Chapter 4. Categorical Data Analysis by Logistic Regression Models

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Contents

§4.1 Introduction

- A **categorical variable** is one for which the measurement scale consists of a set of categories. Each category in the variable is called a **level**.
- We consider situations in which the response variable is categorical with continuous and/or categorical explanatory variables, where the continuous explanatory variable can be converted into a category variable as well.
- We also consider situations where it is not appropriate to distinguish between response and explanatory variables.

§4.1 Methods

- **Logistic regression** and **log-linear models** are two closely related, methods for analysing categorical data.
- Logistic regression can be used when we have a uni-variate binomial or proportion response variable, whereas log-linear models can be used in more general situations.
- Logistic regression will be shown to be a special type of log-linear model.

§4.2 Example 1: Car Preferences

The following data were obtained from a study of car preference (local or imported) among city and country residents.

1. Males:

Residence	Preference	
	Local	Imported
City	168	68
Country	32	12

2. Females:

Residence	Preference	
	Local	Imported
City	84	16
Country	164	24

(to be continued)

§4.2 Example 1: Car Preferences Data ctd...

3 Combined (or collapsed):

	Preference	
Residence	Local	Imported
City	252	84
Country	196	36

- Three factors: `Gender`, `Residence` and `Preference`, forming a 3-way contingency table (i.e. a $2 \times 2 \times 2$ contingency table). It can be `collapsed` over `Gender` to get a 2-way table.
- Reasonable to set `Preference` as the response variable.
- Want to see the association between `Preference` and the other two factors.

§4.2 Example 2. Voting behaviour (1)

The voting behaviour of white voters in the 1980 Presidential Vote versus their political views rated on a scale from 1 to 7 where 1 = extremely liberal and 7 = extremely conservative.

Political Views	Reagan	Carter or other	Total
1	1	12	13
2	13	57	70
3	44	71	115
4	155	146	301
5	92	61	153
6	100	41	141
7	18	8	26

- It is reasonable to assume `Reagan`, the number of voters voting for Reagan, to follow a binomial distribution.
- The aim is to find how the probability of voting for Reagan is related to `Political Views`.

§4.2 Example 2. Voting behaviour Plot

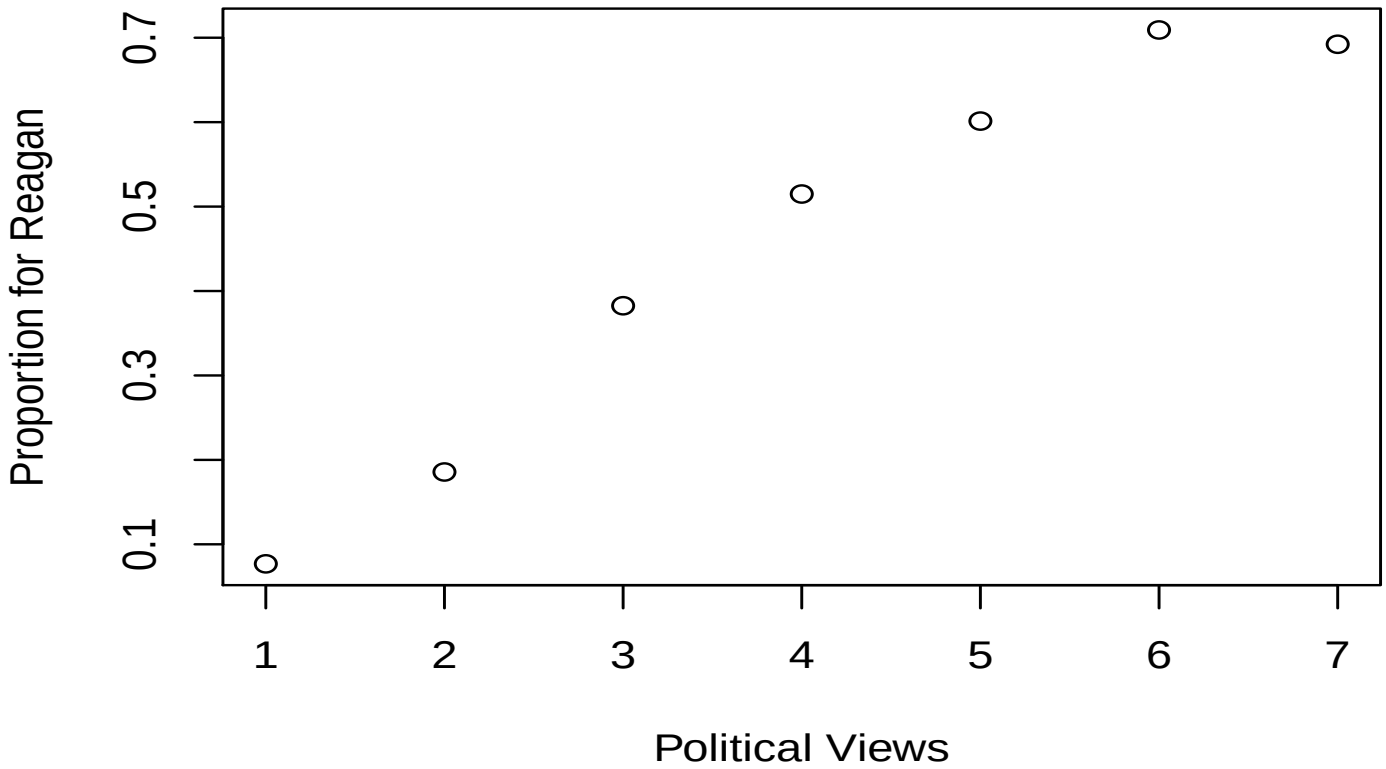

Figure 4.1: Proportion of voters who voted for Reagan plotted against their stated political views

```
voting<-read.csv("C:/Users/qguoqi/OneDrive - The University of Melbourne/
                  My Documents/MAST90139/data/voting.csv")
plot(voting$views, voting$reagan/voting$total,
                   xlab="Political Views", ylab="Proportion for Reagan")
```

§4.3 Bernoulli trials and the binomial distribution

- **Bernoulli trials:** a sequence of trials each with just two possible outcomes, "success" and "failure" and Pr(success) $= \pi$, independent of the outcome of any other trial.
- **Binomial distribution:** the number (Y) of successes in m Bernoulli trials with Pr(success) $= \pi$; $Y \overset{d}{=} \text{Bin}(m, \pi)$

$$\Pr(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y} \quad \text{for} \quad y = 0, 1, 2, \cdots, m$$

$\mathbb{E}(Y) = m\pi$; $\text{Var}(Y) = m\pi(1-\pi)$.

- **Normal approximation:** If $Y \overset{d}{=} \text{Bin}(m, \pi)$ and $X \overset{d}{=} N(m\pi, m\pi(1-\pi))$ then $Y \overset{d}{\approx} X$

$$\Pr(Y = y) \approx \Pr(y - 0.5 \leq X \leq y + 0.5)$$

provided both $m\pi$ and $m(1-\pi)$ are ≥ 5.

§4.4 Link functions for binomial data

- **Logit** link:

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

- **Probit** link:

$$\text{probit}(\pi) = \Phi^{-1}(\pi),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution.

