Analysing categorical data using logit models

Graeme Hutcheson, University of Manchester

The lecture notes, exercises and data sets associated with this course are available for download from:

www.Research-Training.net/Manchester

These notes cover the basic logit model for modelling categorical data. These notes are comprehensive and provide reference material for the technique of logistic regression and how it can be computed using R and Rcmdr.

Logistic regression

Data that are constrained are relatively common and may consist of dichotomous classifications such as ‘survived–died’, ‘waged–unwaged’, ‘employed–unemployed’, ‘succeed–fail’, or measures of proportion constrained between 0 and 1 (proportion killed, proportion who succeed, the proportion of questions correctly answered, score on an examination, etc.). Although the response variable is categorical in this case, the form of the model is identical to OLS regression as we are still just attempting to model (or predict) a single response variable using other variable(s). For example,

Variable Y can be predicted by Variable X

To demonstrate modelling a binary response variable we use a hypothetical data set showing the success of a treatment for a particular disease given the severity of the infection when the treatment was started and the hospital where the treatment was administered. The response variable is binary and indicates two different treatment outcomes (those patients who survived and those who died); one explanatory variable is continuous and indicates the severity of the infection (in this case, by measuring the concentration of bacteria present in a measure of blood) and the other is unordered categorical and indicates one of three hospitals (A, B or C) where the treatment was administered. These data are provided on Statlib (http://lib.stat.cmu.edu/datasets) in a zipped file named hutsof99.zip (the file you want is tab4_01.dat - a plain text data file)1.

---

1The files in Statlib are the data sets that have been used in Hutcheson and Sofroniou, 1999.
Table 1: Treatment Outcome as a Function of Infection Severity and Hospital (raw data)

<table>
<thead>
<tr>
<th>Infection</th>
<th>Treatment</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.3</td>
<td>survived</td>
<td>C</td>
</tr>
<tr>
<td>18.2</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>22.7</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>32.9</td>
<td>survived</td>
<td>C</td>
</tr>
<tr>
<td>38.0</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>39.9</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>44.0</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>44.9</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>46.8</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>47.7</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>49.1</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>49.7</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>50.9</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>50.9</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>53.7</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>60.1</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>61.7</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>62.2</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>64.1</td>
<td>died</td>
<td>A</td>
</tr>
<tr>
<td>68.9</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>71.2</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>72.3</td>
<td>died</td>
<td>B</td>
</tr>
<tr>
<td>73.0</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>73.2</td>
<td>died</td>
<td>B</td>
</tr>
<tr>
<td>74.6</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>75.3</td>
<td>died</td>
<td>B</td>
</tr>
<tr>
<td>76.3</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>77.7</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>79.1</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>80.6</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>85.0</td>
<td>survived</td>
<td>C</td>
</tr>
<tr>
<td>86.0</td>
<td>died</td>
<td>A</td>
</tr>
<tr>
<td>89.4</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>90.0</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>90.9</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>91.5</td>
<td>survived</td>
<td>B</td>
</tr>
<tr>
<td>96.6</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>97.2</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>101.6</td>
<td>survived</td>
<td>A</td>
</tr>
<tr>
<td>104.8</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>108.6</td>
<td>died</td>
<td>A</td>
</tr>
<tr>
<td>109.9</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>111.3</td>
<td>died</td>
<td>B</td>
</tr>
<tr>
<td>113.2</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>118.8</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>127.0</td>
<td>died</td>
<td>B</td>
</tr>
<tr>
<td>142.0</td>
<td>died</td>
<td>C</td>
</tr>
<tr>
<td>157.4</td>
<td>died</td>
<td>A</td>
</tr>
<tr>
<td>169.2</td>
<td>died</td>
<td>C</td>
</tr>
</tbody>
</table>

An initial analysis will model treatment outcome using infection severity as an explanatory variable. This can be shown using the following descriptive model representation:

Treatment Outcome can be predicted by severity of infection

It is useful at this point to consider whether a linear model (OLS regression) may be appropriately applied to this research question. Figure 1 shows an OLS regression model applied to these data (the thick diagonal line drawn on the graph). One can see clearly see that the data are constrained between 0 and 1 on the Y axis, but the regression model is not. For these data, an OLS regression model provides a poor picture of the relationship between the two variables and gives some predictions that are clearly impossible. For example, at an infection of 0, the probability of someone dying is about -0.2 (this is impossible), and at an infection severity of about 250, the probability of dying is 2 (this is also impossible).

