β, is also accurate (true value 1.5).

To get the posterior estimate for the random effects i.e. $\widehat{\boldsymbol{\nu}} := \mathrm{E}(\boldsymbol{\nu}|\boldsymbol{y}, \widehat{\boldsymbol{\tau}})$:

```
> ranef(mylme)
```

```
$Z
  (Intercept)
1   0.8867433
2   1.1064969
3  -0.5706097
```

If we wish to predict the response of an observation given the covariate $x_\star = 1$, but do not know which group it belongs to, then the best predictor is $x_\star^T \widehat{\beta} = \widehat{\beta}$; for this example this is 1.51026.

If we want to predict the response of an observation from, say, group $j = 1$, then we use the estimate of the random effect to get the estimate

```
> fixef(mylme)+ranef(mylme)$Z
```

```
  (Intercept)
1   2.3970000
2   2.6167536
3   0.9396471
```

4.6 Inference Continued

4.6.1 Inference about the fixed effects, β

Let us return to the un-restricted log-likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 V_{\boldsymbol{\tau}}\right| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V_{\boldsymbol{\tau}}^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}). \tag{4.4}$$

For fixed $\boldsymbol{\tau}$, taking derivative of this log-likelihood with repsect to $\boldsymbol{\beta}$ gives

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = X^T V_{\boldsymbol{\tau}}^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})$$

so that the MLE of $\boldsymbol{\beta}$ is the solution of

$$X^T V_{\boldsymbol{\tau}}^{-1} X\widehat{\boldsymbol{\beta}} = X^T V_{\boldsymbol{\tau}}^{-1}\boldsymbol{y},$$

which we recognise. Recall that $V_{\boldsymbol{\tau}} := (I_n + Z\Psi Z^T)$ where $\Psi = \frac{\sigma_v^2}{\sigma_\epsilon^2} I_m$ — therefore $V_{\boldsymbol{\tau}}$ depends on the variance components $\boldsymbol{\tau} = (\sigma_\epsilon^2, \sigma_v^2)^T$. The (profile) log-likelihood of the variance parameter $\boldsymbol{\tau}$ is

$$\ell(\boldsymbol{\tau}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 V_{\boldsymbol{\tau}}\right| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})^T V_{\boldsymbol{\tau}}^{-1}(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}).$$

and the Fisher information of $\boldsymbol{\beta}$ is

$$\mathcal{J}(\boldsymbol{\beta}) = X^T V_{\boldsymbol{\tau}}^{-1} X.$$

In practice, the estimated value of $\widehat{\boldsymbol{\tau}}$ is plugged into the Fisher information, from which we can find the standard error for the MLE $\widehat{\boldsymbol{\beta}}$ in the form

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\tau}}}$$
$$\mathcal{J}\left(\widehat{\boldsymbol{\beta}}\right) = X^T V_{\widehat{\boldsymbol{\tau}}}^{-1} X.$$

Problem

The standard errors computed from this plugin formula do not take into account the uncertainty in the estimation of $\boldsymbol{\tau}$; but this approach is nevertheless commonly used.

Note

We can plug in any estimator of the variance components $\boldsymbol{\tau}$ in the above. This is how we get estimate of the fixed effect, $\widehat{\boldsymbol{\beta}}$, using the REML estimators.

4.7 Diagnostics

For normal mixed linear models, the two main distributional assumptions pertain to the normality of the random effects $\boldsymbol{\nu}$ and the errors $\boldsymbol{\epsilon}$.

Normality of Random Effects

For normal mixed linear models, we assumed

$$\boldsymbol{\nu} \sim N(\mathbf{0}, \sigma_v^2 I_m).$$

To check this assumption, some "estimates" of the random variables $\boldsymbol{\nu}$ are required. Typically, the conditional expectations of the random effects, given the observations, are used:

$$\widehat{\boldsymbol{\nu}} = \mathrm{E}(\boldsymbol{\nu} \mid \boldsymbol{y}, \boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}) = \Psi Z^T V_{\widehat{\boldsymbol{\tau}}}^{-1}(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}).$$

In turns out that using QQ plots of the predicted random effects for the purpose of checking their normality are not useful. This is because the observed distribution of $\widehat{\boldsymbol{\nu}}$ does not necessarily reflect the true distribution of $\boldsymbol{\nu}$. In practice, checking the normality assumption for $\boldsymbol{\nu}$ should be based on the results for a linear mixed model with and without assuming this normality (outside scope of course).

Residuals

As for linear models, the main tools for checking the assumption of the normality of the errors, $\boldsymbol{\epsilon}$, are based on residuals. First, recall the normal linear mixed model:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon} \quad \text{or} \quad \boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma_\epsilon^2(I_n + Z\Psi Z^T)).$$

One set of residuals is the conditional residuals defined as

$$\boldsymbol{\epsilon}^{(c)} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}} - Z\widehat{\boldsymbol{\nu}},$$

where $\widehat{\boldsymbol{\nu}} = \mathrm{E}(\boldsymbol{\nu}|\boldsymbol{y}, \boldsymbol{\tau} = \widehat{\boldsymbol{\tau}})$.

Another set is the marginal residuals defined as

$$\boldsymbol{\epsilon}^{(m)} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}.$$

These raw residuals are useful to check heterogeneity of the conditional or marginal variance. However, they are not useful for checking normality assumptions and/or detecting outlying observations. This is because the raw residuals are typically correlated and their variances will differ. Moreover, standardised and Pearson residuals are not appropriate for checking normality. This is because the linear mixed model allows for correlation between the errors. A work-around (approximate)

solution is to transform the raw residual such that the resulting transformed residual is approximately Normal. Then a QQ plot of the transformed residuals should show an approximate straight line. This transformation can be based upon the Cholesky decomposition of the variance-covariance matrix. Some recommend the following diagnostic plots: plot of the marginal residuals against covariates to check the linearity assumptions for the covariates and; plots of the conditional residuals against the estimated conditional means can be used to detect outlying observations or heteroscedasticity of the errors.

Question: Does the REML estimator depend on the choice of transformation, L?

4.7.1 Summary: Fixed or Random Effects

In certain cases, random effects are almost obligatory:

- when we wish to generalise for unseen levels of a factor (i.e., we view the observed levels as samples from a population).
- when we have reason to believe that the grouping effects are actually "tightly" distributed around an overall population mean.
- when the design forces correlation within groups (e.g,. block designs where one treatment only is tested per plot/block).