NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

ST3131 - Regression Analysis

(Semester 1 AY2025/2026)

Midterm Test

Time Allowed: 80 minutes

INSTRUCTIONS TO STUDENTS

1. Students are required to complete this test individually.
2. This is an OPEN BOOK BLOCK INTERNET exam. You may refer to your lecture notes, tutorials, textbooks or any notes that you have made, but you are not allowed to perform any online search.
3. The use of offline Large Language Models (LLMs) is prohibited for this test.
4. Both programmable calculators and non-programmable calculators are allowed.
5. Students are required to answer ALL questions. Total mark is 50.
6. Hand-in your answer booklet to the invigilators at the end of the test.
7. Report numerical answers to at least three significant figures if it's absolute is smaller than one (e.g. 0.0123) and to three decimal places if it's absolute is larger than one (e.g. -2.345).

1. (*40 points*) Data set `house-prices-OR.csv` was given on Canvas. Data were collected in the US that contain the price of houses and some information about the houses when they were sold. The description of variables of our concern are given in the table below.

Column's Name	Description
HP.in.thousands	price when the house was sold ($1000)
House.Size	house's size (sqft)
Acres	the total area of the land has has the house (acre)
Bedrooms	number of bedrooms
Garage	1 = Yes; 0 = No
Age.Category	N = new; M = medium; O = old; VO = very old

- please run function `setwd()` in a separate line (if you need it) when importing the data set into R.
- **the names given in bold below MUST be used in your R code**.
- In all the tests, use significance level $\alpha = 0.05$.

Load the file `house-prices-OR.csv` into R and name it as **house**. Our purpose is to fit a linear model that helps to predict house's price.

1. (2 points) Rename the first column of data frame **house** from **HP.in.thousands** to **price**.
2. (2 points) Form a QQ plot and a histogram for **price** and comment if it is suitable to be the response for a linear model. Explain.
3. (4 points) Plot a scatter plot between **price** and **House.Size**. Derive the correlation coefficient between them. Give your comments.
4. (2 points) Fit a multiple linear model for **price** that depends on all the input variables in the data given, name it as **M1**. Write down the fitted model.
5. (4 points) Test the significance of variable **Age.category** in model **M1** by reporting the hypotheses, the test statistic value, the null distribution and the p-value.
6. (2 points) Report the coefficient of variable **Garage** in model **M1** and interpret it.
7. (4 points) Let **SR1** denotes the standardized residuals of model **M1**. Give your comments on the adequacy of model **M1** based on the residual plots.
8. (2 points) Does model **M1** have any outliers? If yes, report the number of outliers.
9. (2 points) Does model **M1** have any influential points? If yes, report the index of influential points in the data set given.

10. (10 points) Given the three houses, A, B and C, listed below. Before using model **M1** to predict the selling price for each of them, we would need to check if any of them is extrapolation.

    A: House.Size = 2551, Acres = 0.4, Bedrooms = 3, Garage = 1, Age.Category = O;

    B: House.Size = 600, Acres = 0.05, Bedrooms = 1, Garage = 0, Age.Category = VO;

    C: House.Size = 3100, Acres = 0.3, Bedrooms = 3, Garage = 1, Age.Category = M;

    (a) Using only the quantitative input variables, form Hat matrix for model **M1** in R, denoted as **H**. Report the value of the largest diagonal h_{ii} of **H**, denote it as **h.max**.

    (b) With the three new points given above, form matrix **x**, then find and report the value h_{00} for each point. Any of the three houses given is extrapolation?

    Hint: You could refer to Slide 50 – 54 of Topic 2 - Description.

    (c) For house A, report the predicted selling price and a 99% confidence interval for its selling price.

11. (2 points) Remove the variable that is most insignificant in model **M1**, and re-fit a new model without it, called **M2**. Report the adjusted R^2 of **M2** to 4 decimal places.

12. (2 points) Compare the adequacy and the goodness of fit of model **M1** and model **M2**. Which model would you prefer among these two? Explain.

13. (2 points) One suggests that we should use square root of the house's price as the response instead of the original price. Explain for that suggestion.

2. (*10 points*) A data set is collected with n observations, $(x_1, y_1), ..., (x_n, y_n)$. A simple linear model $y = \beta_0 + \beta_1 x + \epsilon$ is fitted using the data set above by OLS method. The equation of the fitted model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators of β_0 and β_1 by OLS method, which yields a coefficient of determination R^2.

1. (3 points) Prove that $\text{Cov}(y, \hat{y}) = \text{Var}(\hat{y})$.

2. (3 points) Prove that $R^2 = [\text{Cor}(y, \hat{y})]^2$.

3. (4 points) Given $n = 25, R^2 = 0.9$. Test the significance of the model at $\alpha = 0.05$.

Recall some definitions:
For a sample $(x_1, ..., x_n)$, sample variance is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
For two samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, the sample covariance is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

For two samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, the correlation coefficient is

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}}.$$

END OF QUESTIONS