STAC67: Assignment 2

Deadline to hand in: **Feb. 22 (Monday) 11:59 pm, 2021**

Total: 85 points

Q. 1 (25 points) Give observations on Y are to be taken when $X = 4, 8, 12, 16, 20$, respectively. The true regression function is $E(Y) = 20 + 4X$, and the ϵ_i are independent $N(0,\ 5^2)$. When you generate a random number, use R code, **set.seed(your student number)** before the R codes of generating a random number, so that we can replicate the result.

(a) (5 pts) Calculate $P(|\hat{\beta}_1 - \beta_1| > 1)$ where $\hat{\beta}_1$ is the least squares estimator of β_1, compute it by hands.

(b) (5 pts) Generate five random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at $X = 4, 8, 12, 16, 20$ and calculate $Y_1, Y_2, \ldots, Y_5$. Obtain the least square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, when fitting a linear regression line to the five cases. Also calculate $\hat{Y}_0$ when $X_0 = 10$ and obtain a 95% confidence interval for $E(Y_0)$ when $X_0 = 10$.

(c) (10 pts) Repeat part (b) 1,000 times, generating new random numbers each time, and make a histogram distribution of 1,000 estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. calculate the mean and standard deviation of the 1,000 estimates. Are the results consistent with theoretical expectations?

(d) (5 pts) What proportion of the 1,000 confidence intervals for $E(Y_0)$ when $X_0 = 10$ include $E(Y_0)$? Is this result consistent with theoretical expressions?

Q. 2 (35 points) The dataset "NHLhtwt.csv" is posted at Quercuse. It contains weights (Y, in pounds) and heights (X, in inches) for 717 National Hockey League players for the 2013/2014 season. Complete the following parts (treating this as a sample from a conceptual population of potential athletes).

(a) (5 pts) Fit a Simple Linear Regression relating Weight (Y) to (X) using R. Construct 95 % confidence interval for the mean weight of all players with $X_0 = 74$. Compute it by hands (use R) and compare the result with the built-in R function.

(b) (5 pts) Construct a 95% prediction interval for a new player with $X_0 = 74$. Compute it by hands (use R) and compare the result with the built-in R function.

(c) (5 pts) Plot the residuals versus fitted values.

(d) (5 pts) Obtain a normal probability plot of residuals and test the hypothesis that the errors are normally distributed with the Shapiro-Wilk test.

(e) (10 pts) We would like to conduct the **Brown-Forsythe test** to determine whether or not the error variance varies with the level of X. Divide the data into the two groups based on the median of X. Use $\alpha = 0.05$. Do not use the built-in R function. Write your own R function to implement this test. What is your test result?

(f) (5 pts) If there is evidence of non-normality or non-constant variance of errors, obtain a Box-Cox transformation, and repeat the previous parts.

Q. 3 (10 points) For the multiple regression, the design matrix $\boldsymbol{X}$, the response vector $\underset{\sim}{Y}$, and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are presented below:

$$\boldsymbol{X} = \begin{pmatrix} 1 & -9 & 3 \\ 1 & -7 & -7 \\ 1 & -5 & 7 \\ 1 & -3 & -9 \\ 1 & -1 & 1 \\ 1 & 1 & 5 \\ 1 & 3 & -1 \\ 1 & 5 & -3 \\ 1 & 7 & -5 \\ 1 & 9 & 9 \end{pmatrix} \qquad \underset{\sim}{Y} = \begin{pmatrix} 34 \\ 16 \\ 26 \\ 35 \\ 32 \\ 11 \\ 24 \\ 1 \\ -3 \\ 15 \end{pmatrix} \qquad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{1067840} \begin{pmatrix} 106784 & 0 & 0 \\ 0 & 3300 & -460 \\ 0 & -460 & 3300 \end{pmatrix}$$

(a) Calculate the least squares estimates of the model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

(b) Calculate the fitted values from your model and hence calculate an unbiased estimate of the error variance σ^2.

Q. 4 (15 points) A foam beverage insulator (beer hugger) manufacturer produces their product for firms that want their logo on beer huggers for marketing purposes. The firms cost analyst wants to estimate their cost function. She interprets β_0 as the fixed cost of a production run, and β_1 as the unit variable cost (or marginal cost). Based on n = 5 production runs she observes the following pairs (X_i, Y_i) where X_i is the number of beer huggers produced in the ith production run (in 1000s), and Y_i was the total cost of the run (in \$1000). Let's assume that the variance of random error is $\sigma^2 = 4$.

i :	1	2	3	4	5
X_i :	3	5	4	6	7
Y_i :	4	6.5	5	7	7.5

(a) (6 pts) Using matrix methods, obtain the following:

(1) Design matrix **X**

(2) vector of estimated regression coefficients, $\underset{\sim}{\hat{\beta}}$

(3) variance-covariance matrix of $\underset{\sim}{\widehat{\beta}}$

(b)(3 pts) From variance-covariance matrix in part (a), obtain the following:

(1) $Cov\{\widehat{\underset{\sim}{\beta_0}}, \widehat{\underset{\sim}{\beta_1}}\}$

(2) $Var\{\widehat{\underset{\sim}{\beta_0}}\}$

(3) $Var\{\widehat{\underset{\sim}{\beta_1}}\}$

(c) (3 pts) Find the hat matrix $\boldsymbol{H}$.

(d) (3 pts) Find $Var\{\underset{\sim}{e}\}$.