**STAT 632, HW 1**
Due: Thursday, February 3

**Reading**: Chapter 2 from *A Modern Approach to Regression.*
Chapter 3, pp. 59–71 from *An Introduction to Statistical Learning.*

**Directions**: Please submit your completed assignment to Blackboard. For the concept questions, your solutions may be typed (using LaTeX or equation editor in Word), or handwritten and then scanned (you can download a scanner application on your smart phone). For the data analysis questions, which require R, you must type your solutions. I suggest using R Markdown and knitting to PDF. If you are using Word, please convert your report to a PDF. Include all R code in your answers to each data analysis question.

**Exercise 0**. Provide a link to your Github page.

## Concept Questions

**Exercise 1**. The following is a regression summary from R for a linear regression model between an explanatory variable $x$ and a response variable $y$. The data contain $n = 50$ points. Assume that all the conditions for SLR are satisfied.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1016     0.4082  -2.699  ______  **
x             2.2606     0.0981   _____  < 2e-16 ***
```

(a) Write the equation for the least squares regression line.

(b) R performs a t-test to test whether the slope is significantly different than 0. State the null and alternative hypothesis for this test. Based on the $p$-value what is the conclusion of the test (i.e., reject or do not reject the null hypothesis)?

(c) Calculate the missing $p$-value for the intercept.

(d) Calculate the missing t-statistic for the slope.

(e) Calculate a 95% confidence interval for the slope of the regression line. Does this interval agree with the results of the hypothesis test?

**Exercise 2**.[1] Consider the linear regression model through the origin given by $Y_i = \beta x_i + e_i$ for $i = 1, \cdots, n$. Assume $e_i \sim N(0, \sigma^2)$, that is, the errors are independent and normally distributed with constant variance.

(a) Show that the least squares estimate of the slope is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(Hint: minimize $R(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$ by taking the derivative and setting the derivative equal to zero.)

(b) Show that $E(\hat{\beta}) = \beta$

(c) Show that $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

[1] From *A Modern Approach to Regression with R*, Chapter 2, Exercise 4, with slight modifications

## Data Analysis Questions

**Exercise 3**.[2] The web site `www.playbill.com` provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11–17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3–10, 2004). The data are available on the book web site `http://gattonweb.uky.edu/sheather/book/` in the file playbill.csv.

Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$ where $Y$ is the gross box office results for the current week (in dollars) and $x$ is the gross box office result for the previous week (in dollars). Complete the following tasks:

(a) Use `read.csv()` to load the `playbill.csv` data file into R. Make a scatter plot of the response versus the explanatory variable, and superimpose the least squares regression line.

(b) Calculate a 95% confidence interval for the intercept and slope of the regression model, $\beta_0$ and $\beta_1$ [hint: use the `confint()` function]. Is 1 a plausible value for $\beta_1$?

(c) Use the fitted regression model to estimate the gross box office results for the current week (in dollars) for a production with \$400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in dollars) for a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000 in gross box office the previous week?

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

[2] From *A Modern Approach to Regression with R*, Chapter 2, Exercise 1, with slight modifications

**Exercise 4.**[3] For this question use the `oldfaith` data set form the `alr4` package. To access this data set first install the package using `install.packages("alr4")` (this only needs to be done once). Then load the package into R with the command `library(alr4)`. Documentation for the data set can be read in the help menu by entering the command `help(oldfaith)`.

The `oldfaith` data set gives information about eruptions of Old Faithful Geyser during October 1980. Variables are `Duration` in seconds of the current eruption, and the `Interval`, the time in minutes to the next eruption. The data were collected by volunteers and were provided by the late Roderick Hutchinson. Apart from missing data for the period from midnight to 6 a.m., this is a complete record of eruptions for that month.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

(a) Use the `lm()` function to perform a simple linear regression with `Interval` as the response and `Duration` as the predictor. Use the `summary()` function to print the results.

(b) Make a scatter plot of `Interval` versus `Duration`. Superimpose the least squares regression line on the scatter plot.

(c) An individual has just arrived at the end of an eruption that lasted 250 seconds. What is the predicted amount of time the individual will have to wait until the next eruption? Calculate a 95% prediction interval for the time the individual will have to wait for the next eruption.

(d) Interpret the coefficient of determination ($R^2$).

[3] From Weisberg S., *Applied Linear Regression*, Fourth edition, Exercise 2.20, with slight modifications