- **Complementary log-log** link:

$$\text{comp-log-log}(\pi) = \ln(-\ln(1-\pi))$$

- **log-log** link:

$$\text{log-log}(\pi) = -\ln(-\ln\pi)$$

§4.4 Voting behaviour example

- Let Y_i denote the number of voters who voted for Reagan out of the n_i voters in group i, (i.e. those with political views = i).
- We assume that $Y_i \stackrel{d}{=} \text{Bin}(n_i, \pi_i)$ and follows a **logistic regression** model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

  where $x_i = i$ is the political view of group i.
- For this model $\hat{\beta}_0 = -2.045$; $\hat{\beta}_1 = 0.496$.
- Next figure plots the empirical logits against political views, together with the fitted logistic regression line. The empirical logits are computed as

$$\text{emp.logit}(\pi) = \ln\left(\frac{y + 0.5}{n - y + 0.5}\right)$$

Scatterplot.

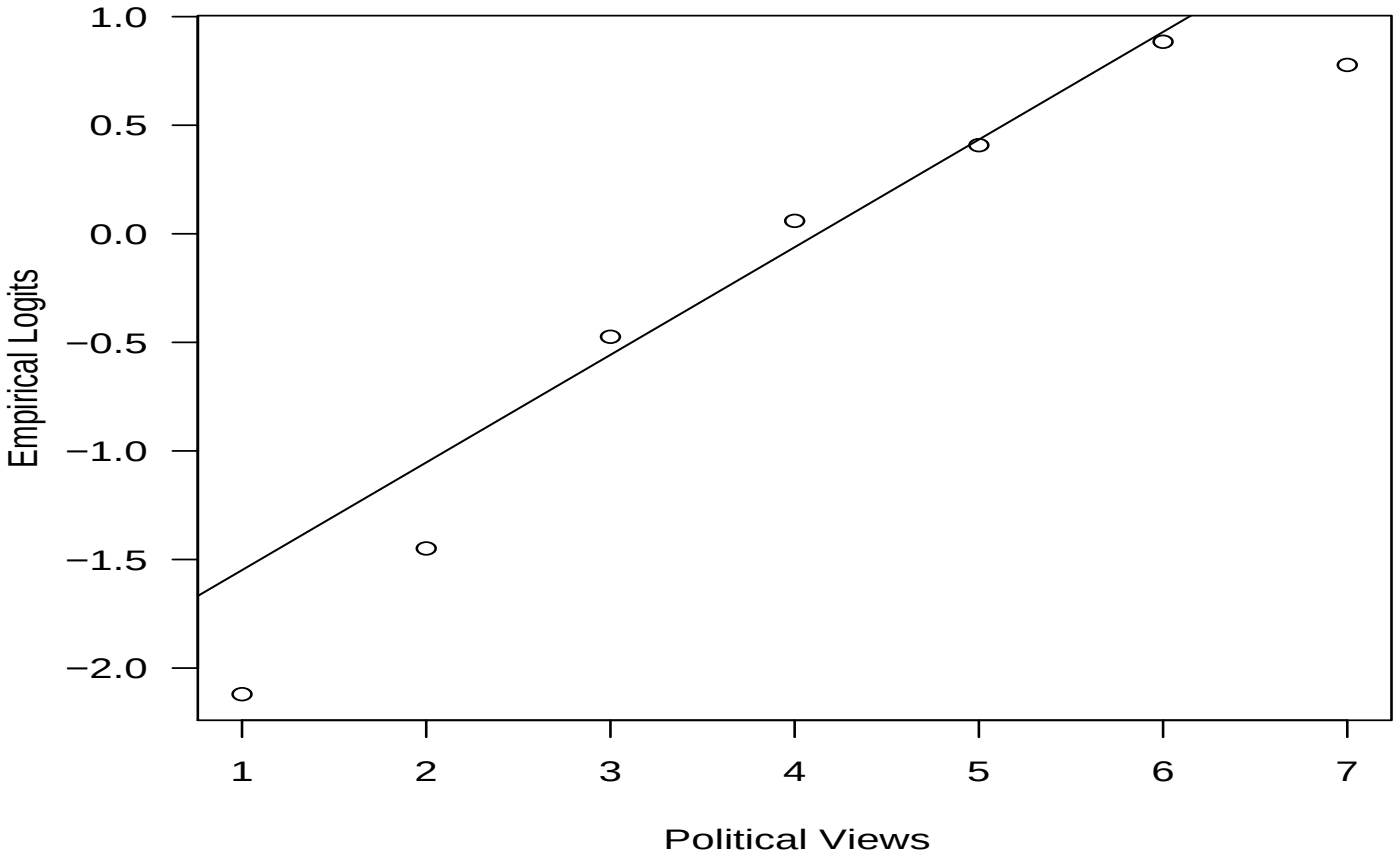

Figure 4.2: Empirical logits vs. Political Views

Voting behaviour example R output (1)

```
vote.dat<-read.csv("C:/Users/qguoqi/OneDrive - The University of Melbourne/
                My Documents/MAST90139/data/voting.csv")
vote.dat
```

```
  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
```

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

Voting behaviour example R output (2)

```
summary(vote.1)

Call:
glm(formula = reagan/total ~ views, family = binomial, data = vote.dat,
    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
```

Voting behaviour example residual plots

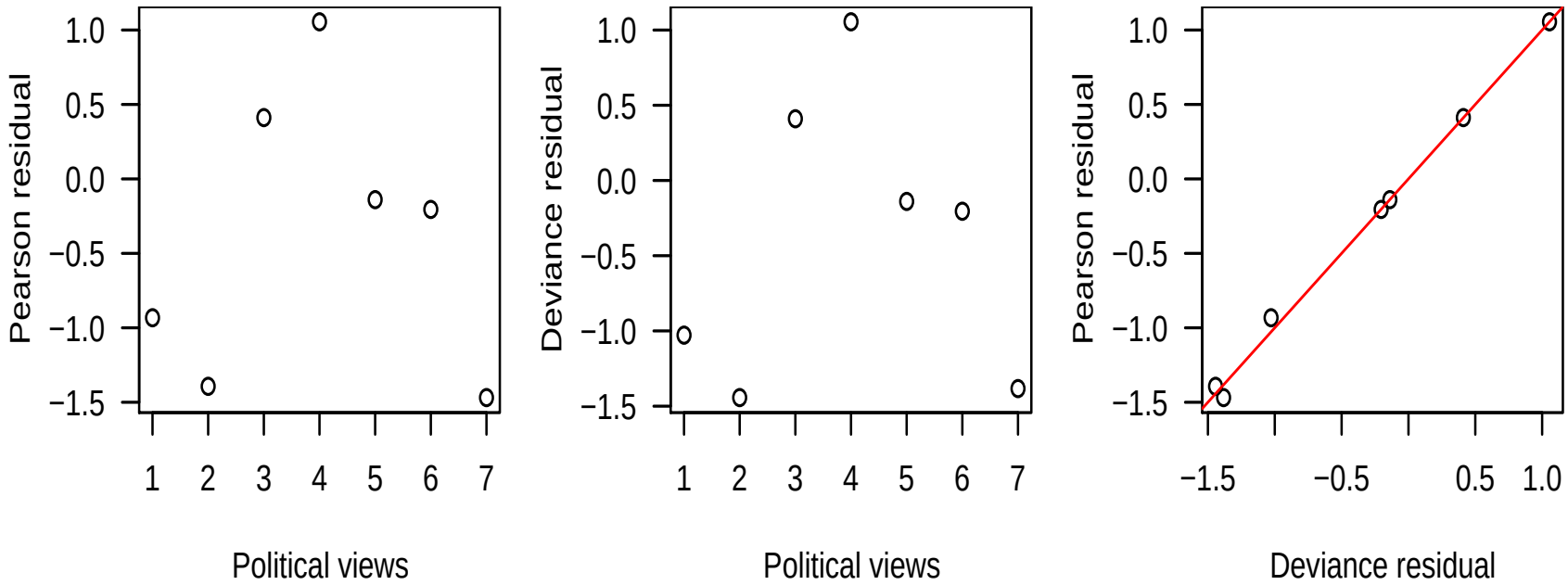

Figure 4.3: Pearson and deviance residuals vs. political views, and vs. each other

Voting behaviour example R output (3)

```
anova(vote.1, test = "Chi")

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 2.926e-18
```

Voting behaviour example analysis

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

$$\hat{\beta}_1 \pm 1.96 \times \text{se}(\hat{\beta}_1) = 0.4957 \pm 1.96 \times 0.06052 = 0.4957 \pm 0.1186.$$

   The CI does not include zero $\Rightarrow \beta_1 \neq 0$, and hence that there is significant association between voting behaviour and political views.

