DSI-06_Homework 4: Chapter 5, pg 221

Julia Gallucci

2023-02-20

6. We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.

```
#install.packages("ISLR") #install package containing Default dataset
library(ISLR) #load library
df <- Default #save dataset as a variable
head(df) #return the column names and first few rows of the dataset
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

(a) Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.

```
glm.fit <- glm(default ~ income + balance, family = "binomial", data = df)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
```

```
## balance      5.647e-03  2.274e-04  24.836  < 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: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

The standard errors of the coefficients are listed in the table above as std.error (Income -> 4.985e-06, Balance -> 2.274e-04)

(b) Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.

```
boot.fn <- function(data,index){ #this function takes two inputs, data and index
  glm.fit <- glm(default ~ income + balance, family = "binomial", data = data[index,]) #fit a glm on da
  coef(glm.fit)[2:3] #extract coefficients of income and balance, intercept info not needed
  }
```

(c) Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.

```
boot.coef <- boot::boot(df, boot.fn, R=1000) # use boot function with function made above on the data,
boot.coef
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = df, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##         original       bias     std. error
## t1* 2.080898e-05 3.696715e-07 4.999887e-06
## t2* 5.647103e-03 1.496030e-05 2.242637e-04
```

(d) Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.

standard error using glm -> income= 4.985e-06, balance = 2.274e-04 standard error using bootstrap -> income= 4.999887e-06, balance = 2.242637e-04

values obtained from bootstrap function are very similar to the standard formula for computing the standard errors in the glm() function!