It is clear that using OLS regression to model a binary response variable is not appropriate. If we look closely at the two variables we can see that the relationship between them is non-linear. This can be seen in Figure 2 which shows the proportion of patients who die at each infection severity. This graph shows that at very low infections, practically everyone survives, whereas at very high infections practically everyone dies. The relationship between the variables ‘flattens out’ at each extreme of infection severity producing an S-shaped curve which is constrained to lie between 0 and 1 (this shape is also known as a sigmoid\(^2\).

\(^2\)Further examples of S-shaped distributions can be found in Agresti, 1996, and Collett, 1991.
Figure 1: An inappropriate model: Modelling the relationship between infection severity and outcome using OLS regression

Figure 2: Modelling the relationship between infection severity and outcome using proportions
This non-linear relationship presents some problems for the analyst as it is not possible to describe the relationship between the variables simply. The descriptive model proposed above...

Treatment Outcome can be predicted by severity of infection

suggests that we want a single parameter for the effect of severity on outcome (i.e., is severity associated with outcome and how significant is this association?). An S-shaped model of probability does not allow us to answer these questions simply. Such a model shows that a unit increase in severity is associated with different changes in outcome probability depending on the level of severity. For example, a unit increase in severity is not strongly related to a change in the fitted outcome at high or low levels of severity (the slope of the model is almost 0 at the extremes). In the middle of the sampled severity readings, however, a unit increase in severity results in a much bigger change in the fitted outcome. In order to answer questions such as ‘Is outcome associated with infection severity?’ and ‘How significant is the relationship between severity and treatment outcome?’ it would be useful to have a single parameter to describe the relationship. A linear model would provide just such a parameter and would be a useful way of modelling the relationship, provided that the relationship between the variables is represented in a linear form. This is essentially what logistic regression does; it estimates the parameters of a linear model that is derived from the data and then transforms this into a non-linear S-shaped model of probability. The process of logistic regression can be seen as an exercise in data transformation; a linear model (the linear predictor) is transformed into a non-linear model (the fitted values of probability) which is more suited to the description of a binary response variable. The transformation in this case is the log-odds, which is commonly known as the logit.

1 The Regression Equation

In order to explain how a linear model ($Y = \alpha + \beta x$) can be used to describe a non-linear relationship (the S-shaped curve), it is useful to look in some detail at the equations and the transformations required to convert between a non-linear model of probability and a linear model that can be simply estimated from the data.

1.1 Modelling probability

The probability that a particular outcome will occur (for our example, this is the probability that a patient will die) can be modelled using Equation 1.

$$\text{PROBABILITY: } Pr(\text{death}) = \frac{e^{(\alpha + \beta x)}}{1 + e^{(\alpha + \beta x)}}$$

where $Pr(\text{death})$ is the probability that the outcome is ‘death’ (this may also be represented as ‘$Y=1$’),
e is the natural logarithm base,
$\alpha$ and $\beta$ are parameters of the linear component of the model,
and $x$ is the value of the explanatory variable.
The relationship between the explanatory variable and the probability that a patient will die is non-linear and this needs to be transformed so that it can be described using a linear model. The transformation of the S-shaped function to a linear one looks complicated, but is relatively straightforward. Firstly, one needs to transform the measure of probability into a measure of odds, which is defined as the ratio of the probability of an event occurring to the probability of an event not occurring.

1.2 Modelling odds

The odds of a patient dying can be represented algebraically as shown in Equation 2.

\[
\text{odds(death)} = \frac{\text{Pr}(\text{death})}{1 - \text{Pr}(\text{death})}
\]  

(2)

where \(\text{Pr}(\text{death})\) is the probability that the outcome is death, and \((1 - \text{Pr}(\text{death}))\) is the probability that the outcome is not death.

For our model of treatment outcome, the odds that the outcome will be death is given in Equation 3.

\[
\text{ODDS: odds(death)} = e^{(\alpha + \beta x)}
\]

(3)

The odds of death are constrained to assume any value above 0 (see Figure 3); a relationship that is also non-linear. It is, however, a simple matter to represent this linearly by taking the natural log of the odds, which is commonly known as the logit.

