CS 5565, LAB2(Linear Regression) 100 pts.

Name _____

1. View the video at the following URL and install R
   `https://www.youtube.com/watch?v=5ONFqIk3RFg`
   You may download the R Code for Labs and the Data Sets to use from the textbook website.
   `http://www-bcf.usc.edu/~gareth/ISL/`

2. This question involves the use of simple linear regression on the Auto data set.

   (a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

      i. Is there a relationship between the predictor and the response?
      ii. How strong is the relationship between the predictor and the response?
      iii. Is the relationship between the predictor and the response positive or negative?
      iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

   (b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

   (c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

3. This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Produce a scatterplot matrix which includes all of the variables in the data set.

   (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

   (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?
      ii. Which predictors appear to have a statistically significant relationship to the response?
      iii. What does the coefficient for the `year` variable suggest?

   (d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

   (e) Use the $*$ and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

   (f) Try a few different transformations of the variables, such as $\log(X), \sqrt{X}, X^2$. Comment on your findings.

4. This question should be answered using the `Carseats` data set.

   (a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.

   (b) Provide an interpretation of each coefficient in the model. Be carefulsome of the variables in the model are qualitative!

   (c) Write out the model in equation form, being careful to handle the qualitative variables properly.

   (d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$

   (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

   (f) How well do the models in (a) and (e) fit the data?

   (g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

   (h) Is there evidence of outliers or high leverage observations in the model from (e)?

5. In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
> set.seed(1)
> x = rnorm(100)
>y=2*x+rnorm (100)
```

   (a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command $\text{lm}(y \sim x + 0)$.)

   (b) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

   (c) What is the relationship between the results obtained in (a) and (b)?

   (d) For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^n x_{i'}^2}}.$$

   (These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{(\sqrt{n-1})\sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i'=1}^n y_{i'}^2) - (\sum_{i'=1}^n x_{i'}y_{i'})^2}}$$

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

(f) In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta_1 = 0.$ is the same for the regression of y onto x as it is for the regression of x onto y.