Intro to Data Science
Homework 6: Due Wednesday October 30 at 2:00pm

Exercises:

1. This exercise involves the use of simple linear regression on the Auto dataset from the ISLR package.

   (a) Create a scatter plot of the variables mpg versus horsepower from the Auto data set.

   (b) Describe your observations from the scatter plot.

   (c) What does the following command do and how should you interpret the result?

   ```
   with(Auto,cor(horsepower,mpg))

   ## [1] -0.7784268
   ```

   (d) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. That is, use the following command:

   ```
   auto_fit <- lm(mpg~horsepower,data=Auto)
   ```

   (e) What do you learn from the information output by the following commands:

   ```
    tidy(auto_fit)

   ## # A tibble: 2 x 5
   ##   term        estimate std.error statistic   p.value
   ##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
   ## 1 (Intercept)   39.9     0.717        55.7 1.22e-187
   ## 2 horsepower    -0.158   0.00645     -24.5 7.03e- 81
   ```

   and

   ```
    glance(auto_fit)

   ## # A tibble: 1 x 11
   ##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
   ##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
   ## 1     0.606         0.605  4.91      600. 7.03e-81     2 -1179. 2363. 2375.
   ## # ... with 2 more variables: deviance <dbl>, df.residual <int>
   ```

   (f) Make a plot of the residuals versus the fitted values from the regression. Recall that you can use the augment function in the broom package to create a data frame that adds the residuals and fitted values from the linear regression to the original data set. What do you conclude from this plot?

   (g) Compute the MSE and RMSE from the regression.

(h) Is there a relationship between the predictor and the response?

(i) How strong is the relationship between the predictor and the response?

(j) Is the relationship between the predictor and the response positive or negative?

(k) Plot the linear regression (remember that you can use geom smooth for this) along with the scatterplot of the data. What do you observe from this plot?

2. Use a bootstrap to approximate a 95% confidence interval for the slope parameter in the linear regression from problem 1. Be sure to make a plot of the bootstrap distribution.

3. Use a permutation test to test the null hypothesis: H_0 : slope is zero versus the alternative hypothesis H_A : slope is not equal to zero in the regression fit for problem 1. A plot is probably very helpful here.