1.3 Modelling log-odds

Logit[\(\text{Pr}(\text{death})\)] has a linear relationship with the explanatory variable and is represented algebraically in Equation 4.

\[
\text{LOG-ODDS: log-odds[Pr(death)] = logit[Pr(death)] = } \alpha + \beta x
\]

(4)

Equations 1 to 4 demonstrate how a non-linear relationship between a binary response variable and a continuous explanatory variable can be represented linearly when the response variable is described in terms of the log-odds of the probability of a particular outcome. Using logit[\(\text{Pr}(\text{death})\)] instead of \(\text{Pr}(\text{death})\) enables linear parameters to be estimated from the data and the relationship between the variables to be modelled using a regression analysis. The transformation of the S-shaped relationship between outcome and severity of infection to the straight line relationship between logit[\(\text{Pr}(\text{outcome})\)] and severity is shown in Figure 3.

---

3Logit[\(\text{Pr}(Y=1)\)] is also commonly known as logit[\(\pi(x)\)], where \(\pi\) = the probability of \(x\) and not 3.1416... , see Agresti, 1996; Hosmer and Lemeshow, 1989.
Figure 3: Transforming an S-shaped distribution of probability into a linear distribution
Probabilities, odds and logits - moving between them

It is a relatively simple matter to convert probabilities into odds and logits and logits back into odds and probabilities. Table 2 shows a number of probabilities converted into logits and a number of logits that have been converted back into probabilities. These calculations can be computed manually, or they can be computed automatically using R.

---

**Working with Probabilities, odds and logits**

Libraries required: `boot`.

**R console: commands**

```r
# Load the package 'boot'
library(boot)

# Computing a logit from a probability can be achieved using the logit(x) function.
logit(0.1)
# answer: -2.197225  # a probability of 0.1 has a logit value of -2.197

logit(0.5)
# answer: 0

logit(0.9)
# answer: 2.197225

# Computing a probability from a logit can be achieved using the inv.logit(x) function.
inv.logit(-3)
# answer: 0.04742587  # a logit of -3 has a probability of 0.047

inv.logit(0)
# answer: 0.5

inv.logit(3)
# answer: 0.9525741

# Computing an odds from a logit can be achieved using the exp(x) function.
exp(2.13)
# answer: 8.414867  # a logit of 2.13 has an odds of 8.415

exp(-0.413)
# answer: 0.6616623
```
Table 2: The logit transformation

<table>
<thead>
<tr>
<th>probability - logit</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x/(1 - x)</td>
<td>log[x/(1 - x)]</td>
</tr>
<tr>
<td>0.1</td>
<td>0.11</td>
<td>-2.20</td>
</tr>
<tr>
<td>0.5</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.9</td>
<td>9.00</td>
<td>2.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>logit - probability</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>log odds</td>
<td>odds</td>
<td>probability</td>
</tr>
<tr>
<td>x</td>
<td>$e^x$</td>
<td>$e^x/[1 + e^x]$</td>
</tr>
<tr>
<td>-3.00</td>
<td>0.05</td>
<td>0.33</td>
</tr>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>0.5</td>
</tr>
<tr>
<td>3.00</td>
<td>20.09</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Note: Do not worry too much about these transformations - they are usually performed automatically by the software package. It is useful, however, to know how they can be computed.

2 Calculating and Interpreting Model Parameters

The logistic regression procedure estimates the parameters of a linear model (parameters relating to $\text{logit}[\Pr(Y=1)]$). The parameters of the model are therefore interpreted in much the same way as they are in OLS regression with the gradient, or slope of the line, indicating the parameter $\beta$ which is interpreted as the change in the log odds of the outcome being in category 1 (compared to 0) that is associated with a unit change in $x$. Put another way, $\beta$ represents the change in the log-odds of the probability of dying (provided that the death category is coded using a higher number) for a unit increase in severity. The parameter $\alpha$ indicates the value of $\text{logit}[\Pr(\text{death})]$ when severity = 0.

Selecting the category to model

First, it is important to know which category we are predicting. This category can be easily identified and changed if needs be using Rcmdr using the procedure shown below.
Identifying the outcome category to be modelled

Data set: Infection.csv (available for download from www.Research-Training.net/Manchester)

Rcmdr: commands

Data ▼
Manage variables in active data set ▸
Reorder factor levels...

Reorder Factor Levels
Factor (pick one) pick Outcome
Name for factor: provide a name or leave as ‘Outcome’
OK
Give permission to over-write variable if asked.
Yes

Reorder Levels
old levels died, survived
new order category to be modelled = 2, other category = 1
OK

If you want to model the category ‘died’, then code this as 2, and ‘survived’ as 1. If you want to
model the category ‘survived’, then code this as 2, and ‘died’ as 1. If you are likely to be modelling
both categories, it is sometimes useful to save the new code to a different variable name. Table
3 shows the parameter estimates for the model of treatment outcome (when the category ‘died’ is
being modelled).