2. "LD(50)": When $\pi = 0.5$, $\text{logit}(\pi) = 0$, hence an estimate of the value of x (political views) for which $\pi = 0.5$ is given by

$$-\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{2.045}{0.4957} = 4.125.$$

3. The value of x for which Pr(response) $= 0.5$, often referred to as the "LD(50)" (lethal dose 50), provides a useful point of reference when thinking about the relationship between a (continuous) explanatory variable (or covariate) and "π" [$=$ Pr(response)].

Odds, log-odds and the logistic function

Prob $= \pi$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Odds $= \frac{\pi}{1-\pi} = \phi$	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9
log(odds) $=$ logit(π)	-2.20	-1.39	-0.85	-0.41	0	0.41	0.85	1.39	2.20

Properties of logit function:

1. $\ln(\phi) = \text{logit}(\pi) = \text{log-odds} \overset{denote}{=} \gamma$
2. $\text{logit}(1-\pi) = -\text{logit}(\pi)$
3. $\text{logit}(0) = -\infty; \quad \text{logit}(0.5) = 0; \quad \text{logit}(1) = \infty.$

Relationships between probability, odds and log–odds

Probability	Odds	log–odds (logit)
π	$\frac{\pi}{1-\pi} : 1$	$\ln(\frac{\pi}{1-\pi})$
$\frac{\phi}{1+\phi}$	ϕ	$\ln(\phi)$
$\frac{e^\gamma}{1+e^\gamma}$	e^γ	γ

Parameter estimation

- Uses the method of maximum likelihood.
- Closed form expressions not possible $\Rightarrow$ iterative procedure needed (iteratively reweighted least squares or IRWLS — see Chapter 3).
- Packages give the number of iterations required for procedure to converge; there can be problems with convergence.
- Exact properties of estimators (distributions, means, variances etc) not known but, **asymptotically**, estimators are normally distributed, unbiased with variance (or dispersion) matrix given by the inverse of the "Information Matrix", see Chapter 3.
- Asymptotic results used to obtain standard errors of estimates and to carry out various forms of inference.

Interpretation of the logistic regression model `vote.1`

- For $\text{logit}(\pi) = \beta_0 + \beta_1 x$, a unit increase in $x \Rightarrow \text{logit}(\pi)$ increases by $\beta_1 \Rightarrow$ odds in favour of a response increase by a factor of $\exp(\beta_1)$.
- **Application to voting example:**

$$\widehat{\text{logit}(\pi_i)} = -2.045 + 0.4957 x_i$$

x	$\hat{\gamma} = \widehat{\text{logit}(\pi)}$	$\hat{\pi} = \frac{e^{\hat{\gamma}}}{1+e^{\hat{\gamma}}}$	$\widehat{\text{odds}} = e^{\hat{\gamma}} \ : \ 1$
1	-1.5493	0.175	0.212
2	-1.0536	0.259	0.349
3	-0.5579	0.364	0.572
4	-0.0622	0.485	0.940
5	0.4336	0.607	1.543
6	0.9292	0.717	2.532
7	1.4249	0.806	4.157

Odds ratios in `vote.1`

Odds ratios:

$$\left.\begin{array}{rcl} \frac{0.349}{0.212} & = & 1.646 \\ \frac{0.572}{0.349} & = & 1.646 \\ \vdots & \vdots & \vdots \\ \frac{4.157}{2.532} & = & 1.646 \end{array}\right\} = e^{0.4957} = e^{\hat{\beta}_1}$$

- When the political view score increases by 1 unit, the odds of voting for Reagan vs. for Carter increases by a factor of 1.646.

Confidence intervals for odds ratios

Approximate 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm 1.96 \times \text{se}(\hat{\beta}_1) = (\hat{\beta}_{1L}, \hat{\beta}_{1U}), \text{ say.}$$

Hence an approximate 95% confidence interval for the odds ratio is

$$\left(e^{\hat{\beta}_{1L}},\ e^{\hat{\beta}_{1U}}\right)$$

Application to voting example:

$$\hat{\beta}_1 \pm 1.96 \times \text{se}(\hat{\beta}_1) = (0.3771,\ 0.6143).$$

Hence an approximate 95% confidence interval for the odds ratio is

$$(e^{0.3771},\ e^{0.6143}) = (1.458,\ 1.848).$$

The interval **does not** include "1", hence the effect of political views is (statistically) significant (at the 5% level).

§4.5 Tests of adequacy (or goodness) of fit

- When the response variable is binomial, it is possible to test the adequacy of fit of a GLM by comparing the observed frequencies ($O_i = y_i$) with those expected ($E_i = m_i\hat{\pi}_i = \hat{y}_i$) if the model is correct. Here m_i is the size of group i.
- There are two, commonly used, tests:
  - Pearson's chi-square test, and
  - the likelihood ratio test (LRT).

Pearson's Chi-square test for goodness of fit

- If y_i is the observed binomial response for group i with group size m_i, the Pearson χ^2 test statistic is

$$X^2 = \sum_i \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} = \sum_i \frac{(y_i - \hat{y}_i)^2}{\widehat{\text{Var}}(y_i)}$$

  where the sum is taken over all "successes".
- If y_i is a count (frequency) number in cell i of a contingency table, the Pearson χ^2 test statistic is

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} = \sum_i \frac{(y_i - \hat{y}_i)^2}{\widehat{\text{Var}}(y_i)}$$

  where the sum is taken over both "successes" and "failures".
- For logistic model `vote.1`, the value of X^2 can be obtained from `R`

```
sum(resid(vote.1,type="pearson")^2)
```

  on a glm "object" (`vote.1` in this case).

