The Retention Dataset

Jerry Kiely

27 March 2017

Introduction

This analysi is concerned with data relating to student retention in the Engineering faculty of DIT. The purpose of the analysis will be to use logistic regression models to identify and quantify relevant risk factors in student retention. The data includes risk factors regarding prior academic performance (e.g. leaving certificate results, leaving certificate maths grade), and personal characteristics (gender, home address, CAO choices made, etc.)

Variable name	Details
passed	Whether the student qualified to enter second year of their degree (0 = did not qualify, 1 = qualified)
gender	Male (1) or Female (0)
lc_points	Leaving certificate points achieved
mathgrd	Leaving certificate mathematics grade
CAO_choice	CAO ranked choice of programme entered
Address	Coded home address; 1 = Dublin, 2 = Dublin commuter belt, 3 = outside Dublin commuter belt

The Data

The data contains some extra columns that we don't need.

```r
colnames(retention)
```

```
## [1] "X"           "gender"      "passed"      "mathgrd"     "CAO_choice"
## [6] "address"     "lc_points"   "lc_points.1"
```

so we remove them:

```r
retention$X           = NULL
retention$lc_points.1 = NULL
```

we convert columns to factors:

```r
retention$mathgrd     = as.factor(retention$mathgrd)
retention$address     = as.factor(retention$address)
```

and we remove rows where NULLs or NAs are present:

```r
retention             = retention[complete.cases(retention),]
```

finally we have a look at the data:

```r
head(retention)
```

```
##   gender passed mathgrd CAO_choice address lc_points
## 1      0      0     80+          1       2       315
## 2      0      0     >20          1       1       270
## 3      0      1   50-60          1       2       370
## 4      0      1   20-30          2       2       295
## 6      0      0   20-30          1       1       260
## 7      0      1   35-45          1       1       280
```

The Model

First thing we do is fit a linear model to the data, including all possible interactions between the predictors.

Using either of the drop1 or the step functions we prune unimportant predictors from the model.

```
## Single term deletions
##
## Model:
## passed ~ gender + mathgrd + CAO_choice + lc_points + gender:CAO_choice
##                   Df Deviance    AIC     LRT Pr(>Chi)
## <none>                 288.38 316.38
## mathgrd            5   303.06 321.07 14.6893 0.011776 *
## lc_points          1   296.24 322.24  7.8668 0.005035 **
## gender:CAO_choice  3   296.23 318.23  7.8501 0.049212 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We are left with the following formula:

```
##
## Call:
## glm(formula = passed ~ gender + mathgrd + CAO_choice + lc_points +
##     gender:CAO_choice, family = binomial(link = "logit"), data = retention)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9148  -1.1097   0.6604   0.9128   2.0125
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.358323   1.038658  -3.233 0.001224 **
## gender             0.082133   0.482321   0.170 0.864784
## mathgrd20-30       1.224932   0.553829   2.212 0.026984 *
## mathgrd35-45       1.958951   0.582856   3.361 0.000777 ***
## mathgrd50-60       1.210271   0.592484   2.043 0.041081 *
## mathgrd65-75       0.753588   0.728172   1.035 0.300714
## mathgrd80+         1.350320   0.919415   1.469 0.141921
## CAO_choice2        2.749213   1.239096   2.219 0.026505 *
## CAO_choice3        1.187253   1.331843   0.891 0.372695
## CAO_choice4        1.114424   1.602384   0.695 0.486755
## lc_points          0.007958   0.002914   2.731 0.006322 **
## gender:CAO_choice2 -3.087517   1.291805  -2.390 0.016845 *
## gender:CAO_choice3 -0.832331   1.411712  -0.590 0.555466
## gender:CAO_choice4 -1.438172   1.649171  -0.872 0.383177
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 331.05  on 247  degrees of freedom
## Residual deviance: 288.38  on 234  degrees of freedom
## AIC: 316.38
##
## Number of Fisher Scoring iterations: 4
```

looking at the coefficient for lc_points, also it's log odds ratio for example, we see a value of 0.0079582 with an odds ratio of 1.00799 which would indicate that the odds of the student entering the second year of their degree would increase by 1.00799 for every leaving certificate points achieved.