Statistical Methods for Big Data

**Further Logistic Regression Examples**

**Logistic regression with a binary response variable and three categorical explanatory variables**

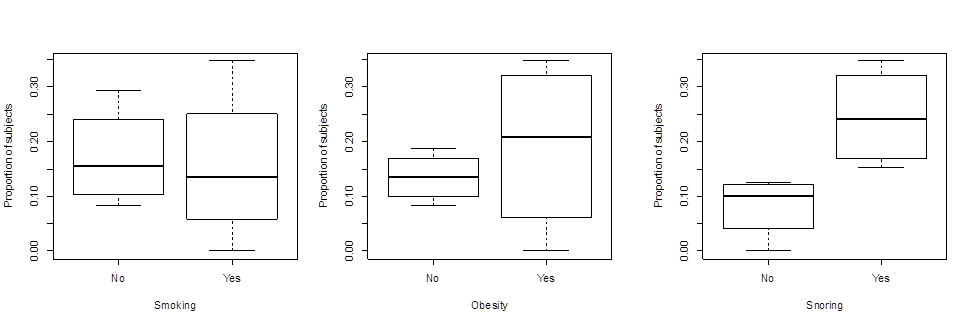
A study to examine the relationship between hypertension (high blood pressure) and smoking, obesity and snoring, produced the following results:

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Hyp** | **Total** | **Prop** | **Smoking** | **Obesity** | **Snoring** |
| 5 | 60 | 0.083 | Yes | Yes | Yes |
| 2 | 17 | 0.118 | No | Yes | Yes |
| 1 | 8 | 0.125 | Yes | No | Yes |
| 0 | 2 | 0 | No | No | Yes |
| 35 | 187 | 0.187 | Yes | Yes | No |
| 13 | 85 | 0.153 | No | Yes | No |
| 15 | 51 | 0.294 | Yes | No | No |
| 8 | 23 | 0.348 | No | No | No |

We refer to each unique combination of explanatory variables as a treatment. To examine the relationship between the proportion of individuals suffering from hypertension and each of the explanatory variables we can use boxplots (see Fig 1).

**Figure 1**



We can examine the distribution of the proportion of individuals suffering from hypertension using a histogram (see figure 2).

**Figure 2**



R is able to perform logistic regression for proportional data in two different ways.

1. Specify the response variable as a response matrix of two columns, where one column is the number of "successes" and one column is the number of "failures".
2. Specify the response variable as a single column containing the proportion of successes - however, a weighting argument containing the total number of trials must also be included in the model specification.

Let’s look at the code for the first case. Suppose the data frame containing all the information shown in Table 1 is called T, we will assume that the variable names correspond to those given in Table 1.

First, we specify the response matrix, (we call it y)

y<-cbind(T$Hyp, T$Total-T$Hyp)

next we specify the model:

model1<- glm(y~Smoking+Obesity+Snoring, binomial, data = T)

Now let’s look at the code for the second case. Again, suppose the data frame containing all the information shown in Table 1 is called T, we will assume that the variable names correspond to those given in table 1.

We specify the model:

model2<- glm(Prop~Smoking+Obesity+Snoring, binomial, data = T, weights = Total)

Note that model 1 and model 2 are identical.

Summarising the model, we get the following output:

Call:

glm(formula = y ~ Smoking + Obesity + Snoring, family = binomial, data = T, weights = Total)

Deviance Residuals:

1 2 3 4 5 6 7 8

-0.0434 0.5415 -0.2548 -0.8005 0.1976 -0.4660 -0.2126 0.5623

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.37766 0.38018 -6.254 4e-10 \*\*\*

smokingYes -0.06777 0.27812 -0.244 0.8075

obesityYes 0.69531 0.28509 2.439 0.0147 \*

snoringYes 0.87194 0.39757 2.193 0.0283 \*

---

Signif. codes: 0 `\*\*\*' 0.001 `\*\*' 0.01 `\*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 14.1259 on 7 degrees of freedom

Residual deviance: 1.6184 on 4 degrees of freedom

AIC: 34.537

Number of Fisher Scoring iterations: 4

**Deviance Residuals**

The deviance residuals show the deviance of each treatment from the fitted model. The sign indicates whether the observation is larger or smaller than expected (we usually examine the residuals in a plot).

**Table of Coefficients**

The table of coefficients gives the estimates of the regression coefficients, their standard errors, corresponding z-values and associated p-values.

The fitted model is of the form:

Where *pi* represents the proportion of subjects with hypertension, represents the categorical variable Smoking, represents the categorical variable Obesity and represents the categorical variable Snoring.