Likelihood ratio test (residual deviance) for goodness of fit

- If y_i is the observed binomial response for group i with group size m_i, the residual deviance statistic of the model is

$$D = 2\sum_i \left[y_i \ln \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \ln \frac{m_i - y_i}{m_i(1-\hat{\pi}_i)} \right]$$

  where the sum is taken over all "successes".
- If y_i is a count (frequency) number in cell i of a contingency table, the residual deviance statistic of the model is

$$D = 2\sum_i \left[y_i \ln \left(\frac{y_i}{\hat{y}_i} \right) - (y_i - \hat{y}_i) \right]$$

  again, the sum is taken over both "successes" and "failures".
- The residual deviance of model `vote.1` can be obtained from `R` by the "deviance" command, eg

```
deviance(vote.1)
```

Properties of goodness of fit tests

- The two tests are asymptotically χ^2 (as all $\mathbb{E}(y_i) \to \infty$) with $(k - r)$ degrees of freedom, where k is the number of groups and r is the number of regression parameters.
- The χ^2 approximation is OK provided all of the expected frequencies are ≥ 1 and no more than 20% of them are < 5.
- **The null hypothesis being tested is that the model provides an adequate fit to the data, which is rejected at 5% level if the calculated test statistic value is larger than $\chi^2_{0.95}(k - r)$.**
- The two tests are asymptotically equivalent.

$$\frac{X^2}{D} \overset{prob}{\to} 1 \text{ as } m_i \to \infty, \ \forall \ i$$

Example: Voting behaviour

	Reagan		Carter or other		
x	observed	expected	observed	expected	total
1	1	2.3	12	10.7	13
2	13	18.1	57	51.9	70
3	44	41.9	71	73.1	115
4	155	145.8	146	155.2	301
5	92	92.8	61	60.2	153
6	100	101.1	41	39.9	141
7	18	21.0	8	5.0	26

$X^2 = 6.31$, $D = 6.39$, $k = 7$ and $r = 2$ so that the χ^2 tests have 5 degrees of freedom. [Critical value $= \chi^2_{0.95}(5) = 11.07$]

Testing the significance of terms in a GLM

- This is a linear hypothesis testing problem which can be tested by three common methods: *Wald test*, *likelihood ratio test* (LRT), and *score test*. The LRT is the most common one used.
- Here **LRT** statistic equals the *change in residual deviance* in comparing two relevant models, where one model is a special case (nested) of the other, and the terms being tested are the difference between these two models.
- If model $M(1)$ is nested within model $M(2)$, then residual deviance $D(2) \leq D(1)$, and $D(1) - D(2)$ can be used to test whether model $M(2)$ is significantly better than model $M(1)$, or equivalently whether the extra terms in $M(2)$ over $M(1)$ are significant or not.
- $\Delta D = D(1) - D(2)$ asymptotically follows a $\chi^2(df_1 - df_2)$ distribution when $M(1)$ is correct. Here df_i is the (residual) degrees of freedom for model $M(i)$.
- **Note:** Change in Pearson's X^2 cannot be used to compare $M(2)$ with $M(1)$, since $X^2(1) - X^2(2)$ can be negative and hence cannot be χ^2.

Example: Voting behaviour

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

Model $M(1) \Longleftrightarrow$ voting behaviour is independent of political views.

Tests of adequacy of fit

- $M(1) : 82.3 > 22.46\ (= \chi^2_{0.999}(6))$, hence $M(1)$ is not adequate at 0.1% level.
- $M(2) : 6.4 < 11.07\ (= \chi^2_{0.95}(5))$, hence $M(2)$ has adequate fit at the 5% level.

Comparison of models [$M(1)$ versus $M(2)$]

- $75.9\ (= 82.3 - 6.4) > 10.83\ (= \chi^2_{0.999}(1))$, hence model $M(2)$ is significantly better than model $M(1)$, i.e. $\beta_1 \neq 0$ at the 0.1% level.

Wald test

- Based on the asymptotic normality of maximum likelihood estimators.
- Amounts to comparing the estimate divided by its standard error to the unit normal distribution (a Z test) or, equivalently, [estimate/se(estimate)]^2 to a χ^2_1 distribution.
- Can be extended to include the simultaneous testing of a set of parameters.
- Usually gives similar results to those obtained using the change in (residual) deviance, but the two tests are not equivalent.

Example: Voting behaviour (continued)

```
> summary(vote.1)$coef
                 Value Std. Error   z value
(Intercept) -2.0446385 0.26745833 -7.644700
          x  0.4957038 0.06050581  8.192665
```

Testing significance of political views (x) amounts to testing whether its coefficient is significantly different from zero,

$$Z\ (=\text{z value}) = 0.4957/.06052 = 8.19 \quad \text{cf. N(0,1)}$$

or

$$W = Z^2 = 8.19^2 = 67.09 \quad \text{cf. } \chi^2_1$$

$$[\text{cf. LRT } \Delta(\text{deviance}) = 82.3 - 6.4 = 75.9]$$

Residuals for GLM with binomial response

There are three kinds of residuals, each with a different role:

1. Pearson residuals
2. Standardised residuals
3. Deviance residuals

Pearson residuals

Raw residuals, $y_i - m_i\hat{\pi}_i$, are not especially useful since variances of the y_is are not constant.

Pearson residuals are defined as

$$r_i = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i/m_i)}}$$

Note

- $X^2 = \sum_{i=1}^{2k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} r_i^2$
- Standard deviations of (Pearson) residuals are close to 1 so that we expect most ($\approx 95\%$) to lie between –2 and +2, if the fitted model is correct.

Standardised residuals

- It is also possible to make an additional adjustment to the residuals (based on leverage as in a linear model) which ensures that the standard deviations of all residuals are equal, and very close to 1.
- The resultant residuals are sometimes referred to as *standardised (or adjusted) Pearson residuals*. However, we will not pursue this in this subject.

Deviance residuals for GLM with binomial response

- **Deviance residuals** in GLM with binomial response are defined as

$$d_i = \text{sign}(y_i - \hat{y}_i)\sqrt{2y_i \ln \frac{y_i}{\hat{y}_i} + 2(m_i - y_i) \ln \frac{m_i - y_i}{m_i - \hat{y}_i}}$$

- Note the residual deviance of the underlying model $D = \sum_{i=1}^{k} d_i^2$
- Deviance residuals can also be "standardised". Not pursued here.
- For a correct model, the distribution of the deviance residuals is "closer" to normal than that of the Pearson residuals.

Examples: Voting behaviour (continued)

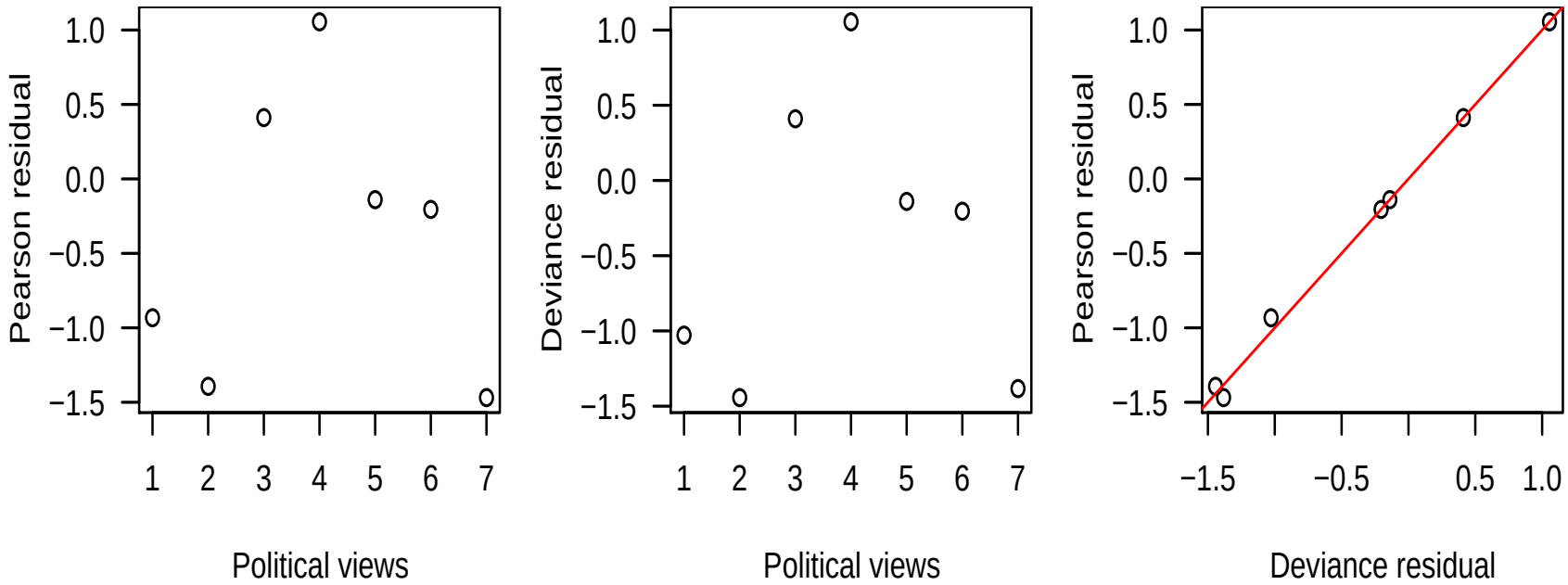

Figure 4.4: Pearson and deviance residuals vs. political views, and vs. each other

