Chapter 8

Logistic regression

8.1 Objectives

1. To understand odds ratios in the analysis of proportions.
2. To write, fit, assess and interpret logistic regression models.
3. To use R in the fitting, checking and comparison of logistic regression models.

8.2 Odds and the odds ratio

Odds

An alternative to probability is "odds". This is a useful concept: bookmakers think so[1]; and so do statisticians. It provides an alternative measure of chance/likelihood, which spans 0 to ∞ instead of 0 to 1. It is defined as follows:

odds of A (odds for A), $\quad \mathcal{O}(A) = \dfrac{\Pr(A)}{\Pr(A')} = \dfrac{\Pr(A)}{1-\Pr(A)}$.

$0 \leq \Pr(A) \leq 1 \quad \Rightarrow \quad 0 \leq \mathcal{O}(A) \leq \infty$.

$\Pr(A)$	0	0.05	0.0909	0.2	0.5	0.6	0.8	0.95
$\mathcal{O}(A)$	0	0.0526	0.01	0.25	1	1.5	4	19

odds against $A = \mathcal{O}(A') = 1/\mathcal{O}(A)$;

(odds of '4 to 1 on' is equivalent to a probability of 0.8; and odds of '4 to 1 against' is equivalent to probability of 0.2.)

A consequence of the above definition is that $\Pr(A) = \dfrac{\mathcal{O}(A)}{1+\mathcal{O}(A)}$.

Note: A further useful extension of this transformation idea is the log-odds, i.e. $\ln \mathcal{O}(A)$, which spans $-\infty$ to ∞ as $\Pr(A)$ spans 0 to 1.

Comparing risks

Suppose that the risk for group 1 is p_1 and for group 2 the risk is p_2.
How should we compare the groups?

risk difference? $p_1 - p_2$ risk ratio? $\dfrac{p_1}{p_2}$ odds ratio? $\dfrac{p_1}{1-p_1}\dfrac{1-p_2}{p_2}$.

Each of these is used in different situations. When the risks are small, the risk difference will be small too: is a difference of 0.001 important? Maybe if $p_1 = 0.002$ and $p_2 = 0.001$, it is. When the risks are large, the risk ratio is relatively diminished. The fairest comparison turns out to be the odds ratio: it's the one the biostatisticians tend to use. It has a lot of advantages as you will see, although it may not be the simplest.

[1] Though, these days, bookmakers tend to use 'unit payout' = 1+odds.against;

e.g. unit payout = \$2.40 $\Rightarrow$ odds.against = 1.4 $\Rightarrow$ odds = 1/1.4 = 0.714 $\Rightarrow$ prob = 0.417.

In chapters 3 and 4 we looked at how we can compare two proportions, using a normal approximation to the binomial distribution, or the chi-squared and Fisher's exact tests. Here we consider another way which can be generalised much more easily than those methods.

Suppose we are comparing the in-hospital death rate of patients in two hospitals in a 3 month period. We define:

p_1 = proportion of deaths among admissions in hospital 1;
p_2 = proportion of deaths among admissions in hospital 2.

To compare these two proportions, we could use the difference between the two proportions, as we did before. But we could also use the *ratio* of the proportions, or the *odds ratio*. These quantities are defined as follows, using the terminology commonly employed in the biomedical sciences:

- $\delta = p_1 - p_2$ ("risk difference");
- $\phi = \dfrac{p_1}{p_2}$ ("relative risk");
- $\psi = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$ ("odds ratio").

The term "odds" for an event (historically, in gambling) is used to express the imbalance between the chance of the event occurring and the chance of it not occurring. This imbalance is represented by the ratio of these two chances or probabilities. When a bookmaker gives the odds against a horse winning as "10 to 1" it means that they regard the chance of the horse losing as being 10 times greater than the chance of it winning. This makes the implied chance of winning equal to $\frac{1}{11}$, since then the chance of losing is $1 - \frac{1}{11} = \frac{10}{11}$. The odds of losing are then $\frac{10/11}{1/11} = 10$.

If the odds are 1, the inherent probability is $\frac{1}{2}$, since $\frac{\frac{1}{2}}{1-\frac{1}{2}} = 1$.

