Simulation Based Regression Applications

YOUR NAME

12 November, 2020

Exercises

We will use the loans data set again to create linear models. Remember this data set represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals.

1. Loans

In this exercise we will examine the relationship between interest rate and loan amount.

a. Read in the data from `loans.csv` in the `data` folder.

b. Create a subset of data of 200 observations with the following three variables `interest_rate`, `loan_amount`, and `term`. Change `term` into a factor and use a stratified sample to keep the proportion of loan terms roughly the same as the original data.

c. Plot `interest_rate` versus `loan_amount`. We think `interest_rate` should be the response.

d. Fit a linear model to the data by regressing `interest_rate` on `loan_amount`. Is there a significant relationship between `interest_rate` and `loan_amount`?

e. Using the t distribution:

   i. Find a 95% confidence interval for the slope.

   ii. Find and interpret a 90% confidence interval for a loan amount of $20000

f. Repeat part e using a bootstrap.

g. Check the assumptions of linear regression.

2. Loans II

Using the `loans` data set of 200 observations from the previous exercise, use the variable `term` to determine if there is a difference in interest rates for the two different loan lengths.

a. Build a set of side-by-side boxplots that summarize interest rate by term. Describe the relationship you see. Note: You will have to convert the `term` variable to a factor prior to continuing.

b. Build a linear model fitting interest rate against term. Does there appear to be a significant difference in mean interest rates by term?

c. Write out the estimated linear model. In words, interpret the coefficient estimate.

d. Construct a bootstrap confidence interval on the coefficient.

e. Check model assumptions.