Examples: Voting behaviour (continued)

None of the residuals are especially large (all < 1.5 in magnitude), but there is evidence of a pattern which suggests that the quadratic model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

may be appropriate.

Model	$\text{logit}(\pi_i)$	residual deviance (D)	df
$M(1)$	β_0	82.3	6
$M(2)$	$\beta_0 + \beta_1 x_i$	6.4	5
$M(3)$	$\beta_0 + \beta_1 x_i + \beta_2 x_i^2$	1.1	4

Since $D_2 - D_3 = 6.4 - 1.1 = 5.3 > 3.84 = \chi^2_{0.95}(1)$, the p-value is < 0.05, and $M(3)$ is significantly better than $M(2)$.

Examples: Voting behaviour (continued)

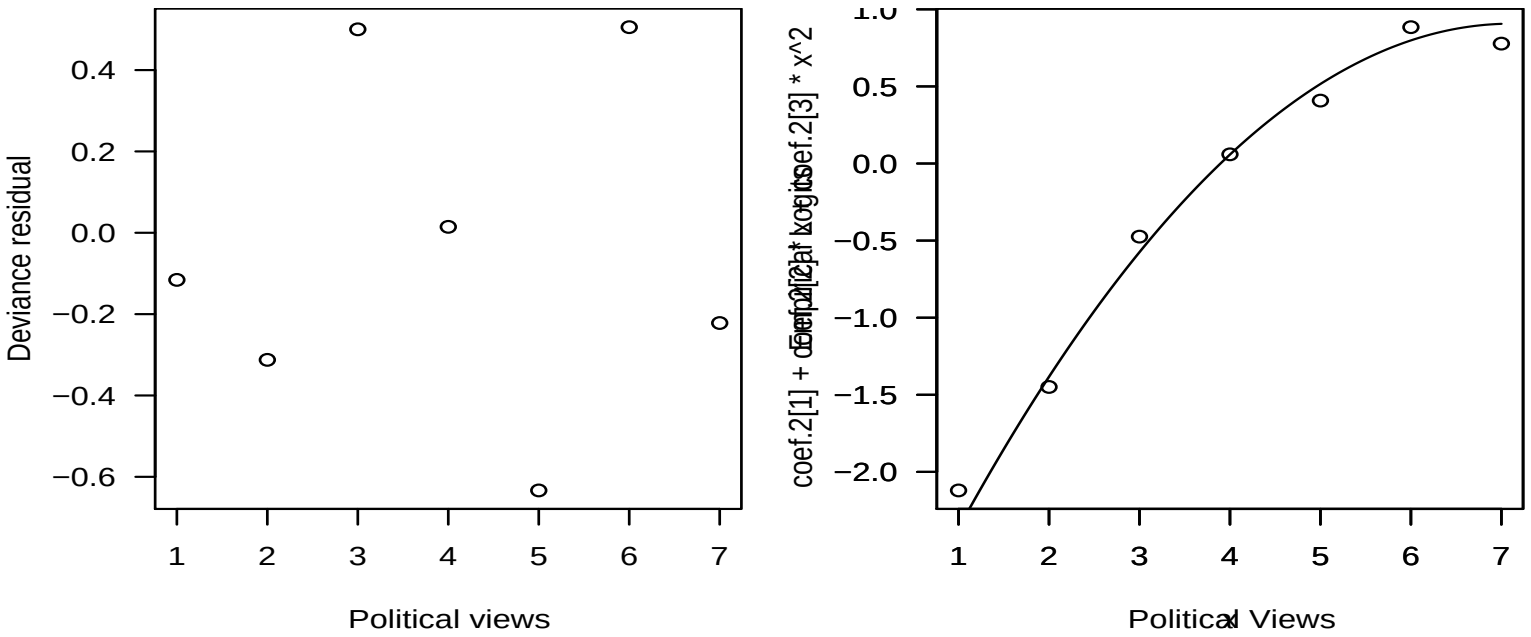

Figure 4.5: Deviance residuals and logit($\hat{\pi}_i$) vs. political views, respectively

- Residuals small in magnitude, with no evidence of a pattern.
- Appropriateness of quadratic model indicated in right hand panel.

More Examples

We now explore more examples having binomial response.

Car preference data

Gender	Residence	Prefer local car	Total
male	city	168	236
	country	32	44
female	city	84	100
	country	164	188

R analysis of car preference data

```
> carpref.dat <- read.csv("../data/cars.csv")
> print(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)
> anova(carpref.1, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: local/total

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         3    19.2363
gender     1  18.6320         2     0.6043 1.585e-05
residence  1   0.4692         1     0.1351    0.4934
```

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

Analysis of Deviance Table

Model: binomial, link: logit

Response: local/total

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         3    19.2363
residence  1   7.6161         2    11.6202    0.0058
gender     1  11.4850         1     0.1351    0.0007
```

Order in which terms are fitted is important. Residence is significant without gender, but not significant after allowing for gender. However, gender is significant with or without allowance for residence.

```
> summary(carpref.1)$coef
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.7127250	0.2336783	7.3294146	2.311603e-13
gendermale	-0.8230744	0.2479970	-3.3188883	9.037657e-04
residencecountry	0.1751783	0.2562225	0.6836958	4.941673e-01

No need to include residence in the model, so omit it:

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8245493	0.1703877	10.708222	9.313376e-27
gendermale	-0.9082586	0.2157127	-4.210501	2.548049e-05

Let's see what happens when the constant term is omitted from the model.

```
> carpref.4 <- glm(local/total ~ gender - 1, family = binomial,
+     weight = total, data = carpref.dat)
> summary(carpref.4)$coef
```

```
               Estimate Std. Error   z value     Pr(>|z|)
genderfemale 1.8245493  0.1703877 10.708222 9.313376e-27
gendermale   0.9162907  0.1322876  6.926507 4.313589e-12
```

```
> deviance(carpref.3)
```

```
[1] 0.6043126
```

```
> deviance(carpref.4)
```

```
[1] 0.6043126
```

Since gender is a factor, the models with and without the constant (carpref.3 and carpref.4) are equivalent.

Extended voting behaviour data

Race	Political Views	1980 Presidential Vote	
		Reagan	Carter or Other
White	1	1	12
	2	13	57
	3	44	71
	4	155	146
	5	92	61
	6	100	41
	7	18	8
Non-white	1	0	6
	2	0	16
	3	2	23
	4	1	31
	5	0	8
	6	2	7
	7	0	4

Political views: 1 = extremely liberal; 7 = extremely conservative.