For the hospital example, suppose that $\hat{p}_1 = 24/100 = 0.24$ and $\hat{p}_2 = 6/60 = 0.10$. Then

- $\hat{\delta} = 0.24 - 0.10 = 0.14$
- $\hat{\phi} = \dfrac{0.24}{0.10} = 2.40$
- $\hat{\psi} = \dfrac{0.24/(1-0.24)}{0.10/(1-0.10)} = 2.84$

All three quantities have their place, depending on the investigation. In this chapter we will concentrate on the last of them, the "odds ratio". The odds against dying in hospital 1 are estimated to be 3.17 to 1, since $\frac{1-0.24}{0.24} = \frac{0.76}{0.24} = 3.17$. In hospital 2, the corresponding odds are $\frac{1-0.10}{0.10} = 9$ or "9 to 1". The odds ratio for comparing hospital 2 to hospital 1 is therefore estimated to be $9/3.17 = 2.84$. If we reverse the order of the hospitals (or consider the odds of dying rather than against dying), the estimated odds ratio (often abbreviated to OR) is $1/2.84 = 0.35$.

The general understanding of odds is not always good, and it is therefore quite common to hear an explanation of the odds ratio that assumes it is really the relative risk: for example, "the odds ratio was 3.4: subjects receiving the new counselling strategy were 3.4 times more likely to report satisfactory harmonisation with their partners". This interpretation of an odds ratio is wrong. The relative risk and the odds ratio are close when the probabilities involved are small, but not otherwise. When the probabilities involved are large—close to 1—the two measures differ greatly.

8.3 Modelling of proportions

The inferences we have considered up until now have involved simple contexts, such as the comparison of two proportions. We now consider the *statistical modelling* of proportions. The modelling techniques we have applied so far have been to normally distributed data, such as analysis of variance, t-tests, and linear regression. However, these are usually inappropriate for proportions because:

1. (Unlike the normal case) the variance of the data is not constant – it is related to the mean;
2. The application of the usual methods can lead to absurd estimates, such as estimated proportions outside the range $(0, 1)$.

The models usually start with the idea that we have data which follow a binomial distribution, and our aim is to *model* p, the probability of "success", as a function of one or more explanatory variables.

So we are envisaging a situation in which we have sample proportions $\hat{p}_i$, where i indexes the distinct proportions. A given sample proportion can be expressed as $\hat{p}_i = y_i/n_i$, where n_i is the total number of units (subjects etc.) and y_i is the number of successes, i.e. the number of units among the total which have the characteristic of interest. The models are analogous to regression models, so we usually use x_1, x_2, etc., to denote explanatory variables.

One simple situation involves a single, continuous explanatory variable, x, and an outcome which is a proportion, y/n for each value of x. For example, a toxic substance is administered to several groups of mice at varying doses, and the proportions of mice developing a tumour are recorded, for each dose. If the substance really is carcinogenic, then the relationship between the dose and the proportion of tumours cannot be linear, because if we increase the dose beyond the minimum level that induces tumours in a whole group, we cannot increase the proportion further. Intuitively, this suggests that any plausible model will show "diminishing returns" as we approach the extremities of 0 and 1.

In other words, we seek an alternative to $p = a + bx$ which has the following features:

- As x increases, so does p (or the opposite);
- No matter what the value of x, p lies between 0 and 1—so that we always get sensible estimates;
- As x gets very small or very large, the effect of a change in x diminishes.

8.3.1 The logistic regression model

One model which has the above properties is the logistic model, whose basic form is:

$$p = \frac{\exp(x)}{1 + \exp(x)} = \frac{e^x}{1 + e^x}$$

Plot the basic logistic function in R to see its shape:

```
> x <-seq(-10,10,.01)
> p <- exp(x)/(1+exp(x))
> plot(x,p,type="l")
```

A simple way of modelling the proportions is to introduce a linear function of x into the model. Now it looks like this:

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Rearranging this gives the model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \qquad \text{(try to show this!)}$$

In words: the logarithm of the odds is a linear function of x. The name given to this particular function is the "logit" function, and the modelling process is known as **logistic regression**. The function logit() is often known as a *link function.*

There are two simple cases to consider:

1. When x is a continuous variable (e.g. age in years), β_1 represents the change in log(odds) corresponding to a change in one unit of x. It follows that e^{β_1} is the odds ratio corresponding to an increase of one unit in x. Also it follows that

   $$\log(\text{odds ratio}) = \beta_1$$

   so we call β_1 the *log-odds ratio.*

2. When $x = 0$ or 1 indexes a particular group (e.g. $x = 1$ for the treatment group and 0 for the placebo group), β_1 represents $\log(OR)$ for treatment relative to placebo. For this case e^{β_1} is the odds ratio for treatment relative to placebo.