Table 3: Logistic Regression Model Parameters

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>the odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>e^β</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>−4.64050</td>
</tr>
<tr>
<td>Severity</td>
<td>0.05921</td>
</tr>
</tbody>
</table>

The statistical model of treatment outcome is:

\[
\text{logit}(\text{death}) = -4.64050 + 0.05921 \times \text{Severity}
\]

From Table 3 it can be seen that \( \beta = 0.05921 \) and \( \alpha = -4.64050 \). As the explanatory variable
increases by 1, \( \text{logit(Pr(death))} \) increases by 0.05921. When severity is equal to 0, the value of
\( \text{logit(Pr(death))} \) equals −4.64050. It is, however, difficult to interpret the regression coefficient \( \beta \), as
we do not tend to think in terms of logits. An easier interpretation can be achieved by using the
odds-ratio \( (e^\beta) \) which represents the change in the odds of dying which is associated with a unit
change in severity. For the example above, \( \beta = 0.05921 \), and \( e^{0.05921} = 1.060998 \) which indicates
that for each unit increase in severity the odds of dying change from 1 to 1.06, an increase of 6%. The
odds-ratio is particularly useful as it is a constant and allows a simple interpretation to be given
to the relationship between the response and explanatory variables.
3 Goodness-of-fit Measures

The overall goodness of fit and the significance of each parameter and group of parameters in a logit model can be derived using essentially the same underlying theory and statistics as were used for OLS regression. The basic statistic used is the residual, which is simply calculated by comparing the observed and predicted values of $Y$. For logit models, the deviance-residual is used which makes use of a simple transformation of the raw residual to help interpretation. The sum of all the deviance-residuals represents the deviance and is used to compare nested models to ascertain significance of single and groups of parameters. The following description will show how the residuals are computed and how these are used to compute a measure of deviance that is used to ascertain significance.

3.1 Computing residuals for logit models

It is useful to demonstrate how the residuals are computed for logit models\(^4\). Although the theory behind the computations and use of the residuals is the same for all models, their computation will be demonstrated for null, simple and multi-variable models.

3.1.1 The null model: $Y \sim \alpha$

If the only information we had was outcome (whether someone died or not), the predicted value of outcome would simply be the probability of someone picked at random being in one or other of the outcome groups. The regression model of this would be...

$$\Pr(\text{death}) = \alpha$$

where $\alpha$ is the probability of being in the died category. For our sample, there are 23 who died and 26 who survived. The probability of dying is therefore...

$$\frac{23}{23+26} = 0.4693878$$

The predicted probability of dying can be confirmed using Rcmdr. First, run a null model of outcome (insert the value 1 to represent the explanatory variables in the model)...\(^4\)

\(^4\)Knowing such things is a source of comfort to some people. If you are not one of these, feel free to skim over this section; all you need to realise is that the computation of the residuals (and by inference, all the model-fit statistics we are to discuss) is based on the same principles as the computation of residuals and model-fits in OLS regression.
Computing a logistic regression model in R

Data set: Infection.csv (available for download from www.Research-Training.net/Manchester)

Rcmdr: commands

Statistics ▼
Fit models ►
Generalized linear model...

Generalized Linear Model
Variables (double click to formula): click on: Outcome and enter 1
Model Formula: this should read: Outcome ~ 1
Family (double click to select): click on: binomial
link function: should read: logit
OK

Rcmdr: output

Call:
  glm(formula = Outcome ~ 1, family = binomial(logit), data = infection)

Coefficients:
  Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -0.1226    0.2863  -0.428   0.668

 Null deviance: 67.745 on 48 degrees of freedom
Residual deviance: 67.745 on 48 degrees of freedom
AIC: 69.745

Now save the fitted values from the model and then view the new variables that have been saved to the data set. Looking at the values under ‘fitted.model’, we can see that the predicted probability of being in the ‘died’ category is 0.4693878, which is identical to the probability calculated above.
Predictions for a logistic regression model...

Compute a logistic regression model of outcome (see above).

Rcmdr: commands

Models ▼
Add observation statistics to data...
Add Observation Statistics to Data
   Fitted values: tick box
   Residuals: tick box
   OK

Rcmdr: output

The fitted values (predictions) are added to the data set. To see these, simply view the data set:

![Data set with fitted values and residuals](image)

The raw residuals

The residuals for this model may be computed by comparing the predicted values with the observed values of Y (Y − Ỹ). This calculation simply results in two different values for the residuals - one when the actual value of Y is 0 and one when it is 1. These simple ‘raw-residuals’ may be calculated as...

