BANA 7042: Assignment 1

Linear Regression

Please clearly circle (or highlight) your multiple choice answers

Question 1: In a multiple linear regression model, which of the following are considered as random?

    a. The continuous response or dependent variable Y
    b. The expected value of Y
    c. The unknown parameter β
    d. The unknown error variance σ^2

Question 2: When developing a linear regression model, adding additional predictors to the model will

    a. Always increase R^2
    b. Always increase adjusted R^2
    c. Sometimes decrease R^2
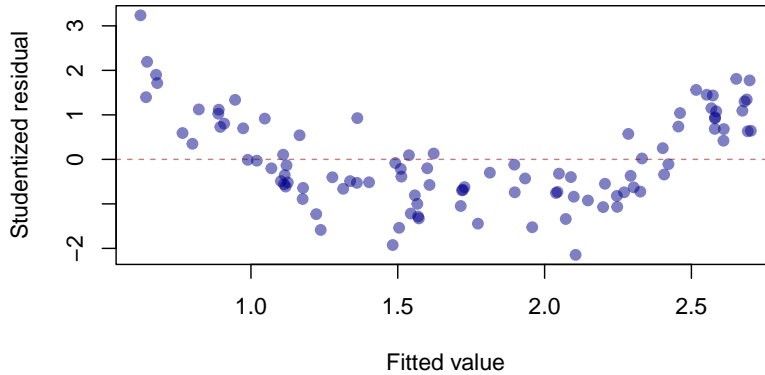   d. Always decrease adjusted R^2

Question 3: What is the best way to identify potential multicollinearity?

    a. Including interaction terms in the linear regression model
    b. Variance inflation factors (VIFs)
    c. Residual diagnostic plots
    d. Adjusted R^2

Question 4: When building a linear regression model of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ for a data set, the presence of multicollinearity in the data may result in _____ standard errors for the slope coefficients (β_1 and β_2) than if the data came from an orthogonal design (i.e., when X_1 and X_2 are uncorrelated).

    a. Smaller
    b. Larger

Question 5: What does the following residual versus fitted value plot suggest about a model between a single predictor X and Y?

a. Heteroscedasticity of the error term
b. A nonlinear between X and Y
c. Non-normality of the error term
d. Satisfactory residuals

Question 6: When a linear regression model is being developed, adding additional variables to the model will

a. always decrease model SSE
b. always decrease model AIC
c. always increase model adjusted R^2
d. always decrease model MSE

Question 7: The ordinary residuals refer to

a. $\bar{Y} - \hat{Y}$
b. $Y - \hat{Y}$
c. $\bar{Y} - \bar{Y}$
d. $\hat{Y} - Y$

Question 8: For a fitted simple linear regression model, which one of the following properties is **NOT** true?

a. The fitted regression line passes through the point $(\bar{X}, \bar{Y})$
b. The residuals sum to zero: $\sum_{i=1}^{n} e_i = 0$
c. $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

d. The residuals, e_i, are guaranteed to be independent

Question 9: The diagonal elements of hat matrix, also referred to as the *hat values* or *leverage values*, measures the influence of observation i on the regression line when removing observation i.
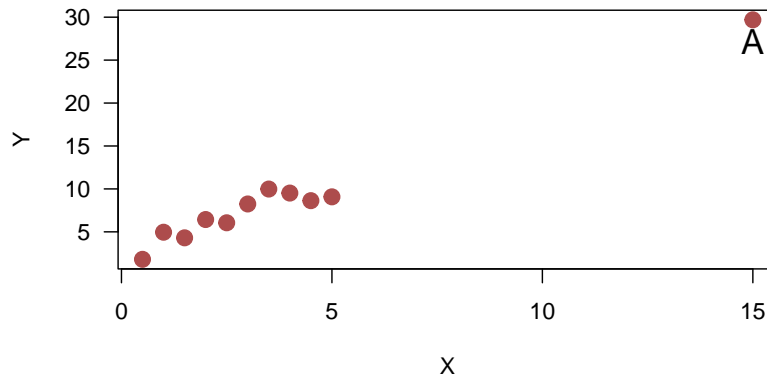
    a. True
    b. False

