Total 52 point
(excluding 2pt bonus)

Dartmouth College Math 50 Fall 2017 Midterm Exam

Name: Solutions

Math 50 Linear Analysis

---

(1) Do not open this exam until you are told to do so.
(2) Before starting write your name, check each page and verify number of questions.
(3) Do not separate the pages of this exam. If they do become separated, write your name on every page and point this out when you hand in the exam.
(4) You can use the back of every page of the exam. However write your answers inside the boxes.
(5) Show your work for each question (except True/False questions). Answers without justification will not get points.
(6) No smart electronic devices.
(7) Turn off all cell phones, smartphones, and other electronic devices, and remove all headphones and smartwatches.
(8) If the given information is not sufficient write **NA.**
(9) For the questions which start with [*T* / *F*] circle true (T) or false (F).

Good Luck !!

2

(1) Fill in the blank with either "always greater or equal" , "always less or equal" or None :
Let (x_i, y_i) denote the observations as usual. In simple linear regression:

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 \ \underline{\text{always greater or equal}}\ \sum_{i=1}^{n}\left[y_i - \left(\hat\beta_0 + \hat\beta_1 x_i\right)\right]^2$$

$\rightarrow \hat\beta_0, \hat\beta_1$ are minimizers of this sum

2

(2) [T / F] If R^2 (coefficient of determination) is near 1 then most of the variation in y is explained by the regression model.

2

(3) [T / F] Multicollinearity occurs when two or more of the regressor variables are highly correlated.

2

(4) Let $\vec\beta = \begin{bmatrix} \vec\beta_1 \\ \vec\beta_2 \end{bmatrix}$. Is it true that $SS_R(\vec\beta) - SS_R(\vec\beta_1|\vec\beta_2) = SS_R(\vec\beta_2|\vec\beta_1)$. If so show, if not then do you know a condition where this statment is true?

$$SS_R(\vec\beta) - SS_R(\vec\beta_1|\vec\beta_2) \stackrel{?}{=} SS_R(\beta_2|\beta_1)$$

$$\cancel{SS_R(\vec\beta) - SS_R(\vec\beta)} + SS_R(\vec\beta_2) = SS_R(\vec\beta) - SS_R(\vec\beta_1)$$

does not hold in general.

Holds if X_1 and X_2 are orthogonal to each other

(5) You want to do a significance of regression test using 12 observations and the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. You choose significance level as $\alpha = 0.05$. The t-statistic you calculated is

$$t_0 = -3$$

For each of the following statements, circle true (T) or false (F).

2 [T / F] It can be concluded that, this test predicts linear relationship between x and y

2 [T / F] Test predicts that: x is of value in explaining the variability in y, and the relationship is not linear

2 [T / F] It can be concluded that, the probability of $\beta_1 < -3$ is less than or equal to 0.025

2 [T / F] The 95% confidence interval on β_1 does not contain 0.

(4)

(6) Choose the one that is most appropriate. All other things being equal,
(a) smaller joint confidence region is more desirable
(b) larger joint confidence region is more desirable
(c) none.

Explain : Smaller confidence region means the parameters can be estimated more accurately

(7) Consider multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

where $\varepsilon_i \sim NID(o, \sigma^2)$. From 20 observations above model is fitted. Following results obtained.

	Coeff. Estimate	$se\left(\hat{\beta}_i\right)$	t-Stat	Other
$\hat{\beta}_0$	10.1	4.0		
$\hat{\beta}_1$	-0.02			
$\hat{\beta}_2$	________	6	4.1	
$\hat{\beta}_3$	________	2.0		Prob($\beta_3 < 1.0$) is 0.01

$SS_T \approx 71.36$ $SS_R \approx 55.2$

Answer the following, if the given information is not sufficient write **NA**.

(2) (a) From the given information we can calculate that $\hat{\beta}_2 =$ 24.6
(2) (b) From the given information we can calculate that $\hat{\beta}_3 =$ 6.16 → See below
(2) (c) [T / F] The value of $\hat{\beta}_1$ is significantly less than other coefficient estimates therefore contribution of regressor x_1 is small.
(2) (d) What is Var(y) ? σ^2
(e) The least squares estimation gives a prediction for Var$(\varepsilon_1) =$ $\frac{71.36-55.2}{16} = 1.01$
(2) (f) Calculate the bounds of 95% confidence interval on β_0