When Y=0 (Outcome = survived), the raw-residual = (0 - 0.4693878) = -0.4693878

When Y=1 (Outcome = died), the raw-residual = (1 - 0.4693878) = +0.5306122

These raw-residuals do not, however, provide an easy distribution to work with (see Collett, 1991 and Hoffmann, 2004) and are usually transformed and reported as deviance residuals.
The deviance residuals

The deviance residuals are computed from the raw-residuals using equations 5 and 6 (see Hosmer, Taber and Lemeshow, 1991).

\[
\text{when } Y=1, \text{ the deviance residual } = \text{sign} \sqrt{2 \times \log(fittedY)} \quad (5)
\]

\[
\text{when } Y=0, \text{ the deviance residual } = \text{sign} \sqrt{2 \times \log(1 - fittedY)} \quad (6)
\]

The 'sign' just means that the sign of the raw residual is preserved (i.e., all residuals when \(Y=0\) are negative, whilst all residuals when \(Y=1\) are positive).

**When \(Y=1\) (Outcome=died)**

The calculation of the deviance residuals when \(Y=1\) (died) is...

\[
\text{sign} \sqrt{2 \times \log(0.4693878)}
\]

Which can be computed in the **R console** using the commands: \(^5\)

\[
\text{sqrt(abs(2*log(0.4693878)))}
\]

which is equal to 1.2299 (preserving the postive sign of the raw residual).

**When \(Y=0\) (survived)**

When \(Y=0\) (survived), the deviance residual is...

\[
\text{sign} \sqrt{2 \times \log(1 - 0.4693878)}
\]

which can be computed in **R** using the code

\[
\text{sqrt(abs(2*log(abs(1-0.4693878))))}
\]

and equals -1.1258 (preserving the minus sign of the raw residual). These values are exactly the same as those that are computed by **Rcmdr** when it is asked to provide residuals (see previous **Rcmdr** output).

---

\(^5\)The command **abs** indicates the absolute value and removes any minus signs.
3.1.2 A simple model: \( Y \sim \alpha + \beta x \)

Residuals for a simple logistic regression model are computed in a similar fashion as they were for a null model. Specifically, the difference between the observed and fitted values is computed and then transformed into a deviance residual. The fitted values for a simple logistic regression model can be computed by software. The procedure in **Rcmdr** is shown below:

### Computing a logistic regression model in R

Data set: **Infection.csv** (available for download from [www.Research-Training.net/Manchester](http://www.Research-Training.net/Manchester))

**Rcmdr: commands**

```r
Statistics ▼
Fit models ►
  Generalized linear model...
    Generalized Linear Model
    Variables (double click to formula): click on: Outcome and Severity
    Model Formula: this should read: Outcome ~ Severity
    Family (double click to select): click on: binomial
    link function: should read: logit
    OK
```

**Rcmdr: output**

```
Call: glm(formula = Outcome ~ Severity, family = binomial(logit), data = infection)

Coefficients:                  Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)             -4.64050   1.38335   -3.355 0.000795 ***
Severity                0.05921   0.01758    3.368 0.000756 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 67.745 on 48 degrees of freedom
Residual deviance: 45.994 on 47 degrees of freedom
AIC: 49.994
```

The process of saving fitted values and residuals for the model is the same as given above. Use the **Rcmdr** menu options ‘Models, Add observation statistics to data, and then select ‘Fitted values’ and ‘Residuals’.

**When Y=1 (Outcome=died)**

The calculation of the deviance residuals when \( Y=0 \) (survived) at different levels of severity can be derived by inserting the fitted value into the deviance-residual equation. When severity = 9.3, the fitted value is 0.01646564 and this equates to a deviance-residual of...

\[
\text{sign} \sqrt{2 \times \log(1 - 0.01646564)}
\]
Which is -0.1822235, after preserving the sign.

The fitted value when severity 49.1 is equal to 0.15014806, which results in a deviance-residual when Y=1 of...

\[
\text{sign}\sqrt{2 \times \log(1 - 0.15014806)}
\]

\[
> \sqrt{\text{abs}(2 \times \text{log}(0.15014806))}
\]

[1] 1.947374

The residuals computed above are the same as those given by Rcmdr.

3.1.3 A multivariable model: \( Y \sim \alpha + \beta_1 x_1 + \ldots + \beta_k x_k \)

It is a simple matter to compute residuals for models with multiple explanatory variables. Including ‘hospital’ in the model of treatment outcome...

