Logistic Regression Example in R.

August 21, 2014

1 Motivation.

- We are interested in modelling the probability that a student will score above 80% in their exam. This is therefore a dichotomous outcome variable i.e. (event occurs / event doesn't occur).
- We use a binary logistic regression to model the log odds of the outcome as a linear combination of the predictor variables.
- We therefore predict from a knowledge of relevant independent variables the probability (p) that it is 1 *(student scoring over or equal to 80%)* rather than 0 *(student scoring below 80%)*

2 Exploratory Data Analysis.

2.1 The Data.

The Data being analysed consist of exam results (%) for 91 students, and a binary variable **(high)** coded as a 1 if a student obtain 80% or above in their exam, and a 0 if not. For each student we also have their sex *(Male (n=51), Female (n=40))*, degree program *(Science, Joint, Arts)*, and their previous four homework results *(hw1, hw2,hw3,hw4)*. Shown below are the data for the first 6 students and the command in R to obtain them.

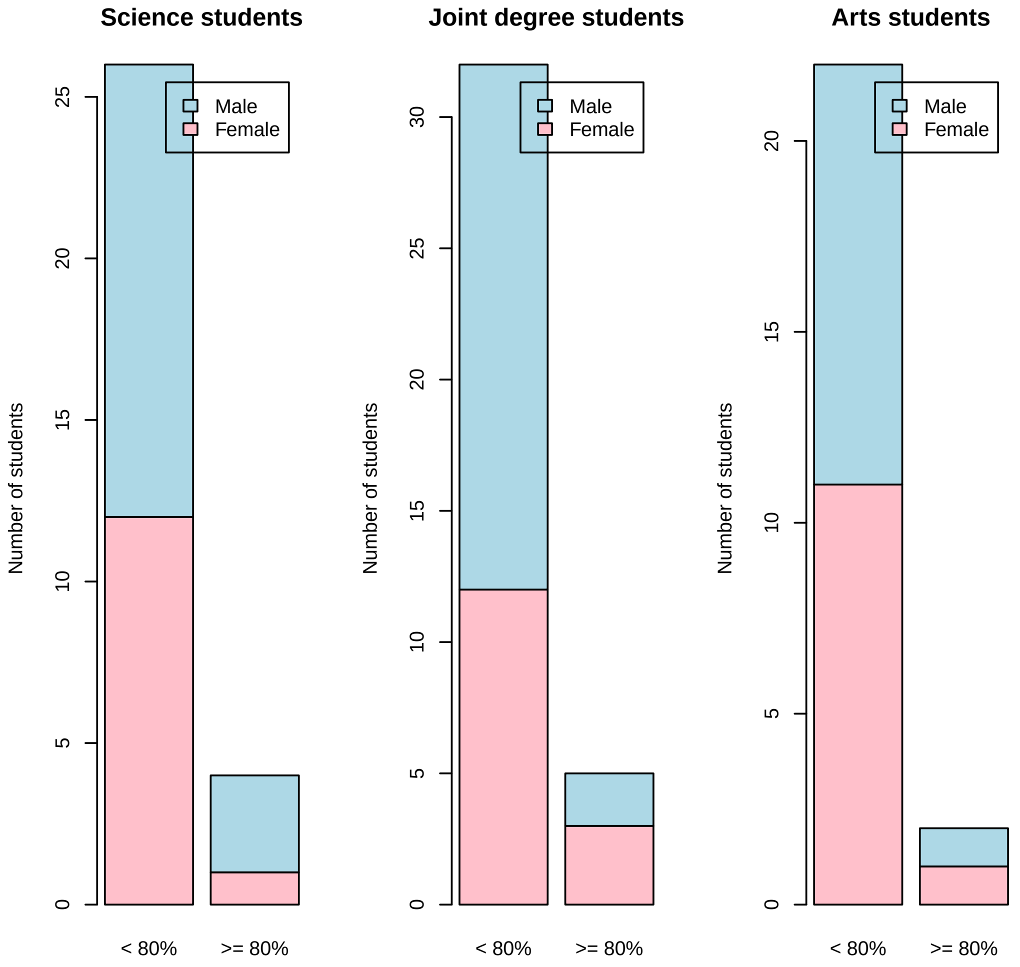
```
> head(data)

  exam sex hw1 hw2 hw3 hw4  degree high
1   34   F  18  41  27  10    arts    0
2   59   F  35  35  75  75   joint    0
3   35   F  58  51  37  42    arts    0
4   62   F  87  38  78  33   joint    0
5   86   F  90  98  90  92   joint    1
6   29   F  22  38   0  41 Science    0
```

