Logistic regression

Maths and Statistics Help Centre

Many statistical tests require the dependent (response) variable (y) to be continuous so a different set of tests are needed when the dependent variable is categorical. One of the most commonly used tests for categorical variables is the Chi-squared test which looks at whether or not there is a relationship between two categorical variables but this doesn't make an allowance for the potential influence of other explanatory (independent) variables on that relationship. For continuous outcome variables, multiple regression can be used for

a) controlling for other explanatory variables when assessing relationships between a dependent variable and several independent variables
b) predicting outcomes of a dependent variable using a linear combination of explanatory (independent) variables

Logistic regression does the same but the outcome variable (y) is binary and leads to a model which can predict the probability of the binary event happening for an individual.

The maths:

For multiple regression a model of the following form can be used to predict the value of a response variable y using the values of a number of explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_q x_q + \varepsilon$$

$$\beta_0 = \text{constant/ intercept}\ , \beta_1, \ldots, \beta_q = \text{coefficients for q explanatory variables } x_1, \ldots, x_q$$

The regression process finds the coefficients which minimise the squared differences between the observed and expected values of y (minimising the residuals). As the outcome of logistic regression is binary, y needs to be transformed so that the regression process can be used. The logit transformation gives the following:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_q x_q$$

p = probabilty of event occuring e.g. person dies following heart attack, $\quad \frac{p}{1-p}$ = odds ratio

If probabilities of the event of interest happening for individuals are needed, the logistic regression equation can be written as: $p = \dfrac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_q x_q\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_q x_q\right)}$, for $0 < p < 1$

Titanic example: On April 14th 1912 the Titianic sank. Only 705 passengers and crew out of the 2228 on board survived. Information on 1309 of those on board will be used to demonstrate logistic regression. The data can be downloaded from *biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls*

Research Question: Using the information on the 1309 people on board the titanic, which factors are most important in the survival of the person on board?

The key variables of interest are:

- Dependent variable: Survival - whether a passenger survived (1) or not (0).
- Possible explanatory variables: Age, gender (recode so that sex = 1 for females and 0 for males), class (pclass = 1, 2 or 3), number of accompanying parents/ children (parch) and number of accompanying siblings/ spouses (sibsp)

Initial analysis

The titanic data will be used to fit a logistic regression model. Firstly the data needs to be downloaded and saved as a comma separated file (.csv), although *R* can also read Minitab, SPSS and Excel files. To read into *R*:

```
> titanic_data <- read.csv("C:/... /titanic_data.csv")
> attach(titanic_data)
```

To check the variables, use structure function

```
> str(titanic_data)
'data.frame':   1309 obs. of 20 variables:
 $ pclass      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ survived    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ name        : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 47 55 98 104 125 131 135 152
...
```

Then, declare the categorical explanatory variables as factors (R automatically treats numerical values as integers).

```
> pclass.f <- factor(titanic_data$pclass)
> Gender.f <- factor(Gender)
```

Most of the variables can be investigated using `table(...,titanic_data$survived)`. Another reason for the cross tabulation is to identify categories with small frequencies as this can cause problems with the logistic regression procedure. The number of accompanying parents/ children (parch) and number of accompanying siblings/ spouses (sibsp) were used to create a new binary variable indicating whether or not the person was travelling alone or with family (1 = travelling with family, 0 = travelling alone). To create a new binary variable `alone` use the following commands.

```
alone<-rep(0,1309) # Initialise the variable alone
for (i in 1:1309) {
   if ((parch[i]>0) | (sibsp[i]>0))
    alone[i]<-1
}
```

Then, declare it as a factor

```
alone.f=factor(alone)
```

When tested separately, Chi-squared tests concluded that there was evidence of a relationship between survival and gender, class and whether an individual was travelling alone.

```
> chisq.test(survived,pclass)
        Pearson's Chi-squared test
data:  survived and pclass
X-squared = 127.8592, df = 2, p-value < 2.2e-16
> chisq.test(survived,sex)
        Pearson's Chi-squared test with Yates' continuity correction
data:  survived and sex
X-squared = 363.6179, df = 1, p-value < 2.2e-16
> chisq.test(survived,alone)
        Pearson's Chi-squared test with Yates' continuity correction
data:  survived and alone
X-squared = 52.4183, df = 1, p-value = 4.485e-13
```

Looking at the percentages of survival, it's clear that women, those in first class and those not travelling alone were much more likely to survive.

	Male	Female	1st class	2nd Class	3rd class	Travelling alone	Travelling with family
% surviving	19.1%	72.7%	61.9%	43%	25.5%	30.3%	50.3%

Logistic regression

Logistic regression will initially be carried out using these three variables. Stage 1 of the following analysis will relate to using logistic regression to control for other variables when assessing relationships and stage 2 will look at producing a good model to predict from.