Extended voting behaviour data example: Questions

(a) Fit a logistic model with nominal main effects. Does there seem to be a trend over the seven levels of political views?

(b) Fit a logistic model (or models) that uses the ordinal nature of political views and interpret the parameter estimates for the model(s) in terms of odds ratios.

R analysis

```
> vote.dat <- read.csv("../data/more_voters.csv")
> vote.dat$total <- vote.dat$reagan + vote.dat$carter
> vote.4 <- glm(reagan/total ~ race + factor(views), family = binomial,
+     weight = total, data = vote.dat)
```

R analysis of the extended voting behaviour data

```
> anova(vote.4, test = "Chi")

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                            13    185.157
race            1   95.494        12     89.662 1.483e-22
factor(views)   6   84.701         6      4.961 3.806e-16
```

(a)

```
> vote.5 <- glm(reagan/total ~ factor(views) + race, family = binomial,
+     weight = total, data = vote.dat)
> anova(vote.5, test = "Chi")

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                           13    185.157
factor(views)  6  102.861         7     82.296 6.344e-20
race           1   77.335         6      4.961 1.443e-18
```

```
> summary(vote.5)$coef
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.401179	1.1347250	-4.7599013	1.936876e-06
factor(views)2	1.017621	1.0828496	0.9397623	3.473395e-01
factor(views)3	2.077408	1.0555007	1.9681733	4.904810e-02
factor(views)4	2.564028	1.0448508	2.4539655	1.412905e-02
factor(views)5	2.908747	1.0514272	2.7664747	5.666597e-03
factor(views)6	3.436971	1.0547046	3.2587051	1.119220e-03
factor(views)7	3.251093	1.1152504	2.9151239	3.555474e-03
racewhite	2.886714	0.4706828	6.1330337	8.621888e-10

Both race and political views (treated as a factor) are highly significant. Also, the parameter estimates for 'views' show an increasing trend.

(b)

```
> vote.6 <- glm(reagan/total ~ views * race, family = binomial,
+     weight = total, data = vote.dat)
> anova(vote.6, test = "Chi")

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                         13    185.157
views       1   92.254        12     92.903 7.623e-22
race        1   80.432        11     12.470 3.008e-19
views:race  1    0.174        10     12.297     0.677
```

```
> summary(vote.6)$coef
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4022397	1.4034154	-3.1368044	0.001708001
views	0.3657522	0.3048999	1.1995809	0.230302132
racewhite	2.3576001	1.4286892	1.6501841	0.098905288
views:racewhite	0.1299519	0.3108494	0.4180543	0.675907453

No evidence of (sig.) interaction between race and political views (x).

Is political views as a factor 'better' than as a variable?

```
> vote.7 <- glm(reagan/total ~ views + race + factor(views),
+       family = binomial,  weight = total, data = vote.dat)
> anova(vote.7, test = "Chi")

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                            13    185.157
views         1   92.254         12     92.903 7.623e-22
race          1   80.432         11     12.470 3.008e-19
factor(views) 5    7.509          6      4.961     0.185
```

After fitting race and political views as a variable, adding views as a factor does not produce a significant improvement. Settle for the model "views + race" and find the parameter estimates.

```
> vote.8 <- glm(reagan/total ~ views + race, family = binomial,
+     weight = total, data = vote.dat)
> summary(vote.8)$coef
```

```
              Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -4.9608575 0.53955873 -9.194286 3.774922e-20
views        0.4908373 0.05926852  8.281585 1.215463e-16
racewhite    2.9369082 0.47227992  6.218575 5.016896e-10
```

Model "`views + race`" provides a good fit to the data, though test may not be very reliable due to small frequencies among non-whites.

- exp(2.937)=18.9
  for the same value of 'views', the odds of a white voting for Reagan are 18.9 times those of a non-white.
- exp(0.4908)=1.63
  the odds in favour of a vote for Reagan increase by a factor of 1.63 for each increase in 'views' of one unit.

Winning favourites in racing example

Track		Distance (metres) 1000	1200	1400	1600	2000
Caulfield	W	27	18	22	17	11
	N	51	75	72	72	36
Moonee	W	25	43	13	39	14
Valley	N	67	116	42	136	38
Flemington	W	12	28	40	17	28
	N	42	86	117	74	86
Sandown	W	17	40	21	21	5
	N	62	106	74	70	21

W = number of winning favourites; N = number of races considered.

R analysis of winning favourites data

```
> winners.dat <- read.csv("../data/races.csv")
> winners.dat$track.f <- factor(winners.dat$track)
> winners.dat$dist.f <- factor(winners.dat$dist)
> attach(winners.dat)

        The following object(s) are masked from winners.dat (position 3) :

         dist dist.f number track track.f win


        The following object(s) are masked from package:stats :

         dist

> winners.1 <- glm(win/number ~ track.f + dist.f, family = binomial,
+     weights = number)
```

```
> summary(winners.1)

Call:
glm(formula = win/number ~ track.f + dist.f, family = binomial, weights = number)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6683  -0.4978  -0.1148   0.2576   2.5284

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.594018   0.176668  -3.362 0.000773 ***
track.fFlemington    -0.007903   0.165515  -0.048 0.961918
track.fMoonee Valley  0.140642   0.164714   0.854 0.393185
track.fSandown       -0.006550   0.171939  -0.038 0.969610
dist.f1200           -0.123231   0.176610  -0.698 0.485328
dist.f1400           -0.199171   0.188371  -1.057 0.290361
dist.f1600           -0.468187   0.184929  -2.532 0.011351 *
dist.f2000           -0.183359   0.215244  -0.852 0.394290
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.559  on 19  degrees of freedom
Residual deviance: 16.095  on 12  degrees of freedom
AIC: 121.63

Number of Fisher Scoring iterations: 3
```

```
> anova(winners.1, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: win/number

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       19    24.5593
track.f  3   0.8723         16    23.6870    0.8321
dist.f   4   7.5922         12    16.0949    0.1077
```

The NULL model provides an adequate fit to the data and neither track nor distance make a significant improvement. Fit the NULL model.

```
> winners.2 <- glm(win/number ~ 1, family = binomial, weights = number)
> anova(winners.2, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: win/number

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    19     24.559
```

```
> summary(winners.2)

Call:
glm(formula = win/number ~ 1, family = binomial, weights = number)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6685  -0.7490  -0.1848   0.5907   3.1265

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.76577    0.05656  -13.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.559  on 19  degrees of freedom
Residual deviance: 24.559  on 19  degrees of freedom
AIC: 116.10

Number of Fisher Scoring iterations: 3
```

Let's find the deviance residuals of the Null model:

```
> matrix(resid(winners.2, type = "deviance"), nrow = 4, ncol = 5,
+     byrow = F)

           [,1]      [,2]       [,3]       [,4]       [,5]
[1,]  3.1264581 -1.476993 -0.2165539 -1.5221388 -0.1531267
[2,]  0.9665050  1.216632 -0.1098292 -0.7744511  0.6670468
[3,] -0.4452849  0.162746  0.5652962 -1.6684543  0.1627460
[4,] -0.7405264  1.306572 -0.6274603 -0.3142441 -0.8013694

> matrix(resid(winners.2, type = "deviance")^2, nrow = 4, ncol = 5,
+     byrow = F)

          [,1]       [,2]       [,3]       [,4]       [,5]
[1,] 9.7747404 2.18150688 0.04689560 2.31690661 0.02344780
[2,] 0.9341320 1.48019221 0.01206246 0.59977455 0.44495148
[3,] 0.1982786 0.02648626 0.31955983 2.78373978 0.02648626
[4,] 0.5483794 1.70712942 0.39370643 0.09874937 0.64219296
```

Create a new variable "group" to allow for distance 1 at track 1.

```
> group <- ifelse((track.f == "Caulfield" & dist.f == "1000"),
+     1, 0)
> group
 [1] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> winners.3 <- glm(win/number ~ group, family = binomial, weight = number)
```

```
> anova(winners.3, test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: win/number

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     19    24.5593
group  1  10.1641         18    14.3952    0.0014
```

```
> summary(winners.3)$coef

              Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -0.8018663 0.05797197 -13.831966 1.634899e-43
group        0.9196494 0.28646894   3.210293 1.325995e-03
```

Conclude that "**group**" is highly significant.