The intercept has a value of -2.37766, the p-value indicates that it is significantly different to 0. **The intercept represents a baseline measure of the log-odds when all the categorical variables are set to “No”.**

In other words, if a subject does not smoke, is not obese and does not snore, then the *log-odds* of the probability that the subject suffers from hypertension is -2.37766.

We can also state that if a subject does not smoke, is not obese and does not snore, then the *odds* of the probability that the subject suffers from hypertension is

The odds in favour of that type of subject (non-smoker, not obese, does not snore) suffering from hypertension are very small.

Note that since the explanatory variables are categorical, the coefficients do not represent slopes, they represent effects on the intercept.

The value of is -0.06777 and it represents the effect that smoking (smokingYes) has on the log-odds. The p-value indicates that it is NOT significantly different to 0 so we question whether to include this variable in our final model.

If a subject **does** smoke, is not obese and does not snore, then the *log-odds* of the probability that the subject suffers from hypertension is

-2.37766-0.06777 = -2.44536.

We can also state that if a subject **does** smoke, is not obese and does not snore, then the *odds* of the probability that the subject suffers from hypertension is

In the current model, smoking is predicted to reduce (very slightly) the probability that a subject will suffer from hypertension.

**Exercise:** Calculate the odds in favour of a subject suffering from hypertension for the following subjects:

1. Does not smoke, is not obese, does snore
2. Does smokes, is obese, does snore

**Null and Residual Deviance**

Recall, the deviance (D) can sometimes be used as a measure of goodness of fit.

The deviance of a model, D, is calculated by comparing the given model with the saturated model - a model with a theoretically perfect fit. Smaller values of D indicate a better fit as the fitted model deviates less from the saturated model.

The residual deviance gives D for the specified model. In this example, the residual deviance is 1.6184 with 4 degrees of freedom (there are four degrees of freedom since there are 8 treatments and 4 parameters estimated from the data).

At the 5% level, the critical value in the chi-square distribution table with 4 degrees of freedom is 9.49. Since the calculated residual deviance is less that this value, we conclude that the model is an adequate fit.

The null deviance gives D for the null model (intercept only). In this example the null deviance is 14.125 with 7 degrees of freedom.

Comparison of the model deviance to the null deviance can be used for model selection. This comparison asks whether the model including the explanatory variables fits significantly better than a model with just an intercept. The test statistic is the difference between the residual deviance and the null deviance. The test statistic has a chi-square distribution with degrees of freedom equal to the differences in degrees of freedom between the fitted model and the null model (i.e., the number of explanatory variables in the model).

In this example, the test statistic is 14.125-1.6184 = 12.5066 with 3 degrees of freedom.

At the 5% level, the critical value in the chi-square distribution table with 3 degrees of freedom is 7.78. Since the test statistic (12.5066)is greater than 7.78 we conclude that there is a significant difference between the fitted model and the null model.

**Model Selection**

The p-value associated with smoking indicated that the effect of smoking was not significantly different to 0. We would like to simplify the model by dropping the smoking term. We try two models:

modelA<- glm(y~Smoking+Obesity+Snoring, binomial, data = T)

modelB<- glm(y~ Obesity+Snoring, binomial, data = T)

Note that the output for modelA was given above.

For modelB we obtained the following output:

Call:

glm(formula = y ~ obesity + snoring, family = binomial, data = hyp)

Deviance Residuals:

1 2 3 4 5 6 7 8

-0.01247 0.47756 -0.24050 -0.82050 0.30794 -0.62742 -0.14449 0.45770

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.3921 0.3757 -6.366 1.94e-10 \*\*\*

obesityYes 0.6954 0.2851 2.440 0.0147 \*

snoringYes 0.8655 0.3967 2.182 0.0291 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 14.1259 on 7 degrees of freedom

Residual deviance: 1.6781 on 5 degrees of freedom

AIC: 32.597

By testing the difference in the residual deviance of modelA and modelB we can decide whether we should retain the Smoking variable.

The test statistic is the difference in the residual deviances of the two models,

1.6781 - 1.6184 = 0.0597

The degrees of freedom is the difference between the degrees of freedom for the two models (5 – 4 = 1)

Using R we can calculate the p-value:

pchisq(0.597,1, lower=F)

Using tables we can test the significance using the critical value method.

At the 5% level, the critical value in the chi-square distribution table with 1 degree of freedom is 3.841. Since the test statistic (0.0597)is less than 3.841 we conclude that there is not a significant difference between the two models and therefore we drop the smoking variable and select simpler model.

**Exercise:**

Test the significance of the snoring variable.