In *R* the command for logistic regression is **`glm (`**y**`~`** x_1 **`+`** x_2 **`+..., data =`** `dataset,` **`family = binomial).`**

Treatment of categorical explanatory variables

When interpreting the output for logistic regression, it is important that binary variables are coded as 0 and 1. Also, categorical variables with three or more categories need to be recoded as dummy variables with 0/ 1 outcomes e.g. class needs to appear as two variables 1st/ not 1st with 1 = yes and 2nd/ not 2nd with 1 = yes. Luckily R does this for you, but the variables need to be declared as factors. To do this use `var1.f <- factor(var1)`

Interpretation of the output

Fit a logistic model to the response variable survived (y) and using the explanatory variables (x_i) p.class, Alone and Gender. The indicator function - `I()`- is used to define the reference category .

```
> model1=glm (survived ~ I(pclass.f==1)+ I(pclass.f==2) + I(Gender.f==1) + I(Alone.f==1), data =
titanic_data,  family = binomial)
summary(model1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1317  -0.6825  -0.4649   0.6968   2.1347
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.1703     0.1332 -16.296  < 2e-16 ***
I(pclass.f == 1)TRUE   1.7034     0.1724   9.882  < 2e-16 ***
I(pclass.f == 2)TRUE   0.8319     0.1779   4.676 2.93e-06 ***
I(Gender.f == 1)TRUE   2.4743     0.1510  16.384  < 2e-16 ***
I(Alone.f == 1)TRUE    0.1560     0.1462   1.067    0.286
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1741.0  on 1308  degrees of freedom
Residual deviance: 1256.1  on 1304  degrees of freedom                    AIC: 1266.1
```

The output gives the coefficients for the parameters, the null deviance and the residual deviance.

The full model can be written as

$$\ln\left(\frac{p}{1-p}\right) = -2.170 + 1.703x_{1st\ class} + 0.832x_{2nd\ class} + 2.474x_{female} + 0.156x_{alone}$$

$x_{1st\ class} = 1$ for 1st class, $x_{2nd\ class} = 1$ for 2nd class, $x_{female} = 1$ for women and $x_{alone} = 1$ for a person not travelling alone

To check whether the variables affect the response the significance of the coefficients (`Pr(>|z|)`)can be used. The p-values are all below 0.05 apart from the test for the variable Alone (p = 0.286). This means that although the Chi-squared test for Survival vs. Alone was significant, once the other variables were controlled for, there is not a strong enough relationship between Alone and survival. Class is tested as a whole (pclass) and then 1st and 2nd class compared to the reference category 3rd class. When interpreting the differences, it is easier to look at look at the $\exp(\beta_i)$ which represents the odds ratio for the individual variable:

```
> exp(model1$coeff)
    (Intercept)   I(pclass.f == 1)TRUE   I(pclass.f == 2)TRUE  I(Gender.f == 1)TRUE  I(Alone.f == 1)TRUE
      0.1141392      5.4927193              2.2976664             11.8728634            1.1688321
```

For example, those in 1st class were 5.49 times more likely to survive than those in first class. With gender, the odds ratio compares the likelihood of a male surviving in comparison to females. The odds for women are a lot higher than for men (11.87 times that of women). Alternatively, the odds of a male surviving over a female using 1/11.87 = 0.084. Females were 11.9 times more likely to survive. Similarly, those travelling with company were 1.2 times more likely to survive.

Although this model does not include negative coefficients, a negative coefficient means that the odds of survival decreases.

The log odds can be obtained:

```
> predict(model1)
```

And the odd ratios

```
> fitted.values(model1)
> predict(model1,type="response")
```

To check how the fitted values of our model match the response variable survived:

```
> table(survived,predict(model1)>0)
 survived FALSE TRUE
       0   682  127
       1   161  339
> (682+339)/1309
[1] 0.7799847
```

This means that 78% of the fitted values are correctly classified.

We compare this with the null model which has no explanatory variables (only includes a constant) such that each person has the same chance of survival.

```
model0=glm (survived ~ NULL, data = titanic_data,  family = binomial)
> model0

Coefficients:
(Intercept)     -0.4812

Degrees of Freedom: 1308 Total (i.e. Null);  1308 Residual
Null Deviance:      1741
Residual Deviance: 1741         AIC: 1743
```

The null model is written as $\ln\left(\frac{p}{1-p}\right)=\beta_0=-0.481$ p=probability of survival=$\frac{\exp(-0.481)}{1+\exp(-0.481)}=0.382$

To check how well the null model fits the data:

```
> table(survived,predict(model0)>0)
 survived FALSE
       0   809
       1   500
> 809/1309
 [1] 0.618029
```

62% of the fitted values are correctly classified, an improvement of 16.2% on the classification.

How good is the model?

In standard regression, the coefficient of determination (R^2) gives an indication of how much variation in *y* is explained by the model. This cannot be calculated for logistic regression, to check the suitability of the model, the deviance, also called G^2 (which corresponds to -2*log-likelihood) is used. A small value of G^2 is preferable. Also a comparison between the deviance for the null model and the full model can be used; this is sometimes called a G^2 test for goodness of fit. This involves comparing to difference of the residual deviances with the difference in number of degrees of freedom using a chi-squared distribution.

```
> pchisq(1741.0-1256.1  ,1308-1304)
[1] 1
```

A large p-value means that the full model is an improvement on the null model.

The deviance can also be used to compare models, for example if the following model is used:

```
> model2=glm (survived ~ I(pclass.f==1)+ I(pclass.f==2) +I(Gender.f==1), data = titanic_data,  family =
binomial)
```

To test which model is better:

```
> anova(model2,model1,test="Chi")
Analysis of Deviance Table

Model 1: survived ~ I(pclass.f == 1) + I(pclass.f == 2) + I(Gender.f ==  1)
Model 2: survived ~ I(pclass.f == 1) + I(pclass.f == 2) + I(Gender.f ==  1) + I(Alone.f == 1)
  Resid. Df  Resid. Dev   Df  Deviance P(>|Chi|)
1      1305     1257.2
2      1304     1256.1    1    1.1322    0.2873
```

This means that the variable `Alone` is not needed in the model as `P(>|Chi|)` is greater than 0.05.