An artificial 3-way classification data example

		B	1		2		3	
		C	1	2	1	2	1	2
A								
1	Y		0.000	5.000	3.000	7.000	5.000	12.000
	N		30.00	30.00	30.00	30.00	30.00	30.00
2	Y		14.000	15.000	18.000	22.000	14.000	19.000
	N		30.00	30.00	30.00	30.00	30.00	30.00
3	Y		26.000	28.000	25.000	26.000	30.000	28.000
	N		30.00	30.00	30.00	30.00	30.00	30.00
4	Y		30.000	30.000	30.000	30.000	30.000	30.000
	N		30.00	30.00	30.00	30.00	30.00	30.00

R analysis

```
> y <- c(0, 14, 26, 30, 5, 15, 28, 30, 3, 18, 25, 30, 7, 22, 26,
+     30, 5, 14, 30, 30, 12, 19, 28, 30)
> a.f <- factor(rep(1:4, 6))
> b.f <- factor(rep(1:3, c(8, 8, 8)))
> c.f <- factor(rep(rep(1:2, c(4, 4)), 3))
> n <- 30 + y * 0
> threeway.1 <- glm(y/n ~ a.f + b.f + c.f, family = binomial, weight = n)
```

```
> summary(threeway.1)$coef

              Estimate   Std. Error      z value     Pr(>|z|)
(Intercept) -2.3133908    0.2991414 -7.733436847 1.046813e-14
a.f2         1.8731837    0.2531687  7.398953332 1.372621e-13
a.f3         3.9317014    0.3318233 11.848782138 2.183409e-32
a.f4        23.7579121 2800.0765191  0.008484737 9.932302e-01
b.f2         0.4677535    0.2698162  1.733600549 8.298898e-02
b.f3         0.7210939    0.2721731  2.649394306 8.063619e-03
c.f2         0.6484499    0.2225661  2.913515425 3.573842e-03

> threeway.2 <- glm(y/n ~ a.f + b.f + c.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> summary(threeway.2)$coef

              Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -2.3133908  0.2991384 -7.733514 1.046179e-14
a.f2         1.8731837  0.2531678  7.398981 1.372334e-13
a.f3         3.9317014  0.3318156 11.849056 2.176281e-32
b.f2         0.4677535  0.2698139  1.733616 8.298628e-02
b.f3         0.7210939  0.2721707  2.649418 8.063052e-03
c.f2         0.6484499  0.2225643  2.913540 3.573561e-03
```

```
> anova(threeway.2)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
a.f   2  215.811        15     38.178
b.f   2    7.238        13     30.940
c.f   1    8.687        12     22.254
```

```
> threeway.3 <- glm(y/n ~ b.f + c.f + a.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> anova(threeway.3)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
b.f   2    4.624        15    249.365
c.f   1    5.512        14    243.854
a.f   2  221.600        12     22.254
```

```
> threeway.4 <- glm(y/n ~ c.f + a.f + b.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> anova(threeway.4)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
c.f   1    5.464        16    248.525
a.f   2  218.896        14     29.630
b.f   2    7.376        12     22.254
```

```
> threeway.5 <- glm(y/n ~ c.f + b.f + a.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> anova(threeway.5)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
c.f   1    5.464        16    248.525
b.f   2    4.672        14    243.854
a.f   2  221.600        12     22.254
```

```
> threeway.6 <- glm(y/n ~ b.f + a.f + c.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> anova(threeway.6)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
b.f   2    4.624        15    249.365
a.f   2  218.425        13     30.940
c.f   1    8.687        12     22.254
```

```
> threeway.7 <- glm(y/n ~ a.f + c.f + b.f, family = binomial, weight = n,
+     subset = -c(4, 8, 12, 16, 20, 24))
> anova(threeway.7)

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    17    253.990
a.f   2  215.811        15     38.178
c.f   1    8.549        14     29.630
b.f   2    7.376        12     22.254
```

Changes in deviance using levels 1 to 3 (only) of A

Factor|(model specification)	Change in deviance
$A \mid -$	215.8
$A \mid B$	218.4
$A \mid C$	218.9
$A \mid (B + C)$	221.6
$B \mid -$	4.6
$B \mid A$	7.2
$B \mid C$	4.7
$B \mid (A + C)$	7.4
$C \mid -$	5.5
$C \mid A$	8.5
$C \mid B$	5.5
$C \mid (A + B)$	8.7

- Where, for example, $B \mid (A + C)$ refers to fitting factor B, after $A + C$, in the model.
- Note that the change in deviance due to adding a factor depends on which of the other factors are already in the model.

Over (and under) dispersion

- It is sometimes found that no reasonable model provides an adequate fit to the data.
- One possible explanation for such behaviour is that the assumptions underlying the binomial distribution are not satisfied.
- In this situation, one may assume the response y_i to follow a **quasi-binomial** distribution with

$$\mathbb{E}(y_i) = m_i\pi_i \quad \text{and} \quad \text{Var}(y_i) = \phi m_i\pi_i(1-\pi_i)$$

  where ϕ is a **dispersion** parameter.
- Refer $\phi > 1$ to as **over-dispersion**; and $\phi < 1$ to as **under-dispersion**.
- Estimate ϕ as $\hat{\phi} = X^2/df$ where X^2 is the Pearson goodness of fit statistic and df is the model residual degrees of freedom.

Example: Germination of stipa seeds

- The data came from a study of the effect of two factors: `treatment` (7 levels: seed cutting, pricking, soaking etc.) and age (3 levels: fresh, 6 months, 12 months), on the germination of stipa seeds.
- Each of the 21 `treatment` $\times$ age combinations were replicated 4 or 5 times, and the response variable was the number of (filled) seeds, out of samples of about 25, that germinated within 30 days of sowing.
- Using logistic regression `glm(..., family=quasibinomial,...)` the following results were obtained:

Model	df	D	X^2
–	97	1358	
T	91	342	
A	95	1263	
$T + A$	89	181	178
$T + A + T : A$	77	146	131

Model	df	D	X^2
–	97	1358	
T	91	342	
A	95	1263	
$T + A$	89	181	178
$T + A + T : A$	77	146	131

- None of these models provide an adequate fit to the data, and there appears to be significant interaction between treatment and age.
- A reasonable conclusion here is that there is over-dispersion, due to a lack of independence between seeds within the samples.
- One way to deal with this problem is to assume that there is an extra unknown **dispersion** parameter, an estimate of which is $\hat{\phi} = 1.70 = \frac{131}{77}$ (preferred) or $\hat{\phi} = 1.90 = \frac{146}{77}$.

