Logistic regression in R

author: Maria Paniw date: 10.02.2016 width: 1620 height: 1080

GLM

The binomial distribution, along with Poisson and negative binomial distributions, which we will cover next, belong to the family of exponential distributions in which the parameters can be transformed into linear predictors using **link functions**. We have covered the *logit* link for the binomial distribution in the last lecture.

These exponential distribution can therefore be described as **generalized linear models (glm)**:

$$g(y) = \beta_0 + \beta_1 X$$

,

where $g()$ is the link function.

In R, there is a function called `glm()` that defines the likelihood function for you and estimates the parameters of your model.

Example: Droso flowering probability as a function of size and time since fire

```
#As always load the data:

data=read.table("G:/Teaching/Rstats/presentations/flDroso.txt",
                header=T)


#Now fit the logistic model:

mod=glm(fl~size+TSF,data=data,family="binomial")
```

What is the output?

```
summary(mod)
```

```
Call:
glm(formula = fl ~ size + TSF, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9549  -0.4956  -0.2565   0.6466   2.7881
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.6140     0.8245  -8.022 1.04e-15 ***
size          1.1106     0.1366   8.130 4.29e-16 ***
TSFthree      0.9422     0.2762   3.411 0.000647 ***
TSFtwo       -2.4206     0.3280  -7.379 1.59e-13 ***

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1223.1  on 1051  degrees of freedom
Residual deviance:  831.6  on 1048  degrees of freedom
  (1444 observations deleted due to missingness)
AIC: 839.6

Number of Fisher Scoring iterations: 6
```

Interpreting the estimate

Our model is defined as:

$$log(\frac{Pr(fl|size, TSF, \beta_0, \beta_1, \beta_2, \beta_3)}{1 - Pr(fl|size, TSF, \beta_0, \beta_1, \beta_2, \beta_3)}) = -6.61 + 1.11size + 0.94TSFthree - 2.42TSFtwo$$

Therefore,

$\beta_0 = -6.61$ - log odds of flowering in TSF>three with size 0

$\beta_1 = 1.11$ - log odds ratio of flowering with each unit increase in size

$\beta_2 = 0.94$ - log odds ratio of surviving in TSF three compared to >three

$\beta_3 = -2.42$ - log odds ratio of surviving in TSF two compared to >three

Interpreting the deviance

The deviance gives us a goodness of fit test for our model by comparing observed and predicted values using loglikelihoods!

The Null deviance is $-2 * logL(log(\frac{Pr(fl|\mu)}{1-Pr(fl|\mu)}))$

The residual deviance is $-2 * logL(mod)$

The relative deviance explains how much variation of the data is explained by adding the predictors *size* and *TSF*.

$$D^2 = \frac{null.deviance - residual.devinace}{null.deviance} = \frac{1223.1 - 831.6}{1223.1} = 0.32$$

Interpreting AIC

The AIC, or **Akaike Information Criterion** is calculated as

$$AIC = -2logL(model) + 2K$$

K = number of estimated parameters in model.

For small sample sizes n, (i.e., n/k < 40), the AIC is corrected:

$$AIC_c = -2logL(model) + 2K + \frac{2K(K+1)}{n-K-1}$$

AIC is a test of relative model significance and is closely tied to hypothesis testing with glms.

That is:

$H_0 : logit(Y_i) = \beta_0$

$H_a : logit(Y_i) = \beta_0 + \beta_1 size + \beta_2 TSF3 + \beta_3 TSFtwo$

We can get the AIC for the null model and compare it to the AIC of the **saturated model** (mod):

```
mod0=glm(fl~1,data=data,family="binomial")

AIC(mod0,mod)
```

```
     df       AIC
mod0  1 1225.0984
mod   4  839.6039
```

```
# a lower AIC means higher likelihood of the data.
#A rule of thumb is that a difference of 5 in AIC is significant.
```

The likelihood ratio test

Besides the AIC, there is another option of testing alternative models, the lilihood ratio test, defined as:

$$G = -2ln(\frac{L(null.model)}{L(full.model)}) = D(null.model) - D(full.model)$$

