BIST 0610J, Spring 2022
Homework #1
due Tue., 2/8/2022

Total 80 Points

1. (40 points) A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 60 women (15 women from each 10-year age group, beginning with age 40 and ending with age 79). Let X be age, and Y a measure of muscle mass. The data "muscle.csv"(the 1st column includes Y and the 2nd for X) can be found on Canvas.

   (a) (5 points) Draw a scatter plot of Y vs. X. Comment on the relationship between Y and X in a sentence or two.

   (b) Is there a linear association between Y and X?

      i. (3 points) Please justify your answer with the estimate of correlation coefficient, and assess it for significance by providing 95% confidence interval.

      ii. (2 points) Summarize the result in a sentence or two.

   (c) (5 points) Fit the following model using PROC REG,

   $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

   Provide $\widehat{\beta}_1$, $\widehat{\beta}_0$, $\mathrm{Var}(\widehat{\beta}_1)$, $\mathrm{Var}(\widehat{\beta}_0)$, and $\widehat{\sigma}^2$.

   (d) (6 points) Give interpretations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ you obtained in (c), and provide 95% confidence intervals for β_0 and β_1, respectively.

(e) (6 points) One wants to know if mean muscle mass is **linearly** associated with age. Perform an appropriate test:

- list H_0 and H_a;
- write out the test statistic (formula);
- state the test statistic's distribution under H_0;
- compute the test statistic;
- draw a conclusion.

(f) (8 points) Please calculate the mean value of X, denote by $\bar{X}$. For each X_i, compute a new variable $X_i^* = X_i - \bar{X}$.

i. Fit a simple linear regression model $Y_i = \beta_0^* + \beta_1^* X^* + \epsilon_i^*$ with $\epsilon_i^* \sim N(0, \sigma^{*2})$, for $i = 1, \cdots, 60$, and provide estimates of β_0^*, β_1^* and σ^{*2}.

ii. Give interpretations for $\widehat{\beta}_0^*$ and $\widehat{\beta}_1^*$.

iii. Compare $\widehat{\beta}_0^*$, $\widehat{\beta}_1^*$ & their standard errors and estimate of σ^{*2} with $\widehat{\beta}_0$, $\widehat{\beta}_1$ & their standard errors and estimate of σ^2 from the model in (c). Summarize your findings.

iv. What is the advantage of using X_i^* over X_i?

(g) (5 points) Determine the ANOVA table for the straight-line regression of Y on X, and test the significance of the straight line regression using an F-test.

2. (18 points) Consider a 'regression-through-the-origin' model, $Y_i = \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \cdots, n$.

(a) (4 points) Derive a least squares estimator of β_1 for this model. Show all your work

(b) (10 points) Provide the mean squared error (mse) for $\widehat{\beta}_1$. Show all your work. (hint: $\text{mse}(\hat{\beta}_1) = \text{Var}(\hat{\beta}_1) + \text{bias}(\hat{\beta}_1)^2$)

(c) (4 points) What is the distribution of $\widehat{\beta}_1$? Justify your answer.

3. (12 points) The following table gives rat body weights (in grams) and latency to seizure (in minutes), following injection of 40 mg/kg of body weight of metrazol.

Latency	Weight	Latency	Weight
2.30	348	2.00	409
1.95	372	1.70	413
2.90	378	2.00	415
2.30	390	2.95	423
1.10	392	1.25	428
2.50	395	2.05	464
1.30	400	3.70	468

(a) (4 points) Using regression-through-the-origin model, provide $\widehat{\beta}_1$ and $\widehat{\sigma}^2$, and test for $\beta_1 = 0$.

(b) (4 points) Consider a regression model with intercept, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, and test for $\beta_0 = 0$ and for $\beta_1 = 0$. Provide $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$, and test for $\beta_0 = 0$ and $\beta_1 = 0$, respectively.

(c) (4 points) Provide comments on your findings from (a) and (b).

4. (10 points) Prove the following expression is true.

$$t^2 = F$$

where t is t-test statistic and F is F-test statistic, both for testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$ in a simple linear regression.