Logistic Regression of the ISLR Default Dataset

Jon Kinsey

Wed Dec 3 20:52:35 2014

```
# Default dataset from Introduction to Statistical Learning with Applications in R
# This is a simulated data set containing information on ten thousand customers.
# http://www.statlearning.com
# http://cran.r-project.org/web/packages/ISLR/
# Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013)
# Introduction to Statistical Learning with applications in R, Springer-Verlag, New York
#
# The data frame has the following 4 variables.
#
# 1. default: A factor with levels No and Yes indicating whether the customer
# defaulted on their debt
#
# 2. student: A factor with levels No and Yes indicating whether the customer is a student
#
# 3. balance: The average balance that the customer has remaining on their credit
#    card after making their monthly payment
#
# 4. income: Income of customer
#
# The goal of this analysis is to predict which customers will default on their
# credit card debt.Cant use linear regression since ouputs are (0,1)
require(ISLR)
```

```
## Loading required package: ISLR
```

```
data(Default)
attach(Default)
head(Default)
```

```
##   default student balance income
## 1      No      No   729.5  44362
## 2      No     Yes   817.2  12106
## 3      No      No  1073.5  31767
## 4      No      No   529.3  35704
## 5      No      No   785.7  38463
## 6      No     Yes   919.6   7492
```

```
str(Default)
```

```
## 'data.frame':    10000 obs. of  4 variables:
##  $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
##  $ balance: num  730 817 1074 529 786 ...
##  $ income : num  44362 12106 31767 35704 38463 ...
```

```r
summary(Default)
```

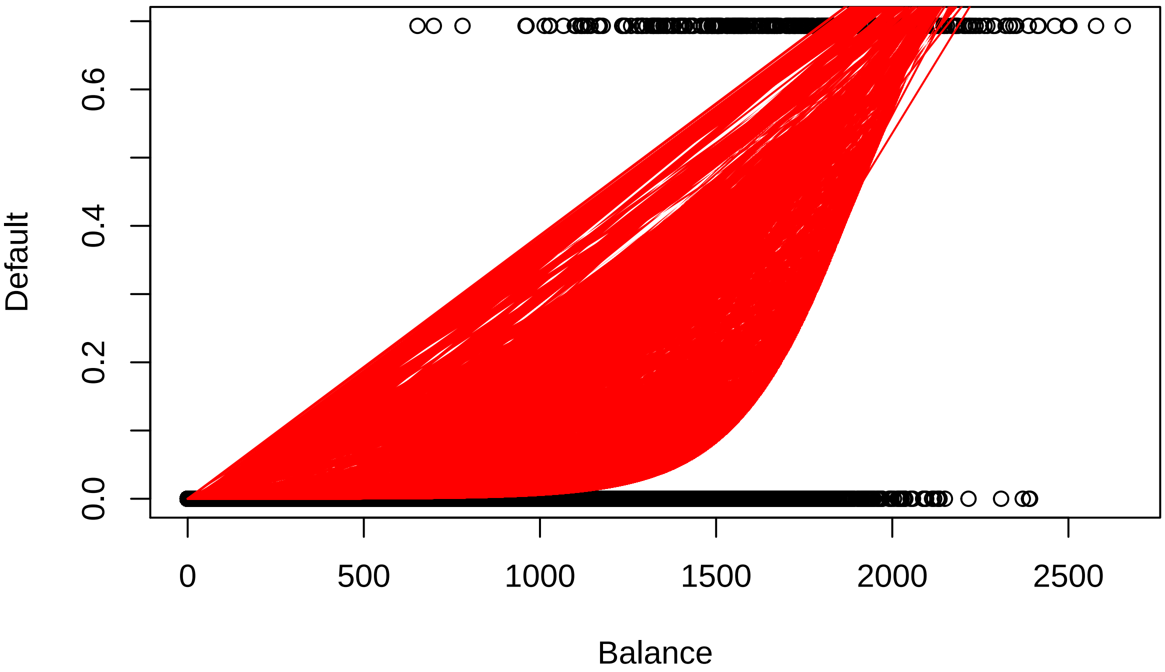
```
##  default    student       balance          income     
##  No :9667   No :7056   Min.   :   0   Min.   :  772  
##  Yes: 333   Yes:2944   1st Qu.: 482   1st Qu.:21340  
##                        Median : 824   Median :34553  
##                        Mean   : 835   Mean   :33517  
##                        3rd Qu.:1166   3rd Qu.:43808  
##                        Max.   :2654   Max.   :73554
```

```r
#
#Let us create the Logistic Regression Model for default vs balance using the glm function
glmodel=glm(formula=default ~ balance, data=Default, family=binomial)
summary(glmodel)
```

```
## 
## Call:
## glm(formula = default ~ balance, family = binomial, data = Default)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -2.270  -0.146  -0.059  -0.022   3.759  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -10.65133    0.36116   -29.5   <2e-16 ***
## balance       0.00550    0.00022    24.9   <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: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600
## 
## Number of Fisher Scoring iterations: 8
```

```r
# We see that beta_1 = 0.0055; this indicates that an increase in balance is
# associated with an increase in the probability of default. To be precise,
# a one-unit increase in balance is associated with an increase in the log
# odds of default by 0.0055 units.
#
# Since the p-value associated with balance in summary is tiny, we conclude that there
# is indeed an association between balance and probability of default.
# The estimated intercept in summary output is typically not of interest; its main
# purpose is to adjust the average fitted probabilities to the proportion of ones
# in the data.
# another: glm(default~student+balance+income,family="binomial",data=Default)
#
# As default has values 'Yes' and 'No' , we convert it to integer 1 and 2
# and take Log of it to limit values between 0 and 1
```

```
#
plot(x=balance,y=log(as.integer(default)),xlab="Balance",ylab="Default")
lines(balance,glmodel$fitted,type="l",col="red")
```



```
#
# Once the coefficients have been estimated, it is a simple matter to compute
# the probability of default for any given credit card balance.
predict(glmodel, newdata=data.frame(balance=2000), type="response")
```

```
##      1
## 0.5858
```

```
# Here, the predicted probability of default for an individual with a balance of
# $2, 000 is much higher, and equals 0.586 or 58.6%.
#
# So how much impact does student status make ?
Default[1:5,]
```

```
##   default student balance income
## 1      No      No   729.5  44362
## 2      No     Yes   817.2  12106
## 3      No      No  1073.5  31767
## 4      No      No   529.3  35704
## 5      No      No   785.7  38463
```

```
default.dummy <- rep(0,length(Default$default))
default.dummy[Default$default=="Yes"] <- 1
df.default <- data.frame(default = default.dummy,student = Default$student,
                         balance= Default$balance, income = Default$income )
mean(df.default$default[df.default$student=="Yes"])
```

```
## [1] 0.04314
```

```r
mean(df.default$default[df.default$student=="No"])
```

```
## [1] 0.0292
```

```r
# has some impact
```