Assignment 1

Last name: DU
First name: MIN
Student ID: 1002602230
Course section: STA302H1F-Summer 2017

Due Date: May 25, 2017, 23:00

Q1 (4 pts) - Typing mathematical notations.

Q1-a: Show that $\sum_i^n (X_i - \bar{X}) = 0$

Proof:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}) &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} \\ &= \sum_{i=1}^{n} X_i - n\bar{X} \\ &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i \\ &= 0\end{aligned}$$

Q1-b (2 pts): Show that $\sum_i^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$

Proof:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\bar{X} + \sum_{i=1}^{n} \bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2\end{aligned}$$

Q1-c (2 pts): Show that $\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_iY_i - n\bar{X}\bar{Y}$

Proof:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}) \\ &= \sum_{i=1}^{n} X_iY_i - 2n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\ &= \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}\end{aligned}$$

Q2 (8 pts) - Answer the following questions

Q2-a (2 pts)

When asked to state the simple linear regression model, a student wrote it as follows

$$E(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Do you agree? And give your reasoning.

Answer: Disagree,

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i).$$

where ϵ_i is an unknown random varaible, according to Gauss Markov Assumption,

$$E(\epsilon_i) = 0$$

Therefore, the model should be

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Q2-b

The **oldfaithful.txt** data set contains data on 21 consecutive eruptions of Old Faithful geyser in Yellowstone National Park. It is believed that one can predict the time until the next eruption (next), given the length of time of the last eruption (duration). That is, Y is the "eruption" and X is the "waiting" in the data set.

- (2 pts) Fit a simple linear regression (show R code)

```r
q2data = read.table("/Users/Joy/Desktop/STA302/oldfaithful.txt",header=TRUE)
str(q2data)     #check the type of each column (variable) in the data set
```

```
## 'data.frame':    272 obs. of  2 variables:
##  $ eruption: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ waiting : int  79 54 74 62 85 55 88 85 51 85 ...
```

```r
head(q2data,10) # have a look of the first 10 data lines
```

```
##    eruption waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
## 7     4.700      88
## 8     3.600      85
## 9     1.950      51
## 10    4.350      85
```

```r
# write R code to fit the data with a simple linear regression
lm_fit = lm(eruption~waiting, data = q2data)
```

- (2 pts) Show the summary output of the simple linear regression.

```r
# Produce the summary output from R
summary(lm_fit)
```

```
##
## Call:
## lm(formula = eruption ~ waiting, data = q2data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

- (2 pts) What is the estimated linear regression model? (replace the following b_0 and b_1 with their estimates)

$$\widehat{eruption} = -1.87 + 0.076 * waiting$$