The estimates of the βs are maximum likelihood estimates, which have the property that they will be approximately normally distributed in large samples.

8.3.2 One binary explanatory variable

Consider again the hospital example, for which we have already estimated the odds ratio as 2.84 (or 0.35 if we reverse the direction). It is good to have a point estimate of the OR — but what is the uncertainty around this estimate? And does it differ significantly from 1? An odds ratio of 1 corresponds to equal proportions (the usual null hypothesis).

Logistic regression in R

We fit a logistic regression in R using the `glm` function (which stands for "generalised linear model"). There are several ways to set it up; we will use the following protocol: the response variable is the observed proportion, and the total number of "trials" for each category is specified in the `weight` argument. Here is how it works for the hospital example:

```
> admissions <- data.frame(hospital = factor(c(1, 2)), death = c(24,
+     6), total = c(100, 60))
> admissions

  hospital death total
1        1    24   100
2        2     6    60

> admissions.1 <- glm(death/total ~ hospital, family = binomial,
+     weight = total, data = admissions)
```

The output of *glm*() gives us the fitted values for β_0 and β_1.

Note the specification of the binomial `family` of distributions; this overrides the default, which is the normal distribution.

We use the `summary` function to learn more about the fitted model.

```
> summary(admissions.1)

Call:
glm(formula = death/total ~ hospital, family = binomial, data = admissions,
    weights = total)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1527     0.2341  -4.923 8.53e-07 ***
hospital2    -1.0445     0.4899  -2.132    0.033 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.1989e+00  on 1  degrees of freedom
Residual deviance: 1.3323e-15  on 0  degrees of freedom
AIC: 12.301

Number of Fisher Scoring iterations: 3
```

The estimated log-odds ratio β_1 is -1.0445, with a standard error of 0.4899. The ratio of the estimate to its standard error is -2.132, which when compared to a standard normal distribution (hence the term `z value`) gives a P-value of 0.033, implying a significant difference between hospitals. (This test of significance based on an approximation to the normal distribution is known as a *Wald test*).

The estimated odds ratio is the exponential of the log-odds ratio, $e^{-1.0445} = 0.35$ which we arrived at before.

An approximate 95% confidence interval for the log-odds ratio is

$-1.0445 \pm 1.96 \times 0.4899 = (-2.0047, -0.0843)$,

which back-transformed to the OR scale gives

$(e^{-2.0047}, e^{-0.0843}) = (0.13, 0.92)$.

This means that the odds of dying in hospital 2 are estimated to be 0.35 times that in hospital 1, with a 95% confidence interval of $(0.13, 0.92)$. The interval does not include 1 which is equivalent to the interval for the log odds ratio β_1 not including 0, since $e^0 = 1$. This confirms that there is a significant difference in proportions between the hospitals. Hospital 1 is taken as the baseline level of the hospital factor, as it was specified first.

If we take the reciprocal of all quantities above, we can say that the odds of dying in hospital 1 are estimated to be 2.84 times that in hospital 2, with a 95% confidence interval of $(1.09, 7.42)$. This interval of course does not include 1 either.

We will discuss other aspects of the output later.

8.3.3 One categorical explanatory variable (more than two levels)

The extension to more than two levels of a categorical explanatory variable is straightforward. Suppose that there was another hospital (No 3) with 14 deaths out of 40 admissions ($\hat{p}_3 = 14/40 = 0.35$).

```
> admissions3 <- data.frame(hospital = factor(c(1, 2, 3)), death = c(24,
+     6, 14), total = c(100, 60, 40))
> admissions3.1 <- glm(death/total ~ hospital, family = binomial,
+     weight = total, data = admissions3)
> summary(admissions3.1)

Call:
glm(formula = death/total ~ hospital, family = binomial, data = admissions3,
    weights = total)

Deviance Residuals:
[1]  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1527     0.2341  -4.923 8.53e-07 ***
hospital2    -1.0445     0.4899  -2.132    0.033 *
hospital3     0.5336     0.4059   1.315    0.189
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  9.7415e+00  on 2  degrees of freedom
Residual deviance: -2.6645e-15  on 0  degrees of freedom
AIC: 18.362

Number of Fisher Scoring iterations: 3
```