**Descriptive Model:** Outcome ~ Severity + Hospital

**Statistical Model:** \( \text{logit} [\text{Pr(Outcome} = 1)] = \alpha + \beta_1 \text{ Severity} + \beta_2 \text{ Hospital B} + \beta_3 \text{ Hospital C} \)

The fitted values and deviance-residuals for this model are easily computed using Rcmdr and can be seen in Figure 5.
When Y=0

The fitted value for Outcome when Severity = 22.7 at Hospital A is 0.008332772. The deviance residual for this is:

\[
\text{sign} \sqrt{2 \times \log(1 - 0.008332772)}
\]

\[
> \text{sqrt(abs(2*log(1-0.008332772)))}
\]

[1] 0.1293652

Which is -0.1293652 after preserving the sign.

3.1.4 When y=1

The fitted value for Outcome when Severity = 49.1 at Hospital C is 0.556305651. The deviance residual for this is:

\[
\text{sign} \sqrt{2 \times \log(0.556305651)}
\]

\[
> \text{sqrt(abs(2*log(0.556305651)))}
\]

[1] 1.082993

3.2 Computing the deviance for logit models

A useful aspect of deviance residuals is that they can be used to calculate the overall deviance in the model. This is particularly useful as it forms the basis for testing variables and groups of variables for significance. The deviance for a logistic regression model is simply to sum of all the squared deviances.

\[
D = \sum_{i=1}^{n} d_i^2
\]  

(7)
When the deviance-residuals are squared and summed for the null model, they give the deviance value for the model. This is easy to illustrate using the regression model above. For the data shown in Figure 5, the deviance can be computed by simply adding up all the squared residuals (which were saved as ‘residuals.model’). This can be achieved easily in R console.

```R
> sum(infection$residuals.model^2)
[1] 34.74169
```

The value of 34.74169 is the deviance value quoted for this model in the logistic regression analysis.

Call:
```
glm(formula = Outcome ~ Severity + Hospital, family = binomial(logit),
    data = infection)
```

Coefficients:
```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.1886     1.8181  -3.404 0.000664 ***
Severity        0.0621     0.0198    3.128 0.001760 **
Hospital[T.B]   0.9831     1.0125    0.971 0.331595
Hospital[T.C]  3.3663     1.2023    2.800 0.005113 **
```

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 67.745 on 48 degrees of freedom
Residual deviance: 34.742 on 45 degrees of freedom

Using the deviance to determine significance

An indication of the goodness-of-fit of a model can be obtained by comparing its deviance value with the deviance for the null model. This statistic, which we will call $-2LL_{diff}$, is shown in Equation 8.

$$-2LL_{diff} = (-2LL_0) - (-2LL_1)$$

where $-2LL_0$ is the measure of deviance in the null model logit($p$) = $\alpha$, and $-2LL_1$ is the measure of deviance in the model logit($p$) = $\alpha + \beta x$.

This equation simply compares the full model (including all the variables) with the null model (including none of the variables). The difference in $-2LL$ between the models indicates the effect that the explanatory variable has on the response variable. To assess the significance of variable $x_1$...

$$-2LL_{diff} = \begin{cases} Y = \alpha + \beta_1 x_1 \\ Y = \alpha \end{cases}$$
The change in the value of $-2LL$ represents the effect that the explanatory variable has on the deviance in the model, an effect which can be evaluated for significance using the $\chi^2$ distribution with the number of degrees-of-freedom equal to the difference in the number of terms between the two models. For the example above, the effect that variable $x$ has on the model can be assessed using the $-2LL_{\text{diff}}$ statistic with 1 degree-of-freedom.

Statistics based on $-2LL$ have a close relationship with the goodness-of-fit statistics used in OLS regression. In fact, the residual sum of squares, which is a measure of deviance in OLS regression, can be viewed as being analogous to $-2LL$ which is a measure of deviance for the logit link. Similarly, the $F$-statistic used in OLS regression can be seen as being analogous to the $\chi^2$ statistic used in logistic regression (see Table 4).

Table 4: OLS and Logistic Regression Goodness-of-fit Statistics

<table>
<thead>
<tr>
<th>Measure of Deviance</th>
<th>Reference Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual Sum of Squares (RSS)</td>
<td>$F$</td>
</tr>
<tr>
<td>log-likelihood ($-2LL$)</td>
<td>$\chi^2$</td>
</tr>
</tbody>
</table>