Session 5: Logistic regression: to binary and beyond…

For both ANOVA and regression we have so far been considering continuous dependent variables or responses. However not all responses are continuous. A common type of response data are binary responses such as yes/no or present/absent data. Examples could be

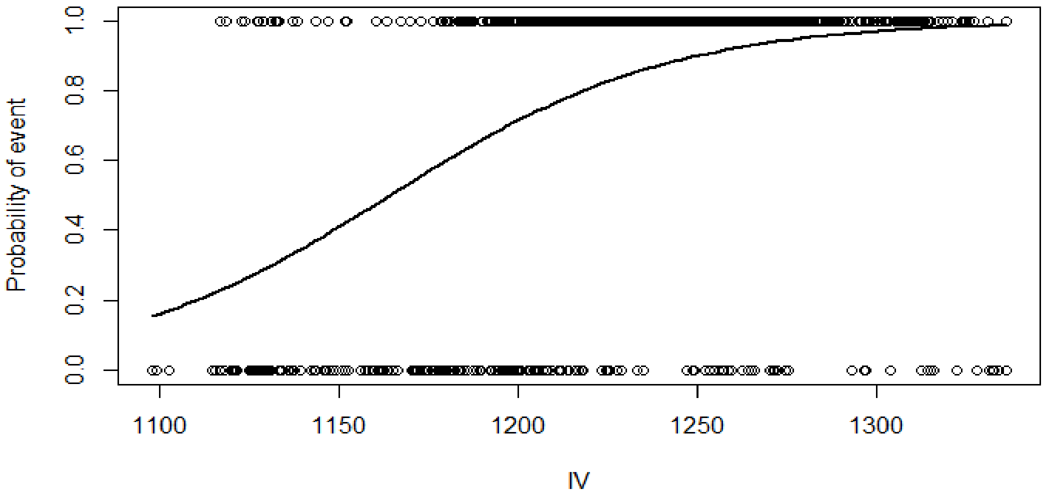
- Gene expression
- Survival of a pathogen
- Presence or absence of specific cell product

These can all be represented as values of 1 (yes, presence etc.) or 0 (no, absence etc.). We can't use linear regression to explore the relationships between these kind of data and different predictors because they are not continuous. Instead we can use a technique known as logistic regression.

A generalisation

Logistic regression is broadly similar to linear regression other than the fact that it deals with binary response data. Standard linear models are based around the expectation of a normal distribution. Logistic regression does not make this assumption. Instead the technique is based around a **binomial** distribution. This is one of a series of alternative distributions that allow linear models to be extended to a wider range of situations. These are called **Generalised** Linear Models and involve using the `glm` command rather than `lm`.

In the case of logistic regression, rather than fitting a straight line to the data, a **logistic** curve, which is a S-shaped or sigmoidal, is used. The fitted line can be interpreted as a prediction of the probability that a response value of 1 will occur for a given value of your IV.



Risky business

Download the chd_risk.csv file from Moodle, start a new section of your script file and write code to import into an object called chd_risk.

These data come from a study of some of the risk factors associated with the occurrence of coronary heart disease in men. Occurrence of a chd event was recorded as a binary variable (1= yes, 0 = no). Age (range from 16-76 years) and serum levels of uric acid and lioprotein A (lpa) were also recorded from all individuals.

Check out the structure of the data (str).

The response column is an integer - it needs to be 1 and 0 values, rather than Yes and No for example - no need to convert to a factor. The rest of the variables (your independent variables) are also continuous.

We also need to install another package - called lmtest. This is used for assessing the overall significance of the logistic regression model. If you are using the AppsAnywhere version of RStudio this will already be installed. if you are using your own version, then use the Packages pane to install this.

Once this is installed, load it up by typing

```r
library(lmtest)
```

in your script. We also need to use car again so add

```r
library(car)
```

and run both lines.

The setup of the code for the analysis follows the same pattern as for regression and ANOVA, with two differences. First we are using glm rather than lm and second we need to add another argument to tell R what distribution to use. This is done through the family = "binomial" argument.

We will start by looking at a single predictor, in this case age.

```r
c1 <- glm(chd~age, data=chd_risk, family="binomial")
```

But, but...

At this point your assumption checking radar should be screaming out! You can relax. Logistic regression makes none of the assumptions of normality of residuals or equality of variance that the previous techniques did.

The main consideration is the breakdown of your response in terms of the frequency of 1 and 0 values. We can check this using

```r
table(chd_risk$chd)
```

```
##
##    0    1
## 1124  262
```

The good news is that most of your sample did not have a chd event (value 0). The numbers of not exactly balanced however, with around 4 times as many No events than Yes. There are no hard and fast rules with regards to how balanced they need to be but if you have less than 10% events in one category you need to be careful of your interpretation.

So now we can look at the results using

```
summary(c1)

##
## Call:
## glm(formula = chd ~ age, family = "binomial", data = chd_risk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7321  -0.6297  -0.4510  -0.2245   2.9655
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.023411   0.362946  -16.60   <2e-16 ***
## age          0.102410   0.007548   13.57   <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: 1343.9  on 1385  degrees of freedom
## Residual deviance: 1100.3  on 1384  degrees of freedom
## AIC: 1104.3
##
## Number of Fisher Scoring iterations: 5
```

Interpretation of the output is the same as for linear regression in terms of the significance and direction of the slopes (`Estimate` column in the summary). So in this case the probability of a chd event occurring increases with age (significant positive slope). The values however do not represent the change in your DV per unit change of your IV as they would for linear regression.