The estimated odds ratio for hospital 3 *relative to hospital 1* is $e^{0.5336} = 1.71$. This means that the odds of dying in hospital 3 are estimated to be 1.71 times that in hospital 1. But this is not significantly different from 1; the P-value is 0.189 for this comparison. Calculate the 95% confidence interval for this OR and confirm that it includes 1.

8.3.4 One numerical explanatory variable

Example Voting behaviour

The data presented below are taken from the 1982 General Social Survey in the USA, and relate the

voting behaviour of white voters in the 1980 Presidential election to political views rated on a scale from 1 to 7 where 1 = extremely liberal and 7 = extremely conservative. Figure 8.1 shows the proportion who voted for Ronald Reagan plotted against political views.

```
> voters <- read.csv("../data/voting.csv")
> voters

  views reagan carter total
1     1      1     12    13
2     2     13     57    70
3     3     44     71   115
4     4    155    146   301
5     5     92     61   153
6     6    100     41   141
7     7     18      8    26

> plot(voters$views, voters$reagan/voters$total, xlab = "Political Views",
+     ylab = "Proportion for Reagan")
```

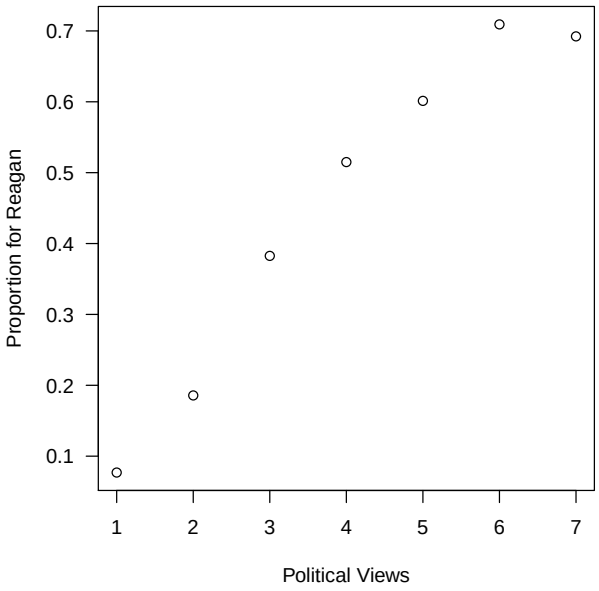

Figure 8.1: Proportion of voters who voted for Reagan vs their stated political views.

Let Y_i denote the number who voted for Reagan out of the n_i voters in group i, i.e. of those with political views $= i, i = 1, \ldots, 7$. Then it is reasonable to assume that $Y_i \sim \text{Bin}(n_i, p_i)$ where p_i is the probability that a voter in group i voted for Reagan. Consider the logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

where $x_i = i$.

Note that this treats the 7-point scale of political views as a continuous variable $i \in [1, 7]$, and assumes a linear scale. This assumption is open to question, but we will accept it for now. In R the model is fitted as follows:

```
> vote.1 <- glm(reagan/total ~ views,
+               family = binomial,
+               weight = total,
+               data = voters)
```

```
> summary(vote.1)

Call:
glm(formula = reagan/total ~ views, family = binomial, data = voters,
    weights = total)

Deviance Residuals:
      1        2        3        4        5        6        7
-1.0277  -1.4430   0.4106   1.0550  -0.1390  -0.2043  -1.3828

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.04464    0.26754  -7.642 2.13e-14 ***
views        0.49570    0.06053   8.190 2.61e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.3323  on 6  degrees of freedom
Residual deviance:  6.3935  on 5  degrees of freedom
AIC: 42.058

Number of Fisher Scoring iterations: 4
```

Interpretation:

1. The coefficients table is essentially the same as we have seen in ordinary regression, but we now use the z value (normal distribution—a Wald test) to determine the P-value instead of t.

