Homework 7

For all of the following examples include your code. I would suggest using RStudio for nice, presentable reports. Producing easy to read reports will tremendously help you explain your results to other people.

1. Load the Auto data set from the package ISLR using the commands

```
install.packages("ISLR")
library(ISLR)
data(Auto)
```

   a. Use the lm() function to perform a simple linear regression with mpg as the response (y variable) and horsepower (x variable) as the predictor. Use the summary() function to print the results. Comment on the output in the following ways.

   b. 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 (using the predict.lm function in R)?

   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.

2. The bureau of economic analysis releases data of gross domestic product (GDP), available at [http://www.bea.gov/national/xls/gdplev.xls]. The table below gives the last 10 years of GDP.

years	GDP
2006	13856
2007	14478
2008	14719
2009	14419
2010	14964
2011	15518
2012	16155
2013	16692
2014	17393
2015	18037

   a. Plot GDP (y-axis) vs Year (x-axis). What trends do you notice?

   b. Regress GDP (dependent variable) on Year (independent variable). Is the coefficient for year significant?

   c. Add the linear fit from part b to the plot in part a.

   d. Now download the full table through the link above. We are interested in the column labeled "GDP in billions of current dollars" and the year. You will have to put this in a form that is readable for R, see

the function read.table for suitable forms. Redo the analysis above, using all the years (1929 - 2015). What trends do you notice? Does a linear regression seem like a good fit?

e. Plot the residuals versus time. Do the errors appear homoscedastic, meaning independent of time?

3. Chapter 22, exercise 5 for $r = 1, 10,$ and 100

4. Chapter 22, exercise 9

5. Sample 10 observations from a uniform$(0, 1)$ distribution. Let's call these $x_1, \ldots, x_{10}$. Define $y_i = x_i^2$. Fit a linear regression model with y as the dependent variable and x as the independent variable. Plot x and y and the corresponding linear fit.

6. Up until the Civil War, the population of the US grew rapidly. Below is a table of census estimates of the population for several different years.

year	population
1790	3.93
1800	5.31
1810	7.24
1820	9.64
1830	12.9
1840	17.1
1850	23.2

a. Assess the fit of regressing population on different functions of time (use 1790 as time 0), including a linear function, quadratic function, and exponential. Plot the fits of every function. Which fit looks the best? Which fit is the best based on R^2? Does this makes sense, and why?

b. Use each fit to predict the population in 1860. The true population was 31.44 million. Which prediction was closest?

7. Suppose that you 100 observations $y_i, i = 1, \ldots, 100$ with covariates x_i. The relationship $y_i = \beta_0 + \beta_1 x_i + \epsilon$ but you fit both the linear and quadratic models. Which model will have the highest R^2 and why?