Homework 2 (87 pts)

Note:

For the following problems, we use the life expectancy data and consider a simple linear regression $Y \sim X$, where $X = X_2$ is the number of nurses and midwives, and Y is the life expectancy.

In the confidence interval problems, note that components in a confidence interval include the point estimate, the critical value, the standard error, and the margin error. The result should be computed toward the end.

For example, compute a CI as $5 \pm 2 * 4 = 5 \pm 8 = (-3, 13)$

In computation problems, a basic rule is that you keep 3 or more significant decimal places for numbers during the working period and keep 2 or more significant decimal places at the number reported at the end.

In the fill-in-the-blank question, when denote or write the formula for a term, show **both the general and the specific form based on the question**. Remember to **fill your answer in the blanks or above the line** to be graded properly.

For example, the critical value for a one-sided t-test, $H_0\ \mu = 0, H_a: \mu > 0$ is denoted by $t(1-\alpha, n-1) = t(0.95, 30)$.
The test statistic, t_s, can be computed with a formula $t_s = \frac{\bar{Y}}{s/\sqrt{n}} = \frac{10}{20/\sqrt{25}}$, and a value of 2.5, where the general form is $t_s = \frac{\bar{Y}}{s/\sqrt{n}}$, and the specific form is $\frac{10}{20/\sqrt{25}}$,
The p value can be computed with a formula $\Pr(t > t_s \mid \mu_0 \text{ is true}) = \Pr(t > 2.5, \text{ given } \mu_0 = 0)$, where the general form is $\Pr(t > t_s \mid \mu_0 \text{ is true})$ and the specific form is $\Pr(t > 2.5, \text{ given } \mu_0 = 0)$,

Unless stated otherwise in each problem, please use a significance level of $\alpha = 0.05$ or a confidence level of $1 - \alpha = 0.95$.

Problem 1 – Compare the hypothesis test between the linear impact and linear correlation (10 pts, no partial credit)

Using the R-generated summary and ANOVA table for the model $Y \sim X$, answer the following questions.

(a) [6 pts] For a two-sided hypothesis test on the linear impact, $Ho: \beta_1 = 0, H_a: \beta_1 \neq 0$ if a T-test is used, the test statistic is computed with the formula: _____, which is computed as _____(value); The critical value has the notation of _____, and a value of _____.

The p-value of the test can be computed with the formula _____, and the value is _____.

(b) [2 pts] **Verify your answer** by highlighting the corresponding p-values in the R output for the T-test .

(c) [2 pts] Adjust the HT components from a two-sided test to a one-sided test. Consider the one-sided HT $Ho: \beta_1 = 0, H_a: \beta_1 > 0$, the test statistic is the same as the two-sided test, but the p-value needs to be adjusted with the formula_____ and computed as _____in this question.

Problem 2 (8 pts)

For a two-sided hypothesis test on the significance of the linear correlation coefficient between X and Y, $H_0: \rho = 0, H_a: \rho \neq 0$

(a) [2 pts] if a T-test is used, the test statistic is computed with the sample correlation, r, with the formula: _____, which is computed as _____(value).

(b) [1 pt] Is this test statistic the same as the t-test in part (a) _____(Y/N).

(c) [2 pts] Discuss when the results of the hypothesis test on the linear impact and the linear association are equivalent.

(d) [3 pts] Use R to compute a 95% confidence interval for the linear correlation coefficient between Y and X. Use the confidence interval to verify the hypothesis test in (c). (hint: if the confidence interval contains the hypothesized value, then the two-sided hypotheses should be rejected or not?)

For the following problems, we focus first on the lack of fit test and then the diagnostic and remedy procedures applied to a Simple Linear Regression (SLR) model.

Specifically, we will gain familiarity with the components involved in the lack of fit test and understand how they vary based on the data.

In the diagnosis procedure of the SLR model, we mainly identify assumption violations by examining the residuals. And for the remedial procedure of the SLR model, we mainly perform simple transformations on either X, Y, or both variables. These methods are still applicable in the diagnosis and remedies of Multiple Linear Regression (MLR) models in future topics. As we delve into MLR analysis, additional techniques will be introduced to build a more efficient model by considering the relationships among multiple predictors and their collective contribution to the model.

Problem 3 (24 pts)

Consider a simple linear regression model with $Y \sim X$ on the following table. Practice doing the **lack-of-fit by hand**.

ID	X	Y
1	0.24	16
2	0.22	40
3	0.23	32
4	0.24	13
5	0.24	1
6	0.24	1
7	0.22	2
8	0.21	3
9	0.24	8
10	0.21	14

(a) [4 pts] Based on a (R-generated) scatter plot of X and Y with the regression line, comment on whether the Simple Linear Regression (SLR) exhibits a lack-of-fit issue.

(b) [10 pts] Compute the components for the lack of fit test: c, $\hat{Y}_i, \bar{Y}_i, \bar{Y}$, SSPE, SSLF, SSE, DFPE, DFLF. DFE.

(c) [5 pts] Next, consider the lack-of-fit test for the SLR. Define Ho/Ha, calculate test statistic, define reject region, compute the p-value, and state the conclusion.

(d) [5 pts] Utilize R to conduct the lack-of-fit test and identify as many components as possible from the ones computed in part (b) in the R output.

Problem 4 (10 pts)

Consider a lack of fit test **on the following data** on $Y \sim X$

X	Y
0.19	65
0.44	30
0.35	22
0.32	31
0.29	9

(a) [2 pts] Can you perform a lack of fit test on this data? Explain.

(b) [4 pts] Suppose a new row is added: X 0.19, Y 59 and the sample size is now 6. Then the SSPE is _____(increased/decreased) by _____. dfPE=_____ and dfLF=_____.

(c) [4 pts] Suppose the data is grouped by the tenth digits as follows,

X	Y
0.1	65
0.4	30
0.3	22
0.3	31
0.2	9

Then the SSPE is _____(increased/decreased) by _____; the dfPE=_____ and dfLF=_____.

Problem 5 – SLR diagnostic process (15 pts)

Utilize the life expectancy data and examine a simple linear regression $Y \sim X$, where $X = X_3$ (pharmacists), and Y represents life expectancy. Employ R to assess assumptions for the SLR.

Conduct screening through the scatter plot, residual plot on X, residual plot on $\hat{Y}$, Shapiro test, Brown-Forsythe test, Breusch-Pagan test, and provide comments on:

- Linear relationship
- Constant variance
- Normal errors
- Outliers

Problem 6 (5 pts)

Do you think it is necessary to transform X or Y? Explain.

Problem 7 – SLR remedy process (15 pts)

Utilize R to conduct a Box-Cox transformation on Y, then proceed with the same diagnostic process as in Problem 5. Compare the transformed model with the original model.