2. An approximate 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm 1.96 \times \text{se}(\hat{\beta}_1) = 0.4957 \pm 1.96 \times 0.0605 = 0.4957 \pm 0.1186 = (0.377, 0.614)$$

   This confidence interval does not include zero, which implies that β_1 is significantly different from zero at the 0.05 significance level, and hence that there is a significant association between voting behaviour and political views. Back-transforming to the odds scale gives an estimate of $e^{0.4957} = 1.64$, with a 95% confidence interval for the OR of $(e^{0.377}, e^{0.614}) = (1.46, 1.85)$, which of course does not include 1.

3. For each unit increase in x (i.e. for each scale point increase in conservatism of political view), $\text{logit}(p)$ is estimated to increase by 0.4957, and hence the odds in favour of a vote for Ronald Reagan increase by a factor of $e^{0.4957} = 1.64$. This estimated odds ratio therefore has a multiplicative interpretation – it means that the odds are estimated to increase by 64% for each unit increase in the score.

4. Residuals in the strict sense that we have understood them are no longer useful. However, they have a corresponding quantity, the so-called *deviance* residuals.

5. The dispersion parameter can be ignored for our purposes. It is important, but beyond the scope of this subject.

6. When $p = 0.5$, $\text{logit}(p) = 0$, hence an estimate of the value of x (political views) for which $p = 0.5$ is given by $-\dfrac{\hat{\beta}_0}{\hat{\beta}_1}$ (show this!) which gives us
   $\dfrac{2.045}{0.4957} = 4.125.$
   The value of x for which Pr(response) = 0.5 is often referred to as the 'LD(50)' (lethal dose 50), which is relevant in toxicity studies (though not in voting behaviour!).

7. The "Number of Fisher Scoring iterations" is the number of times R had to iterate the algorithm to find the maximum-likelihood estimates for the fitted values. Logistic regression requires an iterative procedure, and occasionally there can be problems with convergence.

8.3.5 More than one numerical explanatory variable

The case of more than one numerical explanatory variable with logistic regression can be handled in much the same way as with standard linear regression (Chapter 7).

- Interpretation of the coefficients in a logistic regression is a bit different from the standard linear regression case, as we have seen in the previous section.
- Comparison of these models can be based on the AIC (Akaike Information Criterion) or the F-statistic.

8.3.6 Parameter estimation

The parameters $\beta_0, \beta_1, \ldots$ are (usually) estimated using the method of maximum likelihood.

The maximum likelihood estimate is the value of the parameter that is most likely to have given rise to the observed data, assuming that the proposed model is correct. The exact properties of these estimators are complicated, but asymptotically, the estimators are normally distributed and unbiased. These asymptotic results are used to obtain standard errors for the estimates and to carry out various forms of inference.

8.4 Testing the significance of terms in a model

In the examples considered so far, we have used the P-value in the table of coefficients to test the significance of a term in the logistic regression model. This has been possible because in all of these cases, the test has involved just two levels of a factor, or a continuous variable. To fit a wider class of models, we need a procedure that can test the overall significance of a factor with more than two levels; in fact, we need a test that can compare any two nested models.

The residual deviance

The residual deviance in a generalised linear model (of which logistic regression is one) is analogous to the residual sum of squares in a linear model. It quantifies the variation in the data which is unexplained by the model. It is defined as

$$D = 2\sum_i y_i \log\left(\frac{y_i}{\hat{y}_i}\right)$$

where the sum is taken over both successes and failures. It is given by the `deviance` command:

```
> deviance(vote.1)

[1] 6.393481
```

Under the null hypothesis, the residual deviance asymptotically (as each $n_i \to \infty$) follows a χ^2 distribution with degrees of freedom equal to the residual df.

Change in residual deviance

In the same way that the change in residual sum of squares is used to compare two nested linear models via an F test, the change in residual deviance can be used to compare two nested generalised linear models, but with a χ^2 test. As with linear models, the test is whether the more complex model is significantly better than the simpler one. For example, $\text{logit}(p_i) = \beta_0$, the null model, is a special case of $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$, (obtained when $\beta_1 = 0$), and so the former model is nested within the latter.