- Comparisons between models in presence of over-dispersion should then be made using $F-$tests, instead of χ^2 tests.
- Here this approach leads to the conclusion that the interaction between `treatment` and age (i.e. $T : A$) is not significant $\left[F = \frac{(181-146)/12}{146/77} = 1.54 \text{ cf. } F_{12,77}\right]$, but that treatment and age are significant.
- Allowance for the over-dispersion also needs to be made when obtaining the standard errors of parameter estimates, by multiplying the (un-adjusted) standard errors by $\hat{\phi}$.
- The `Std.Error` values shown in `summary(glm(...))` are adjusted already. It is then appropriate to obtain confidence intervals based on the normal distribution approximation.
- "`summary(glm(...))$cov.unscaled`" gives un-scaled variance matrix of $\hat{\boldsymbol{\beta}}$.

Potential causes of over-dispersion include:

1. The systematic component of the model is inadequate in some way. For example, we may need to include (more) interaction terms, or a quadratic effect or transform one or more of the explanatory variables in order for the response to follow a binomial distribution.
2. The presence of one or more outliers in the data set.
3. An inappropriate choice of link function.
4. A sparse data set, i.e. too many "small" (expected) frequencies.
5. Correlation between the binary responses.
6. So called "random effects", which lead to variations in response probabilities within a group.

§4.8 Logistic regression for ungrouped data: Overview

- It is not always possible or appropriate to group data.
  E.g., when no 'subjects' share the same covariate values.
- Need to treat each 'subject' as a separate group, thus $m_i = 1$ for all i.
- It is still possible to use logistic regression, with π_i denoting the probability of success for subject i.
- It is no longer possible to use either the Pearson X^2 statistic or the residual deviance to test the adequacy of a model.
  Hosmer-Lemeshow statistic is used here instead.
- **But**, the χ^2 test for the **change in residual deviance** between two (nested) models is still valid.

Example: Compliance with screening test for bowel cancer

- Interested in how the probability of **compliance** is related to various factors/covariates.
- **Factors/covariates**: Sex, age (in years), smoking (2 levels), symptoms (3 levels), family history, etc.
- Many interactions between sex and other factors were significant and it was decided to carry out separate analyses for males and females.

Results for females

Model	residual deviance	df
—	439.2	318
Age	424.4	317
Smoking	428.0	317
Symptoms	434.8	316
Smoking + Age	416.6	316
Smoking + Age + Age.Smoking	416.3	315
Smoking + Age + Symptoms	413.3	314

Parameter	Estimate	Standard error	odds ratio [$=e^{\text{estimate}}$]
1	0.655	0.29	
smok(2) [no]	0.824	0.27	2.28
age (in years)	0.042	0.013	1.043 1.52 (for 10 years)

Example: Attitudes towards non means-tested age pension

A survey of attitudes towards the introduction of a non means-tested age pension resulted in the following data.

	Age				
	20	30	40	50	60
Sample size	12	12	12	12	12
Number in favour	4	6	9	11	12

(A) Fit a (straight line) logistic regression model to the data.

(B) Carry out a formal test of whether a quadratic logistic regression model provides a better fit to the data than the straight line model.

(a) Analysis with the data treated as grouped data

```
> y <- c(4, 6, 9, 11, 12)
> n <- c(12, 12, 12, 12, 12)
> age <- c(20, 30, 40, 50, 60)
```

(a) Analysis as grouped data

```
> anova(glm(y/n ~ age, family = binomial, weight = n), test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     4    21.0117
age   1  19.9825         3     1.0292 7.815e-06
```

```
> summary(glm(y/n ~ age, family = binomial, weight = n))$coef

              Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -3.1576993 1.06970977 -2.951922 0.0031580317
age          0.1112453 0.03093199  3.596447 0.0003225927
```

(a) Analysis as grouped data

```
> anova(glm(y/n ~ age + I(age^2), family = binomial, weight = n),
+       test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: y/n

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          4    21.0117
age        1  19.9825         3     1.0292 7.815e-06
I(age^2)   1   0.7521         2     0.2771    0.3858

> summary(glm(y/n ~ age + I(age^2), family = binomial, weight = n))$coef

                Estimate  Std. Error    z value  Pr(>|z|)
(Intercept) -0.426533579 3.412301296 -0.1249988 0.9005245
age         -0.061121914 0.211229868 -0.2893621 0.7723043
I(age^2)     0.002462193 0.003070851  0.8017951 0.4226715
```

(b) Analysis with the data treated as ungrouped data.

```
> y.1 <- c(1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,
+      0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1,
+      1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
+      1, 1, 1, 1)
> age.1 <- 10 * (rep(2:6, c(12, 12, 12, 12, 12)))
```

(b) Analysis as ungrouped data.

```
> anova(glm(y.1 ~ age.1, family = binomial), test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: y.1

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     59     73.304
age.1  1   19.983        58     53.321 7.815e-06

> summary(glm(y.1 ~ age.1, family = binomial))$coef

             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -3.1576993  1.0697077 -2.951927 0.0031579727
age.1        0.1112453  0.0309319  3.596458 0.0003225798
```

(b) Analysis as ungrouped data.

```
> anova(glm(y.1 ~ age.1 + I(age.1^2), family = binomial), test = "Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: y.1

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         59     73.304
age.1       1   19.983        58     53.321 7.815e-06
I(age.1^2)  1    0.752        57     52.569     0.386
```

```
> summary(glm(y.1 ~ age.1 + I(age.1^2), family = binomial))$coef

                Estimate  Std. Error    z value  Pr(>|z|)
(Intercept) -0.426533594 3.412221139 -0.1250017 0.9005222
age.1       -0.061121913 0.211223186 -0.2893712 0.7722973
I(age.1^2)   0.002462193 0.003070725  0.8018279 0.4226525
```

(b) Analysis as ungrouped data.

For analysis of the ungrouped data, allowing for rounding errors, the parameter estimates and their standard errors, and the changes in residual deviance are **(exactly) the same** as for the analysis of the grouped data.

Pensions example: residual deviances (df)

Model	Grouped data	Ungrouped data
μ	21.012 (4)	73.304 (59)
$\beta_0 + \beta_1 \text{age}$	1.029 (3)	53.321 (58)
$\beta_0 + \beta_1 \text{age} + \gamma \text{age}^2$	0.277 (2)	52.569 (57)

Pensions example: parameter estimates (se)

Model	Parameter	Grouped data	Ungrouped data
$\beta_0 + \beta_1 \text{age}$	β_0	-3.158 (1.070)	-3.158 (1.070)
	β_1	0.1112 (0.0309)	0.1112 (0.0309)
$\beta_0 + \beta_1 \text{age}$	β_0	-0.427 (3.412)	-0.427 (3.412)
$+\gamma \text{age}^2$	β_1	-0.0611 (0.2112)	-0.0611 (0.2112)
	γ	0.0025 (0.0031)	0.0025 (0.0031)

Remarks

1. For grouped data, both the Pearson X^2 statistic and the residual deviance can be used to test the adequacy of fit of a model.
2. Only the residual deviance can be used to compare two models, with grouped or ungrouped data.
3. The R **'step'** procedure can be applied to generalised linear models as follows:

```
> step(glm.object)
```

Remarks

4 The 'goodness-of-fit' of a logistic regression model fitted to ungrouped data can be tested using the 'Hosmer-Lemeshow' test.

The test statistic is a X^2 type statistic $\left(\sum \frac{(O-E)^2}{E}\right)$ for a $2 \times g$ contingency table where the rows correspond to success ($y = 1$) and failure ($y = 0$) and the columns are g intervals for the estimated success probabilities.

The test statistic is distributed approximately as χ^2_{g-2} when the fitted model is the true model.

The Hosmer-Lemeshow test has already been introduced in Chapter 2.