Question 10: In a regression study, a 95% confidence interval for β_1 was given as: $(-5, 2)$. What does this confidence interval mean?

    a. The interval $(-5, 2)$ contains the true β_1 with 95% probability
    b. 95% of all possible β_1's are in $(-5, 2)$
    c. 95% of the interval $(-5, 2)$ contains the true β_1
    d. If we were to repeat the experiment many times (i.e., repeatedly take a new sample of size n and compute the same confidence interval), roughly 95% of the generated intervals would contain the true β_1

Question 11: In a regression study, a 95% confidence interval for β_1 was given as: $(-5, 2)$. Which of following is correct?

    a. The 90% interval will be wider than the 95% interval
    b. The 99% interval will be wider than the 95% interval

Question 12: Point A in the far right is likely to be

    a. High influential point of the regression line
    b. High leverage point

Question 13: Which of the following cases can lead to multicollinearity?

    a. There are indicator variables being used as predictors
    b. A predictor can be expressed as (or well-approximated by) a linear combination of the other predictors
    c. The variances across all predictors are not the same
    d. The predictors are not normally distributed and are positively skewed

Question 14: What does the last line of output "F-statistic….p-value: …" indicate in the following multiple regression output?

```
Call:
lm(formula = mpg ~ cyl + disp + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0562 -1.4636 -0.4281  1.2854  5.8269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.82854    2.75747  14.807 1.76e-14 ***
cyl         -1.29332    0.65588  -1.972 0.058947 .
disp         0.01160    0.01173   0.989 0.331386
hp          -0.02054    0.01215  -1.691 0.102379
wt          -3.85390    1.01547  -3.795 0.000759 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.513 on 27 degrees of freedom
Multiple R-squared:  0.8486,    Adjusted R-squared:  0.8262
F-statistic: 37.84 on 4 and 27 DF,  p-value: 1.061e-10
```

    a. All of the coefficients are significantly different from 0 at the $\alpha = 0.05$ level
    b. None of the independent variables explains any of the variation in Y
    c. At least one of the independent variables explains some of the variation in Y
    d. The model explained 1.061e-10 of the variability in Y

Question 15: If one wishes to incorporate seasonal dummy variables for monthly data into a regression model, how many dummy variables should be in the model?

    a. 12
    b. 11
    c. 10

d. 1

Question 16: Consider a simple linear regression model where $\log(Y) = \beta_0 + \beta_1 X + \varepsilon$. If $\beta_1 = 0.05$, which best describes this coefficient's interpretation?

    a. A one-unit increase in X is associated with a 0.05 unit increase in Y

    b. A one-unit increase in X is associated with an approximate 5% increase in Y

    c. A one-unit increase in X is associated with a 5-unit increase in Y

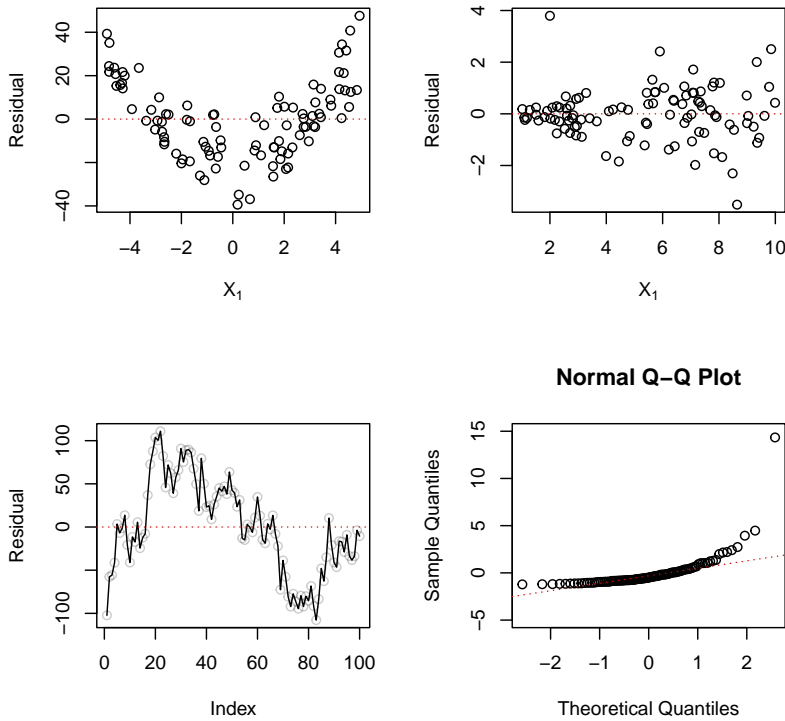   d. A one-unit increase in X is associated with a 0.05% increase in Y

