Statistics 206

Homework 3

NOT DUE

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/Homework/hwX") before its due time.
- For the written part of the homework, please make it into a **.pdf file**. It may be prepared by a word processor (e.g., Latex), or by writing on a tablet or on pieces of papers. You may need to scan your work.
- Please make sure the pages are in order and the answers are numbered and legible.
- **Optional Problems** are not counted towards the grade.
- By submitting homework under your name, you acknowledge that you are the person who prepared the submitted work.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. **A simple linear regression case study by R**. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 *women from each of the four* 10*-year age groups, beginning with age* 40 *and ending with age* 79*. Two variables being measured are: age* (X) *and the amount of muscle mass* (Y)*. Data are stored in the file "muscle.txt".*

   (a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?

   (b) Use the Box-Cox procedure to decide whether a transformation of the response variable is needed.

   (c) Perform linear regression of the amount of muscle mass on age and obtain the R "summary" output. From the summary, obtain the estimated regression coefficients and their standard errors, the mean squared error (MSE) and its degrees of freedom.

   (d) Write down the fitted regression line. Add the fitted regression line to the scatter plot. Does it appear to fit the data well?

(e) Obtain the fitted values and residuals for the 6th and 16th cases in the data set.

(f) Draw the residuals vs. fitted values plot and the residuals Normal Q-Q plot. Write down the simple linear regression model with Normal errors and its assumptions. Comment on these assumptions based on the residual plots.

(g) Construct a 99% confidence interval for the estimated regression intercept. Interpret your confidence interval.

(h) Conduct a test at level 0.01 to decide whether or not there is a negative linear association between the amount of muscle mass and age. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. *(Hint: Which form of alternatives should be used here?)*

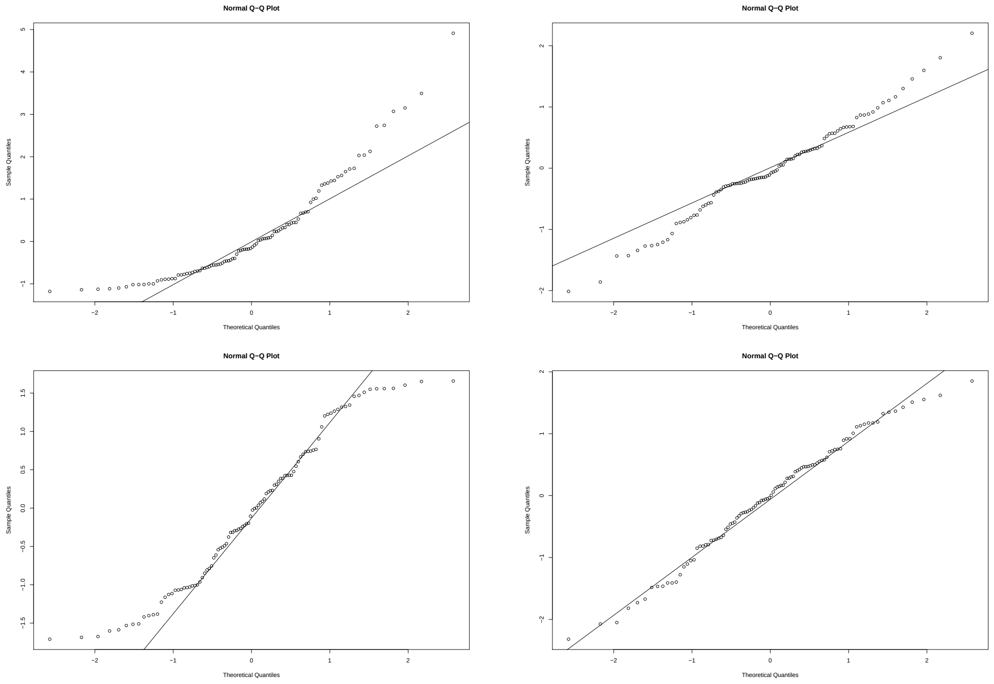
(i) Construct a 95% prediction interval for the muscle mass of a woman aged at 60. Interpret your prediction interval.

(j) Obtain the ANOVA table for this data. Test whether or not there is a linear association between the amount of muscle mass and age by an F-test at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

(k) What proportion of the total variation in muscle mass is “explained” by age? What is the correlation coefficient between muscle mass and age?

2. **Q-Q plots**. For each of the Q-Q plot in Figure 1, describe the distribution of the data (whether it is Normal or heavy tailed, etc.).

Figure 1: Q-Q plots



3. **Coefficient of determination.** Show that

$$R^2 = r^2, \quad r = \text{sign}\{\hat{\beta}_1\}\sqrt{R^2},$$

where R^2 is the coefficient of determination when regressing Y onto X and r is the sample correlation coefficient between X and Y.

4. Confirm the formula for inverting a 2×2 matrix.

5. **Projection matrices**. Show the following are projection matrices, i.e., being symmetric and idempotent. What are the ranks of these matrices? Here $\mathbf{H}$ is the hat matrix from a simple linear regression model with n cases (where the X values are not all equal) , $\mathbf{I}_n$ is the $n \times n$ identify matrix, and $\mathbf{J}_n$ is the $n \times n$ matrix with all ones.

   (a) $\mathbf{I}_n - \mathbf{H}$

   (b) $\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$

   (c) $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$

6. Under the simple linear regression model, using matrix algebra, show that: The residuals vector $\mathbf{e}$ is uncorrelated with the fitted values vector $\hat{\mathbf{Y}}$ and the LS estimator $\hat{\boldsymbol{\beta}}$.

   (Hint: If $\mathbf{Z}$ *is an* $r \times 1$ *random vector,* $\mathbf{A}$ *is an* $s \times r$ *non-random matrix, and* $\mathbf{B}$ *is a* $t \times r$ *non-random matrix, then* $Cov(\mathbf{AZ}, \mathbf{BZ}) = \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{Z}\}\mathbf{B}'$*.)*