This exam contains 5 questions. The total number of points for the questions is 100. The total time allowed for the exam is 75 minutes. It is suggested that you allocate to each question an amount of time proportional to the number of points that the question is worth. You should provide explanations for your answers. Right answers with no explanation will receive no credit. On the other hand, wrong answers with thoughtful explanations may receive some credit. Partial credit will be given. **GOOD LUCK!**
QUESTION 1 (20 POINTS)

Suppose that you are the head of a company which makes flour. Your quality control expert tells you that the net weight of a bag of flour produced by your factory is expected to be 5 pounds with a standard deviation of 0.5. However, you suspect that he might have overstated the expected net weight. To test for this, you sample 4 bags and find the average weight of these 4 bags to be 4.8. Carefully state the null and alternative hypotheses in this situation and carry out a statistical test at the 5% significance level based on the information provided, assuming that net weight follows a normal distribution. What is the p-value in this case? Can you reject the null hypothesis at the 5% significance level? Suppose now that the true expected net weight for a bag of flour is 4.9 pounds. What is the probability of a Type II error (i.e., the error of failing to reject the null hypothesis when it is false) for your test? Do you see a problem here? If so, what might have contributed to the problem? Explain your answers carefully.

QUESTION 2 (20 POINTS)

Let $Y_1, Y_2, ..., Y_n$ be $i.i.d.$ random variables with mean $\mu_Y$ and variance $\sigma^2_Y$ such that $0 < \sigma^2_Y < \infty$. Consider an estimator of the population mean $\mu_Y$ of the form

$$\tilde{\mu}_Y = \frac{1}{2n} \sum_{i=1}^{n} Y_i.$$

(a) Derive a formula for the variance of $\tilde{\mu}_Y$.

(b) Compare the variance formula you have derived in part (a) with the variance for

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Which estimator has the bigger variance, $\tilde{\mu}_Y$ or $\bar{Y}_n$? Is this a violation of the Gauss-Markov Theorem? Why or why not? Support your answer carefully with the relevant algebraic calculations.
QUESTION 3 (10 POINTS)
Consider the linear regression

\[ Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n. \]

Suppose that the following three assumptions hold:

A1 \((X_i, u_i)\) are \(i.i.d.\) for \(i = 1, 2, \ldots, n\)

A2 \(E[u_i | X_i] = c \neq 0\) with probability one for \(i = 1, \ldots, n\)

A3 \(E[X_i^4] < \infty\) and \(E[u_i^4] < \infty\), i.e., large outliers are rare.

Under this setup, is the OLS estimator of \(\beta_1\) unbiased? Provide the relevant algebraic calculations to support your answer.

QUESTION 4 (20 POINTS)

Suppose that you are interested in the relationship between the height of children and that of their parents. To do some research on this problem, you collect data from 110 college students and estimate the following linear regression

\[
\text{Studenth} = 19.6 + 0.73 \times \text{Midparh}, \quad R^2 = 0.45, \ SER = 2.0,
\]

where \(\text{Studenth}\) is the height of students in inches, and \(\text{Midparh}\) is the average of the parental heights. Values in parentheses are the heteroskedasticity robust standard errors.

(a) Test the null hypothesis that average parental height has no effect on the height of their children versus the alternative hypothesis that there is a non-zero effect. Carefully state the null and alternative hypotheses in terms of hypotheses concerning the parameters of the linear model, and carry out a test at the 5% significance level.

(b) Construct a 95% confidence interval for the slope coefficient of the regression above.

(c) Is the result in part (b) consistent with the test result obtained in part (a)? Explain your answer carefully.
QUESTION 5 (30 POINTS)

Consider estimating the linear regression

\[ Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n, \]

using ordinary least squares (OLS). To assess how well the regression line estimated by OLS fit the data points, a commonly-used statistic is the following squared (sample) correlation coefficient

\[ \hat{\rho}^2 = \frac{\left[ \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}) (Y_i - \overline{Y}) \right]^2}{\left[ \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right] \left[ \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \right]}, \]

where

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad \overline{\hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i. \]

(a) Show that, in this case, \( \overline{Y} = \overline{\hat{Y}} \). Use this result to show that \( \hat{\rho}^2 \) is just the usual \( R^2 \) measure. Justify your answer carefully.

(b) Suppose that \( \hat{\beta}_1 = 0 \); what is the numerical value of \( R^2 \) in this case? Justify your answer with the relevant algebraic calculations. Does your answer have an intuitive interpretation?

(c) Suppose that

\[ \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = 3, \]
\[ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 5. \]

Calculate the numerical value of the \( R^2 \) measure in this case. Show your work.