Barplots are a good way to display categorical data, and are great way of assesing if there may be any trends worth exploring in the data. For example from the barplots below we may "unofficially" think that there may be a difference in students gaining over 80% in their exam depending on both their sex and degree program, i.e. more science male students seem to score over 80% over female students on the same degree. To asses if the expected proportions of people in each category differ significantly from the observed we would use a chi-squared test *(not explained here)*.

```
> par(mfrow=c(1,3))
> barplot(table(subset(data,data$degree=="Science")$sex,
+               subset(data,data$degree=="Science")$high),
+         names.arg=c("< 80%",">= 80%"),col=c("pink","lightblue"),
+         legend.text=c("Female","Male"),main="Science students",ylab="Number of students")
> barplot(table(subset(data,data$degree=="joint")$sex,
+               subset(data,data$degree=="joint")$high),
+         names.arg=c("< 80%",">= 80%"),col=c("pink","lightblue"),
+         legend.text=c("Female","Male"),main="Joint degree students",ylab="Number of students")
```

```
> barplot(table(subset(data,data$degree=="arts")$sex,
+               subset(data,data$degree=="arts")$high),
+         names.arg=c("< 80%",">= 80%"),col=c("pink","lightblue"),
+         legend.text=c("Female","Male"),main="Arts students",ylab="Number of students")
```



2.2 Correlation.

A correlation matrix can be used to display the correlations between each of the variables with each other, and is a good way to check for multi-collinearity.

```
> cor(data[,c("hw1","hw2","hw3","hw4","high")])

           hw1       hw2       hw3       hw4      high
hw1  1.0000000 0.3994821 0.4722623 0.5660572 0.3266803
hw2  0.3994821 1.0000000 0.5069044 0.5365741 0.3442244
hw3  0.4722623 0.5069044 1.0000000 0.5151317 0.3538881
hw4  0.5660572 0.5365741 0.5151317 1.0000000 0.3106489
high 0.3266803 0.3442244 0.3538881 0.3106489 1.0000000
```

3 The Logistic Regression Model.

$$ln(\frac{prob(event)}{1 - prob(event)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

The quantity to the left of the equal sign is called a logit *the log of the odds that an event occurs.* The odds that an event occurs is the ratio of the number of people who experience the event to the number of people who do not. The coefficients in the logistic regression model tell you how much the logit changes based on the values of the predictor variables.

3.1 The models and assesing fit.

```
> glm = glm(high ~ hw1 + hw2 + hw3 + hw4 + sex*degree, family=binomial(logit), data=data)
> summary(glm)
```

```
Call:
glm(formula = high ~ hw1 + hw2 + hw3 + hw4 + sex * degree, family = binomial(logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.54379  -0.26034  -0.06789  -0.00800   2.28083

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -18.21948    5.57684  -3.267  0.00109 **
hw1                  0.05991    0.03464   1.729  0.08376 .
hw2                  0.04188    0.03331   1.257  0.20869
hw3                  0.05467    0.03415   1.601  0.10935
hw4                  0.04238    0.02899   1.462  0.14379
sexM                 1.59376    1.83332   0.869  0.38467
degreejoint          1.62778    1.53061   1.063  0.28756
degreeScience       -0.39911    1.79086  -0.223  0.82365
sexM:degreejoint    -0.49087    2.24417  -0.219  0.82686
sexM:degreeScience   0.79878    2.33059   0.343  0.73180
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.098  on 90  degrees of freedom
Residual deviance: 36.502  on 81  degrees of freedom
AIC: 56.502

Number of Fisher Scoring iterations: 8
```

The above gives us regression coefficients estimates with standard errors and a z-test. None of the coefficients are significantly different from zero *(apart from the intercept term).* The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. So for example for every one unit change in homwwork 1 socre, the log odds of achieving above 80% in their exam (versus not obtaining above 80%) increases by 0.05991. The deviance was reduced by 31 points on 9 degrees of freedom, which gives us a p.value of 0.000296.

```
> 1 - pchisq(31, df=9)
```

```
[1] 0.0002960319
```

This indicates that the model appears to have performed quite well, showing a significant reduction in deviance significant at the (**0.01**%) level i.e. *(a significant difference from the null model).*

```
> anova(glm,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: high

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          90     67.098
hw1         1  12.7320        89     54.366 0.0003595 ***
hw2         1   7.9205        88     46.446 0.0048878 **
hw3         1   4.6906        87     41.755 0.0303288 *
hw4         1   0.5783        86     41.177 0.4469945
sex         1   2.1671        85     39.010 0.1409921
degree      2   2.1034        83     36.907 0.3493489
sex:degree  2   0.4050        81     36.502 0.8166960
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The print out above shows the same overall reduction in deviance, from 67.098 to 36.502 on 9 degrees of freedom. Here, the reduction in deviance is shown for each term, added sequentially first to last. Of note are the three first homework terms, which produced a significant reduction in deviance of 12.7320, 7.9205, and 4.6906 respectively on 1 degree of freedom each *(i.e. $p = 0.00003595$, 0.0048878, aand 0.0303288 respectively).*

3.2 Prediction.

Logistic regression, being based on the probability of an event occurring, allows us to calculate an odds ratio, which are the ratio of the odds of an event occurring to it not occurring, however in R we can also easily predict the probability of a student obtaining >= 80%. Lets for example predict the probability of a Female Science student scoring above 80% in her exam having achieved scores of 22, 38, 0 and 41 in homeworks 1 to 4 respectively.

```
> newdata<-data.frame("F",22,38,0,41,"Science")
> colnames(newdata)<-c("sex","hw1","hw2","hw3","hw4","degree")
> predict(glm,type="response",newdata=newdata)

           1
8.555251e-07
```

This tells us that such a student has a 0.0000008555251 probability of achieving equal to or above 80% in her exam.