You will notice that there is no overall test of the significance of the fitted model as there is for linear regression. This is where the `lmtest` package comes in, specifically the `lrtest` function.

```
lrtest(c1)

## Likelihood ratio test
##
## Model 1: chd ~ age
## Model 2: chd ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -550.13
## 2   1 -671.96 -1 243.65  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The important parts are the Chisq value (this is the test statistic), the absolute value of the Df column (so in this case 1 - just ignore the minus) and the Pr(>Chisq), which is the p-value of the overall logistic regression. Overall we have a highly significant model for explaining the variation in occurrence of chd events which shows that the probability of an event occurring increases with age.

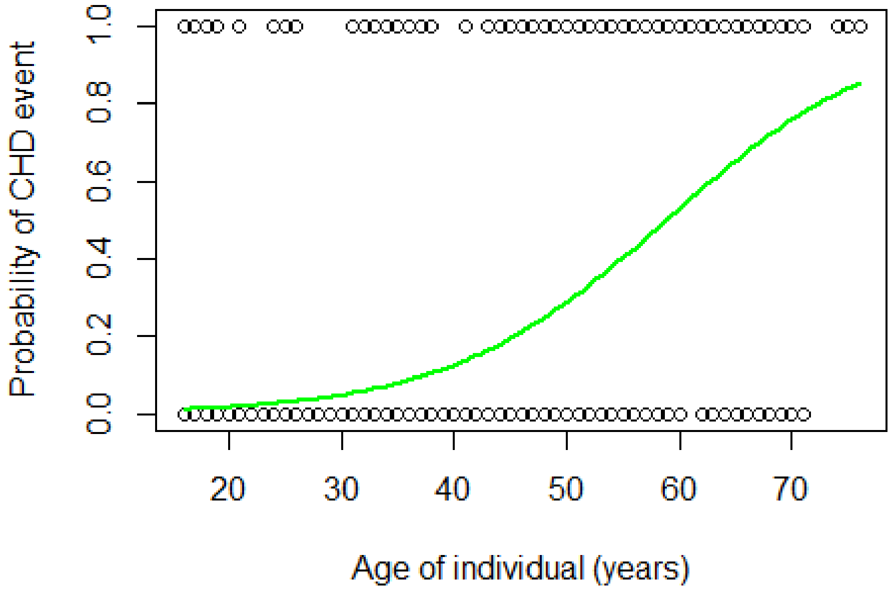
Plotting the results

Plotting the data is achieved through a scatter plot, but the distribution of values looks strongly differnt to a linear regression as there are only values at the 0 and 1 points on the y axis.

```
plot(chd~age, data=chd_risk, xlab="Age of individual (years)", ylab="Probabil
ity of CHD event")
```

In order to plot the fitted line (S-shaped curve) we can use the following:

```
curve(predict(c1, data.frame(age=x),type="response"), add=TRUE, col="green",
lwd=2)
# lwd makes the line thicker - you can use other values
```



More multiples

In exactly the same way as we did for linear regression, we can look at the influence of multiple predictors on our response. In this case we will evaluate the influence of serum uric acid and lipoprotein A concentrations.

```
c2 <- glm(chd~age+uric_acid+lpa, data=chd_risk, family="binomial")
```

Correlation of predictors can be a problem as for linear regression so you should check this using

```
vif(c2)

##       age uric_acid       lpa 
##  1.126910  1.126811  1.001117
```

No problems evident here (values of 5 or more would indicate a potential issue), so look at the outcome.

```
summary(c2)

## 
## Call:
## glm(formula = chd ~ age + uric_acid + lpa, family = "binomial", 
##     data = chd_risk)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7869  -0.5698  -0.3574  -0.1534   3.1896  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  1.322e+00  1.038e+00   1.273    0.203    
## age          7.268e-02  8.135e-03   8.934  < 2e-16 ***
## uric_acid   -1.714e-02  2.194e-03  -7.813 5.57e-15 ***
## lpa          2.164e-04  3.816e-05   5.671 1.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1343.92  on 1385  degrees of freedom
## Residual deviance:  987.36  on 1382  degrees of freedom
## AIC: 995.36
## 
## Number of Fisher Scoring iterations: 5

lrtest(c2)

## Likelihood ratio test
## 
```

```
## Model 1: chd ~ age + uric_acid + lpa
## Model 2: chd ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -493.68
## 2   1 -671.96 -3 356.56  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the results indicate that the probability of a chd event occurring increases with age and lipoprotein A concentration, and decreases with a higher serum uric acid concentration.

In this case all three predictors are significant, but just in the same way as with linear multiple regression, if there was one or more non-significant predictors then you would remove them and refit the model until only significant predictors remain.

For reporting you would need the chi-squared, df and p-values from the `lrtest` output, and would also report the detail of the estimates of slope and significance of the individual variables from the summary information.

You will notice that there is no R-squared value provided with the output either. For logistic regression this is a bit more difficult to define. There are several possibilities but none are entirely satisfactory measures so generally it is not included.

Logistic regression is a widely used technique, not just in biological sciences, but much more broadly, when looking to predict whether or not an event is likely. It forms the basis of a lot of the ‘machine learning’ techniques that are currently being applied in a whole range of areas, so now that you know how to do logistic regression in R, if biology does not work out, you can set yourself up as a big data machine learning consultant!

Make sure that you have written sufficient comments in your script to be able to go back to it in future sessions - you will need to be able to interpret your commands for other examples and for the tests in due course.