In general, if $M(1)$ and $M(2)$ are two models such that $M(1)$ is a special case of $M(2)$ (i.e. $M(1)$ is nested within $M(2)$), then $D(2) \leq D(1)$, where $D(i)$ is the residual deviance for model $M(i)$. $D(1) - D(2)$ can be used to test whether model $M(2)$ is significantly better than model $M(1)$. The test is a χ^2 test with $(df_1 - df_2)$ degrees of freedom, where df_i is the residual degrees of freedom for model $M(i)$.

Note: If there are multiple explanatory variables, then as is the case for multiple regression, the order in which terms are included in the model statement will affect the order in which they are tested, and in turn may affect their statistical significance by this test.

Example Voting behaviour

From the output on page 6 we can obtain the residual deviance for both the null model (listed as `Null deviance` and for the more complex model. They are as follows:

Model	logit(p_i)	residual deviance (D)	df
$M(1)$	β_0	82.3	6
$M(2)$	$\beta_0 + \beta_1 x_i$	6.4	5

Here, model $M(1)$, which is a special case of model $M(2)$, implies that logit(p) is the same for all groups, which is equivalent to saying that voting behaviour is independent of political views.

The change in residual deviance (which is often abbreviated to "change in deviance") is $82.3 - 6.4 = 75.9$, and this is compared to a χ^2 distribution on $6 - 5 = 1$ df. Quantiles of the chi-squared distribution are found in R by `qchisq`. For example:

```
> qchisq(0.999, 1)

[1] 10.82757

> qchisq(0.99, 1)

[1] 6.634897
```

75.9 is much larger than 10.83 ($= \chi^2_{1(0.999)}$), so $M(2)$ is significantly better than $M(1)$, at the 0.001 level.

This comparison can be done entirely in R:

```
> anova(vote.1, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: reagan/total

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     6     82.332
views  1   75.939         5      6.393 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

8.5 Two categorical explanatory variables

Example **Car preferences**

The data are presented in the table below. There are two factors: gender and residence (city vs country). Preference for local or imported cars is the binary response variable.

Gender	Residence	Prefer local car	Prefer import	Total
male	city	168	68	236
	country	32	12	44
female	city	84	16	100
	country	164	24	188

We apply the *step*() function from Chapter 7 using first backward then the forward method, which will compare the AIC for the various models.

```
> carpref.dat <- read.csv("../data/cars.csv")
> carpref.dat
```

```
  local total gender residence
1   168   236   male      city
2    32    44   male   country
3    84   100 female      city
4   164   188 female   country
```

```
> carpref.1 <- glm(local/total ~ gender + residence,
+                  family = binomial,
+                  weight = total,
+                  data = carpref.dat)
> step(carpref.1,~.)
```

```
Start:  AIC=25.21
local/total ~ gender + residence

            Df Deviance    AIC
- residence  1   0.6043 23.677
<none>           0.1351 25.208
- gender     1  11.6202 34.693

Step:  AIC=23.68
local/total ~ gender

            Df Deviance    AIC
<none>           0.6043 23.677
+ residence  1   0.1351 25.208
- gender     1  19.2363 40.309

Call:  glm(formula = local/total ~ gender, family = binomial, data = carpref.dat,
    weights = total)

Coefficients:
(Intercept)   gendermale
     1.8245      -0.9083

Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
Null Deviance:            19.24
Residual Deviance: 0.6043         AIC: 23.68
```

```
> carpref.2 <- glm(local/total~1,
        +family = binomial,
        +weight = total,
        +data = carpref.dat)
> step(carpref.2,~.+gender+residence)
```

```
Start:  AIC=40.31
local/total ~ 1

            Df Deviance    AIC
+ gender     1   0.6043 23.677
+ residence  1  11.6202 34.693
<none>          19.2363 40.309

Step:  AIC=23.68
local/total ~ gender

            Df Deviance    AIC
<none>           0.6043 23.677
+ residence  1   0.1351 25.208
- gender     1  19.2363 40.309
```

```
Call:  glm(formula = local/total ~ gender, family = binomial, data = carpref.dat,
    weights = total)

Coefficients:
(Intercept)   gendermale
     1.8245      -0.9083

Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
Null Deviance:            19.24
Residual Deviance: 0.6043         AIC: 23.68
```

Both the forward and backward methods yielded the same result — that we could obtain the "best" model (according to the AIC criterion) using just *gender* as the explanatory variable.

