stcp-karadimitriou-logitR

The following resources are associated:

Scatterplots and correlation, Checking normality in R and the csv dataset Birthweight reduced.csv'

Logistic regression in R

Dependent variable: Categorical (two groups)

Independent variables: Continuous / Categorical

Common Applications: The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (independent) variables. It can be used to predict the survival probability of an individual based on the independent variables, or it can classify patients into having a certain type of disease.

Data: On April 14th 1912 the Titanic sank. Only 705 passengers and crew out of the total 2228 population on board survived. Information 1309 of those on board will be used to demonstrate logistic regression. Open the data are 'Titanic.csv' and use as dependent variable the survival outcome (Survival, 1=survived, 0=not) and we will use the possible explanatory variables, e.g. Age (scaled), sex, class (pclass=1,2 or 3), whether each passenger was alone (Alone=Yes) or not (Alone=No) and Country of Residence(Residence=American,British,Other).

```
#Reading the data and define them as a data frame
TitanicR<-data.frame(read.csv('...\\Titanic.csv',header=T,sep=','))
```

	pclass	survived	Residence	name	sex
1	3	0	0	Abbing, Mr. Anthony	male
2	3	0	0	Abbott, Master. Eugene Joseph	male
3	3	0	0	Abbott, Mr. Rossmore Edward	male
4	3	1	0	Abbott, Mrs. Stanton (Rosa Hunt)	female
5	3	1	2	Abelseth, Miss. Karen Marie	female
6	3	1	0	Abelseth, Mr. Olaus Jorgensen	male
7	2	0	2	Abelson, Mr. Samuel	male
8	2	1	2	Abelson, Mrs. Samuel (Hannah Wizosky)	female
9	3	1	2	Abrahamsson, Mr. Abraham August Johannes	male
10	3	1	2	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female

```
#Attaching the data to use immediately the variables by their column name
attach(TitanicR)
```

Most of the variables can be investigated using crosstabulations with the dependent variable 'survived'. Another reason for the cross tabulation is to identify categories with small frequencies as this can cause problems with the logistic regression procedure. For building contingency tables and testing relationships between categorical variables via Chi Squared tests, check the 'Chi-Squared test Using R' help sheet.

```
#Specifying the categorical variables
pclassf<-factor(data$pclass)
survivedf<-factor(data$survived,levels=c(0,1),labels=c('NotSurvived','Survived'))
Residencef<-factor(data$Residence,levels=c(0,1,2),labels=c('American','British','Other'))
Alonef<-factor(data$Alone,levels=c(0,1),labels=c('Alone','NotAlone'))
```

Conducting Logistic Regression

Logistic regression will be initially carried out using as dependent the binary survival outcome and as independent the sex, Residence, pclass and Alone. All of the explanatory variables are categorical and so there will be a different interpretation of the coefficients when there is a continuous one. The indicator function `I()` is used to define the reference category.

```
#Fitting the logistic regression
fit<-
glm(I(survived=='Survived')~I(pclass==1)+I(pclass==2)+Residence+Alone+sex,family
=binomial)
```

In the output we see the quantiles of the **deviance residuals**, which are a measure of model fit and we will discuss further on this matter afterwards.

```
> summary(fit)

Call:
glm(formula = I(survivedf == "Survived") ~ I(pclassf == "First") +
    I(pclassf == "Second") + Residencef + I(Genderf == "Male"),
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1478  -0.6450  -0.4845   0.6779   2.2262

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.5646     0.2283   2.473   0.0134 *
I(pclassf == "First")TRUE  1.6370     0.1994   8.210  < 2e-16 ***
I(pclassf == "Second")TRUE 0.9259     0.1910   4.847 1.26e-06 ***
ResidencefBritish         -0.4442     0.2373  -1.871   0.0613 .
ResidencefOther           -0.1369     0.2079  -0.658   0.5103
I(Genderf == "Male")TRUE  -2.5107     0.1470 -17.080  < 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: 1741.0  on 1308  degrees of freedom
Residual deviance: 1253.1  on 1303  degrees of freedom
AIC: 1265.1

Number of Fisher Scoring iterations: 4
```

The next part of the output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values.

Both pclass (reference group is the third class) and sex with reference group the females are statistically significant (p-values<0.0001). However, Country of Residence (with reference group the Americans) and Alone are not statistically significant p-values>0.05. This means that *there is no difference in the log odds of survival between the Residence group.*

However, it contradicts the Chi-squared test in Chi-squared in R sheet where there was a significant association between survival and Residency. This happens because whilst controlling for other variables, the significance of the independent can be changed.

The interpretation of the coefficients is, for example, being of a class 1 versus a class 3, changes the log odds of survival by **1.63** and being a class 2 passenger versus a 3, changes the log odds of survival by **0.92**. Hence, for a male passenger, the log of odds of survival changes by **-2.51** versus a female passenger.

Below the table of coefficients are fit indices, including the null and deviance residuals and the AIC.

