CMSC 478 — Spring 2017 — C. S. Marron
Lab 5: Resampling Methods

Data Description

In this lab, you will work with the `Default` dataset of credit card default data. The data set consits of four variables:

- `default`: Yes/No; did the individual default.
- `student`: Yes/No; was the cardholder a student.
- `balance`: Cardholder's balance.
- `income`: Cardholder's income.

The dataset is part of the `ISLR` package; if you have not already installed the package, you may download the CSV file of the `Default` dataset.

Exercises

Exercise 1: Fit a logistic regression model that uses `income` and `balance` to predict `default` and assess the model using validation sets:

1. Fit the logistic regression model to all of the data; compute the training error of the model.
2. Use the validation set approach to estimate the test error of the model.
3. Use the validation set approach three more times *using different splits of the data* and compare the test error estimates from the four computations.

Exercise 2: Use 5-fold cross-validation to compare your model from Exercise 1 to a logistic regression model using `income`, `balance`, and a dummy variable for `student` to predict `default`.

1. Fit the logistic regression model from Exercise 1 and compute the 5-fold cross-validation error rate.
2. Fit the model using the dummy variable for `student` in addition to `balance` and `income`; compute the 5-fold cross-validation error rate.
3. Is there any evidence that the model including the dummy variable for `student` is better?

Exercise 3: Continuing with the logistic regression model to predict `default` using `income` and `balance`, use the bootstrap to estimate the standard error of the coefficients for `income` and `balance`:

1. Write a function `boot.fn()` that takes as input the `Default` dataset and an index of the observations, and that outputs the coefficient estimates of `income` and `balance`.

2. Use the `boot()` function along with your function `boot.fn()` to estimate the standard errors of the coefficients.

Exercise 4: Compare your boostrap estimates to those produced by standard R functions:

1. Use `summary()` and `glm()` to determine the estimated standard errors for for the coefficients of `income` and `balance`.

2. Compare the estimates with the bootstrap estimates from Exercise 3. Is there reason to trust one estimate over the other?