```
> carpref.3 <- glm(local/total ~ gender, family = binomial, weight = total,
+     data = carpref.dat)
> summary(carpref.3)

Call:
glm(formula = local/total ~ gender, family = binomial, data = carpref.dat,
    weights = total)

Deviance Residuals:
       1         2         3         4
-0.08227   0.19158  -0.59829   0.45044

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8245     0.1704  10.708  < 2e-16 ***
gendermale   -0.9083     0.2157  -4.211 2.55e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19.23630  on 3  degrees of freedom
Residual deviance:  0.60431  on 2  degrees of freedom
AIC: 23.677

Number of Fisher Scoring iterations: 3
```

The output provides us with an estimate of β, however, we are more interested in $\exp(\beta)$, or the odds ratio.

```
> exp(summary(carpref.3)$coef[2, 1])

[1] 0.4032258
```

(coef[2,1] selects the value in the second row and first column of the table of coefficients (-0.9083).

We can interpret this odds ratio as follows: the odds of a male preferring a local car are less than the odds of a female preferring a local car, by a factor of 0.40. Or equivalently, the odds of a female preferring a local car are greater than the odds of a male preferring a local car by a factor of $1/0.4$ or 2.5.

8.6 Exercises

1. **Aspirin and stroke** – Lab 3 Exercises, Q6.

   (a) Perform a hypothesis test to compare the aspirin and control groups using logistic regression.

(b) Estimate the odds ratio and calculate a 95% confidence interval for it. Examine whether it gives results which are broadly consistent with the confidence interval derived in Lab 3 Q6 for the difference between proportions.

2. **Control of budworm**

An experiment was conducted to examine the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid trans–cypermethrin. Batches of 20 moths were exposed for three days to six doses of the pyrethroid and the number in each batch that were dead was recorded.

```
> moths <- data.frame(dose = c(1, 2, 4, 8, 16, 32), dead = c(1,
+     4, 9, 13, 18, 20), total = rep(20, 6))
```

(a) Investigate whether or not there is an effect of the dose of the pyrethroid on the proportion of deaths.

(b) Give an interpretation of the coefficient of dose in the fitted model.

(c) Estimate the LD50 (the dose which kills half the insects).

(d) Note that the doses increase in a multiplicative way rather than additively. To make the increase additive, take logs of the doses, and fit the model again with log(dose) as the explanatory variable. Does this model provide a better fit?

3. **Powdery mildew on broccoli**

A glasshouse experiment was conducted to examine the control of the fungal disease powdery mildew on broccoli. Three chemical treatments were used (control plus two active fungicides), replicated 5 times in a completely randomised design. The experimental unit was a pot planted with 18 broccoli seeds in soil inoculated with the fungus. Not all the seeds emerged, but for those that did, the number of seedlings alive after two weeks was recorded. The variable of interest was the proportion of live seedlings out of those that emerged. The results were as follows:

	control					fungicide 1					fungicide 2				
No. Alive	3	2	2	1	1	10	8	15	14	8	17	14	13	10	16
No. Emerged	17	15	16	11	15	16	11	18	16	12	18	15	16	13	16

(a) Read the data into R, and calculate the average proportion of live seedlings for the three treatments.

(b) Using logistic regression, investigate the effect of the treatments on the proportion of live seedlings.

(c) Estimate the odds ratio, together with a 95% confidence interval, for the effect of fungicide 1 relative to the control.

4. **Aboriginal deaths in custody**

These data (from OzDASL[2]) show the numbers of persons in prison custody in Australia in 1990 and 1995, and the number of deaths. The data are categorised into indigenous and otherwise.

Year	Indigenous	Prisoners	Deaths	Population
1990	Yes	2041	6	168317
1995	Yes	2907	17	190438
1990	No	12264	27	13141817
1995	No	14501	42	13995940

Read the data into R.

(a) Did the mortality rate among prisoners increase from 1990 to 1995?

[2] `http://www.statsci.org/data/index.html`

(b) Was the overall mortality rate for indigenous prisoners greater than for non-indigenous prisoners?

(c) If the answer to (a) and/or (b) was yes, express the relative mortality rate using an appropriate quantity and give an interpretation of this quantity.

(d) For prisoners in 1990, compare the relative risk (indigenous vs non-indigenous) with the odds ratio, and comment on their closeness or otherwise.