We can use the `confint()` function to obtain confidence intervals for the coefficients but here we will exponentiate the coefficients and interpret them as odds-ratios. To get the exponentiated

coefficients, we use the `exp()` command, to derive just the coefficients of the model you can do it through `coef()` command. We can use the same logic to get odds ratios and their confidence intervals, by exponentiating the confidence intervals.

```
#getting the exponential values of the coefficients
exp(coef(fit))
#building confidence intervals for the coefficients
confint.default(fit)
#Extracting the Odds Ratio and the 95% CI
exp(cbind(OR = coef(fit), confint(fit)))
```

```
> exp(cbind(OR = coef(fit), confint(fit)))
Waiting for profiling to be done...
                                    OR      2.5 %     97.5 %
(Intercept)                 1.75873932 1.12615110 2.7581550
I(pclassf == "First")TRUE   5.13961470 3.48896283 7.6287561
I(pclassf == "Second")TRUE  2.52401432 1.73771757 3.6768688
ResidencefBritish           0.64134883 0.40228692 1.0208227
ResidencefOther             0.87207812 0.58065153 1.3126507
I(Genderf == "Male")TRUE    0.08121089 0.06060765 0.1078807
```

The output for Odds Ratio can be interpreted as, by holding Residence, Alone and sex the odds of survival for class 1 (pclass = 1) over the odds of survival for 3rd class (pclass = 3) is **exp(1.631) = 5.10**. In terms of percent change, we can say that the odds for 1st class are 410% higher than the odds for 3rd class. Equivalently, the odds of survival for 2nd class versus the odds of survival for the 3rd class is **2.52** (151%). Finally, holding the rest of independent variables at fixed category, the odds of survival for males over the odds of survival for females is 0.08, i.e., the females are **1-0.08=0.92** (92%) more likely to survive.

```
> table(survivedf,predict(fit)>0)

survivedf  FALSE TRUE
  Died       682  127
  Survived   161  339
> (682+339)/1309
[1] 0.7799847
```

The classification table shows us that 78% of the fitted values were correctly classified. In terms of accuracy that means that our model predicts way better than just a random guess, i.e., the chance of predicting correctly for survival or not by random is 50% so our model is more accurate than random guessing.

Model Selection

We will try and fit a model without the insignificant variables and then we will compare both of them through a deviance hypothesis testing between models. This expression is simply **−2 times the log-likelihood ratio of the reduced model compared to the full model**. The deviance is used to compare two generalised linear models and it has a similar role to residual variance from ANOVA in linear models. The test involves the comparison of differences of the residuals deviances with the difference in number of degrees of freedom using a chi-squared distribution.

```
#Fitting the reduced model
fit2<-
glm(I(survivedf=='Survived')~I(pclassf=='First')+I(pclassf=='Second')+I(Genderf=
='Male'),family=binomial)
#Comparing the models
anova(fit2,fit,test='Chi')
```

The first argument of the output gives us the two models to be assessed. Model 1 indicates the reduced and Model 2 indicates the full one.

```
> anova(fit2,fit,test='Chi')
Analysis of Deviance Table

Model 1: I(survivedf == "Survived") ~ I(pclassf == "First") + I(pclassf ==
    "Second") + I(Genderf == "Male")
Model 2: I(survivedf == "Survived") ~ I(pclassf == "First") + I(pclassf ==
    "Second") + Residencef + I(Genderf == "Male")
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1305     1257.2
2      1303     1253.2  2   4.0771   0.1302
```

The table shows the residual degrees of freedom ($df_{reduced}$=1305,df_{full}=1302), the residual deviance of each model, the difference in degrees of freedom in which case is 3, the difference of the deviances (5.2747) and the p-value=0.1528. This means that under the hypothesis that the reduced model is better, with a test statistic of X^2_3=5.2747 we do not reject that the reduced model fits better.

We may also wish to see measures of how well our model of choice fits. This can be particularly useful when comparing competing models. One measure of model fit is the significance of the overall model. This test asks whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). The test statistic is the difference between the residual deviance for the model with predictors and the null model.

```
#Fitting the intercept model
fitnull<-glm(I(survivedf=='Survived')~1,family=binomial)
#Comparing them
anova(fitnull,fit2,test='Chi')
```

Since the output gave us a p-value<0.0001 we can state that our model fits better than assuming that each person has the same chance of survival (null).

The maths for the final model

The regression process finds the coefficients which minimise the squared differences between the observed and the expected values of the dependent variable. As the outcome of logistic regression is binary, the dependent needs to be transformed. The logit transformation gives the following:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

with p being the probability of event occurring and p/1-p the odds ratio. If the probabilities of the event of interest happening for individuals are needed, the logistic regression can be written for 0<p<1 as:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_q x_q}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_q x_q}}$$

So for our model we have

```
> coef(fit2)
               (Intercept)  I(pclassf == "First")TRUE I(pclassf == "Second")TRUE
                 0.3860037                  1.7231291                  0.8423059
I(Genderf == "Male")TRUE
                -2.5150035
```

$$log\frac{P(Survived)}{P(Not\ Survived)} = 0.386 + 1.723(pclassf = 'First') + 0.842(pclassf = 'Second') - 2.61(Genderf = 'Male')$$

and so the probability of survival for a 2nd class female passenger can be calculated as:

$$p = \frac{e^{0.386+0.842}}{1+e^{0.386+0.842}} = 0.704$$

Notes: In logistic regression there are not as restricted assumptions as in linear regression. The only assumption is that the dependent variable is identically, ndependently and binomially distributed and the residuals of regression should be independent. To check them, a scatter plot of predicted vs residuals can be used to assess whether they have a cloud form (independence) or a specific pattern (no independence).