Part II: short answer response

For the multiple linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{p-1,i} + \epsilon_i$

Question 17: (2 pts.) Recall that the variance inflation factor for X_j is defined as $VIF_j = \left(1 - R_j^2\right)^{-1}$. Explain in one sentence what R_j^2 is. Explain in one sentence why it is better to use VIFs than pairwise correlations when trying to detect the presence of multicollinearity in a regression data set.

Question 18: (2 pts.) For variable selection criteria based on residuals, among $PRESS$, SSE, R^2, R_{adj}^2, and MSE, which are "good" criteria for variable selection? Explain your reason in two sentences.

Question 19: (4 pts.) Please state one possible violation of normal linear regression model assumptions for each of the residual plots below (one sentence for each plot):

Question 20: (4 pts.) Suppose you want to build a linear regression for response variable weight (Y) using covariate height (X_1) and gender ($X_2 = 0$ for female and $X_2 = 1$ for male).

a. Suppose you want to allow the slope (of height) to be the same for different gender but different intercept, how would you build the linear regression model? Please specify the model in one line.

b. Suppose you want to allow both slope (of height) and intercept to be different for different gender, how would you build the linear regression model? Please specify the model in one line.

Question 21: (2 pts.) An engineer has stated: "Reduction of the number of candidate explanatory variables should always be done using the objective forward stepwise regression procedure." Discuss.

Question 22: (4 pts.) A junior investment analyst used a polynomial regression model of relatively high order in a research seminar on municipal bonds and obtained an R^2 of 0.991 in the regression of net interest yield of bond (Y) on industrial diversity index of municipality (

X) for seven bond issues. A colleague, unimpressed, said: "You overfitted. Your curve follows the random effects in the data." Comment on the colleague's criticism.

Question 23: (4 pts.) In a regression study of factors affecting learning time for a certain task (measured in minutes), gender of learner was included as a predictor variable (X_2) that was coded $X_2 = 1$ if male and $X_2 = 0$ if female. It was found that the estimated coefficient of X_2 was $\hat{\beta}_2 = 22.3$ with a standard error of 3.8. An observer questioned whether the coding scheme for gender is fair because it results in a positive coefficient, leading to longer learning times for males than females. Comment.

Question 24: (2 pts.) A student stated: "Adding predictor variables to a regression model can never reduce R^2, so we should include all available predictor variables in the model." Comment.

Question 25: (2 pts.) Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

Question 26: (2 pts.) What is a residual? Why are residuals important in regression analysis?

Question 27: (3 pts.) Explain the purpose of k-fold cross-validation in regression modeling. Why is it generally preferred over a single train-test split for assessing model performance? (2-3 sentences)

Question 28: Consider a dataset where the response variable Y is binary (0 or 1). Explain why ordinary linear regression is not appropriate for modeling this relationship, and what problem arises if you try to use it. (2-3 sentences)

Question 29: In the updated review slides, we discussed linear regression within the framework of Generalized Linear Models (GLMs). Identify and describe the three components of a GLM (Random Component, Systematic Component, and Link Function) specifically as they apply to the normal linear regression model.

Question 30 (Bonus): (5 pts.) An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{p-1,i} + \epsilon_i$ by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit using standard statistical software (e.g., R, Python, or SAS)? No need to run any code, just describe in general how you could accomplish this.