(2) $10.1 - 4\times 2.12 \leq \beta_0 \leq 10.1 + 4\times 2.12$

$1.62 \leq \beta_0 \leq 18.56$

↳ $Var(\varepsilon_1)$ is σ^2
its prediction is $\hat{\sigma}^2$

(g) Test for significance of individual regression coefficient β_0 and β_2 (use $\alpha = 0.05$). Give an interpretation for these results:
For β_0: N/A (typo in the question won't be graded)

For β_2: $t_0 = 4.1 > 2.12 \Rightarrow$ Reject null hypothesis $\beta_2 = 0$
(2) Regressor x_2 is significant (has value in explaining y)

7(b): $\text{Prob}\left(\frac{\hat{\beta}_3-\beta_3}{se(\hat{\beta}_3)} > 2.58\right) = 0.01 \Rightarrow \text{Prob}\left(\beta_3 < \hat{\beta}_3 - 2.58\times 2\right) = 0.01$

$\Rightarrow 1.0 = \hat{\beta}_3 - 2.58\times 2 \Rightarrow \hat{\beta}_3 = 2.58\times 2 + 1 = 6.16$

(4)

(8) Suppose that the fitted regression line (using least squares) for a given data set is

$$\hat{y} = 1 + 3x$$

so that $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 3$. You are trying to understand the effect of multiplying x_i and y_i (observations) with some constant, and therefore you multiple each x_i and y_i values with D. Then the fitted regression line for this new data set is :

$\hat{y} =$ ____$D + 3x$____

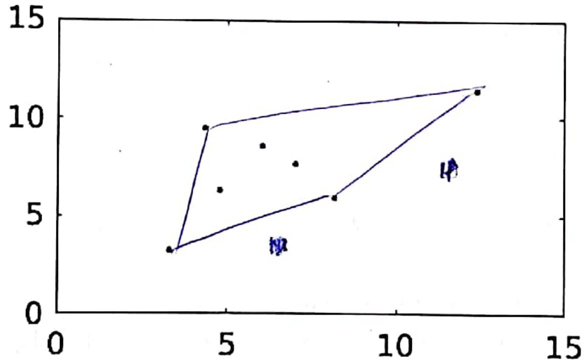
(if it is not possible to calculate write **NA**)

From the formulas of $\hat{\beta}_0$ and $\hat{\beta}_1$

(alternatively geometric approach is also valid i.e, streching in both directions)

(2)

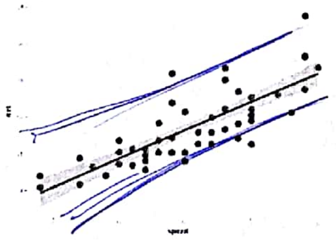
(9) For a multiple linear regression with regressors x_1 and x_2, the scatter plot of x_1 and x_2 for a given data is below (dots denote the observations). Given two points (squares in the plot) determine the extrapolation points (fill the square if that is an extrapolation point).



both are hidden extrapolation points

(2)

(10) Is it likely that below plot be 90% prediction interval of y? If so explain why, if not then draw a more realistic one.



roughly we expect %10 of the points to be outside of prediction interval

4

(11) Consider an unbiased and a biased estimators of an unknown (say σ^2). An unbiased estimator is not always better then the biased one because biased estimator might have a much smaller variance therefore provide more accurate estimations.

(12) An analyst tries to understand the relation between regressors x_1, x_2 and response variable y. She considers two models

$$\begin{aligned} \textit{Model } 1: \quad y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ \textit{Model } 2: \quad y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + \varepsilon. \end{aligned}$$

2

(a) Analyst suspects that at different values of x_2, the rate of change in y with respect to x_1 vary significantly. Which model do you propose to use? Explain.

Model-2, since interaction term provides varying slope $\left(\frac{\partial y}{\partial x_1}\right)$ at different values of X_2

2

(b) Write down a change of variable, and using it transform Model 2 into the below linear regression Model 3.

$$X_3 = X_1 X_2$$

$$\beta_3 = \beta_2$$

$$\textit{Model } 3: \quad y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$$

2

(c) Check that the Model-3 and Model-1 has similar form. Do you expect the resulting fitted regression surface to be different? Why?

They use same model and solutions but with different data. Observations must be converted/transformed for model 3.

2

(d) [Bonus] Analyst then decides to study observations carefully at distict x_2 values. She chooses several x_2 values and looks at scatter plot between y and x_1 at each of these x_2 values. Her visual investigations suggests that, there is a strong linear relationship between y and x_1 such that the slope and the intercept are both changing depending on what x_2 value she chooses. Can you propose a new model that might be better than the above two? Explain.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2 + \varepsilon$$

provides varying slope at different values of X_2

can model varying intercept.