The statistic G is asumed to be distributed as $G \sim \chi^{\in}(df)$

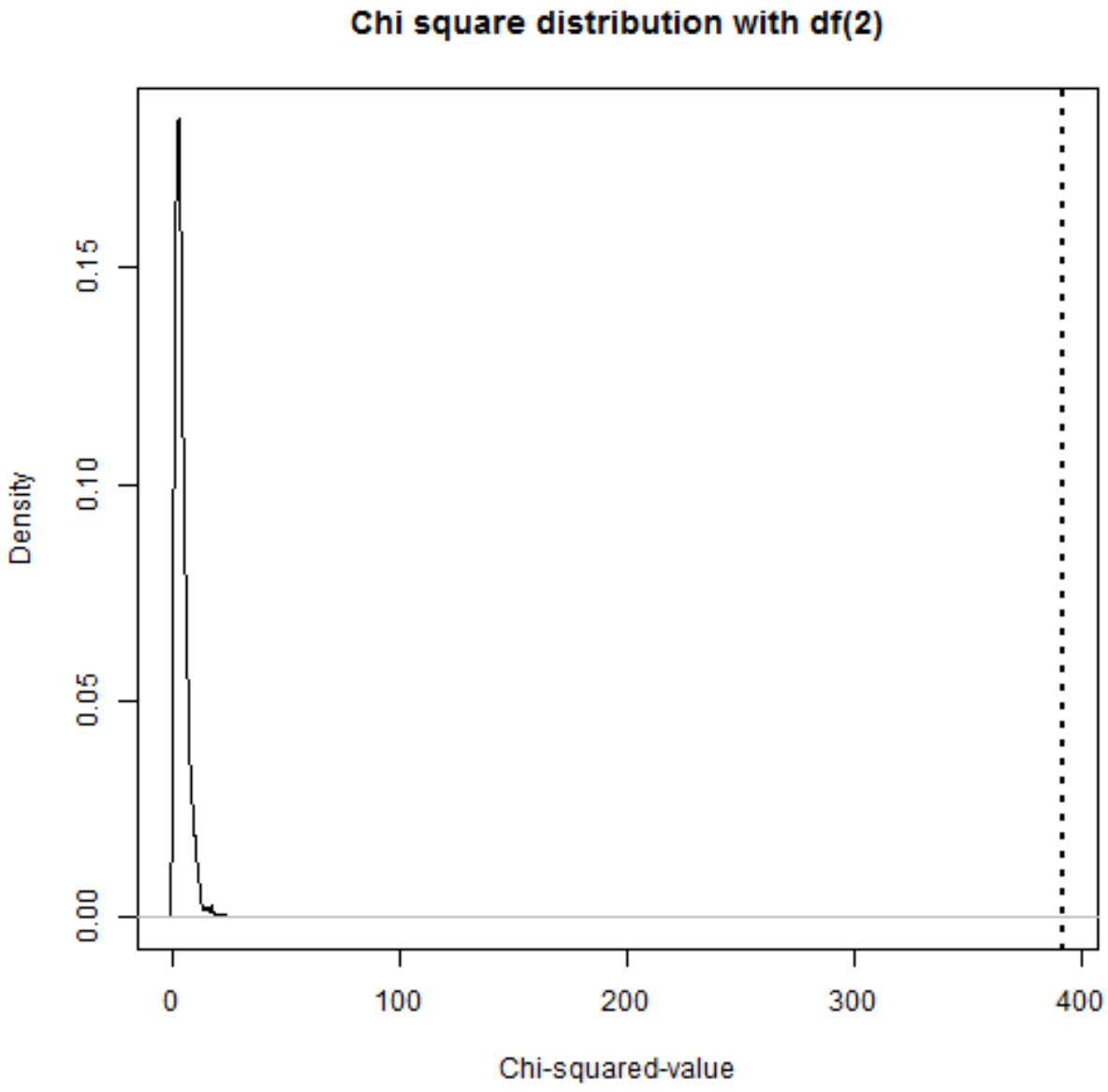
So, let's check our model:

$$G = 1223.1 - 831.6 = 391.5$$

Make the distribution plot

```r
Chidens=density(rchisq(n = 1000, df = 4))

#Plot
plot(Chidens,
     xlim = c(0,391.5), main = "", xlab = "Chi-squared-value")
title("Chi square distribution with df(2)")
abline(v = 391.5, lwd = 2, lty = 3)
```



ANOVA

You can do the likelihood ratio test using ANOVA in R

```r
anova(mod0,mod,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: fl ~ 1
Model 2: fl ~ size + TSF
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1051     1223.1
2      1048      831.6  3   391.49 < 2.2e-16 ***

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

A last remark on logistic regression

The response variable Y does not have to be input as 0-1. Instead we can do it the following way:

```r
# We will work with an example in the package faraway
#(download it if you don't have it).
# Here, we want to know how pesticide use affects the beetles on corn plants.

library(faraway)

data(beetle)
#The data contain the total number of beetles exposed to pesticides (exposed)
# of different concentration (conc)
# and how many of the exposed got affected (affected)

str(beetle)
```

```
'data.frame':   10 obs. of  3 variables:
 $ conc    : num  24.8 24.6 23 21 20.6 18.2 16.8 15.8 14.7 10.8
 $ affected: num  23 30 29 22 23 7 12 17 10 0
 $ exposed : num  30 30 31 30 26 27 31 30 31 24
```

This is fundamentally a binomial response because we have only two outcomes, exposed (p) or not (1-p). But there are no 0s or 1s. Now what?

Well, in R you can input the data as:

```r
affected=beetle$affected #positive response 1

not.affected=beetle$exposed-beetle$affected # 0

response=cbind(affected,not.affected) # bind the two into a matrix

# Fit the model
glm.beetle <- glm(response ~ conc, family = binomial,data=beetle)
```