STAT346: Statistical Data Science I

HW#6 – Due: December 21, 2020 by 6 p.m.

December 4, 2020

Instruction: Answer to the following questions and write your report using R Markdown. You should submit two files, through KU Black Board system (https://kulms.korea.ac.kr), which should have the following naming format:

- `stat346_hw6_your_id.rmd`
- `stat346_hw6_your_id.pdf` or `stat346_hw6_your_id.docx`

---

1. We will use logistic regression to predict the probability of default using income and balance on the `Default` data set.

```
library(ISLR)
head(Default)
```

```
##   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
```

   We now estimate the test error of this logistic regression model using the validation set approach.

   (a) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

   (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

      i. Split the sample set into a training set and a validation set.
      ii. Fit a multiple logistic regression model using only the training observations.
      iii. Obtain a prediction of `default` status for each individual in the validation set by computing the posterior probability of `default` for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.
      iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

(d) Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for `student` leads to a reduction in the test error rate

2. 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. Use `set.seed(1)` for this work.

(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.

(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.

(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`. See

```
library(boot)
? boot
```

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

3. Introduction to Data Science – Exercise 28.5: #1-2 (link).

4. Introduction to Data Science – Exercise 31.9: #1-4 (link).