 Name: 50 points

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 clearly if the question asks it otherwise 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).
(10) Unneccesary information might reduce the points you get from the question.

Good Luck !!

Each horizontal line denotes 1 point.

NOTE

- Unless otherwise specified $\hat{y}$ denotes the fitted values for a simple linear regression using least squares estimation. As usual x_i denotes the x coordinate, and y_i denotes y value of the i^{th} observation.

(1) Choose the one that is most appropriate. All other things being equal,
  (a) a biased estimator of β_1 with a smaller confidence interval is more desirable
  (b) an unbiased estimator of β_1 with a larger confidence interval is more desirable
  (c) none.

Explain : Each can be desirable depending on:
- the differences in biases
- difference in confidence interval lengths.

(2) 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 10 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			Condition index $\kappa_1 = 1$
$\hat{\beta}_2$	24.6	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**.
(a) The intercept estimation is 10.1.
(b) [T / F] The contribution of regressor x_1 cannot be concluded from the above data.
(c) The least squares estimation gives a prediction for Var(β_3) = $(2.0)^2$

Typo: supposed to be $Var(\hat{\beta}_3)$

(also full credit if you said $Var(\beta_3) = 0$)

(d) Calculate the bounds of 95% confidence interval on β_2

$\underbrace{24.6 - 6 \times 2.45}_{} \leq \beta_2 \leq 24.6 + 6 \times 2.45$

$\underbrace{6}_{t_{0.25,6}}$

$9.9 \leq \beta_2 \leq 39.3$

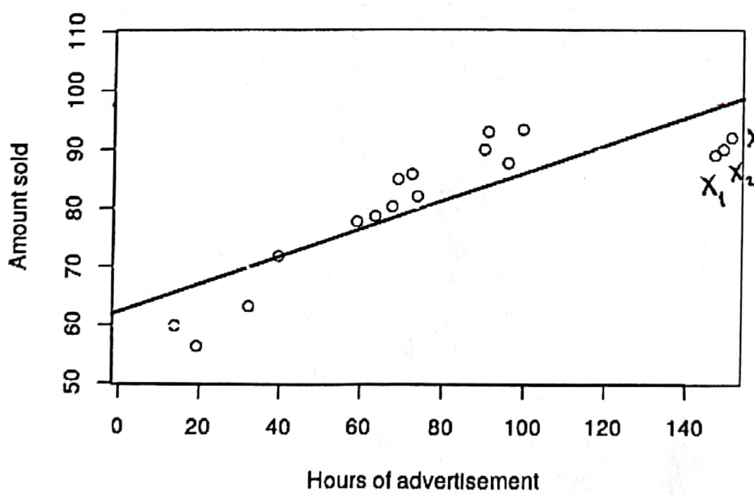
(e) Check the following null-hypothesis with $\alpha = 0.05$

$$\beta_2 = 0$$

0 is outside of CI (from part d)

Thus: reject the nullhypothesis.

(3) The following scatter diagram and fitting shows the relationship between y and x where x denotes number of hours of advertisement of a product on various media and y denotes amount of sales. The line in the plot denotes fitted simple linear regression model.



Which of the following we can deduce from the plot.

(a) From the plot it seems like

$\hat{\beta}_0$ is : 62 (61, 63, 64 also ok)

(b) The expression $\frac{SS_R}{SS_T}$ can be interpreted as the proportion of total variation explained by x. Which of the following is/are likely to be this proportion. (Circle the ones that apply)

-1.25

1.25

0.11

0.70

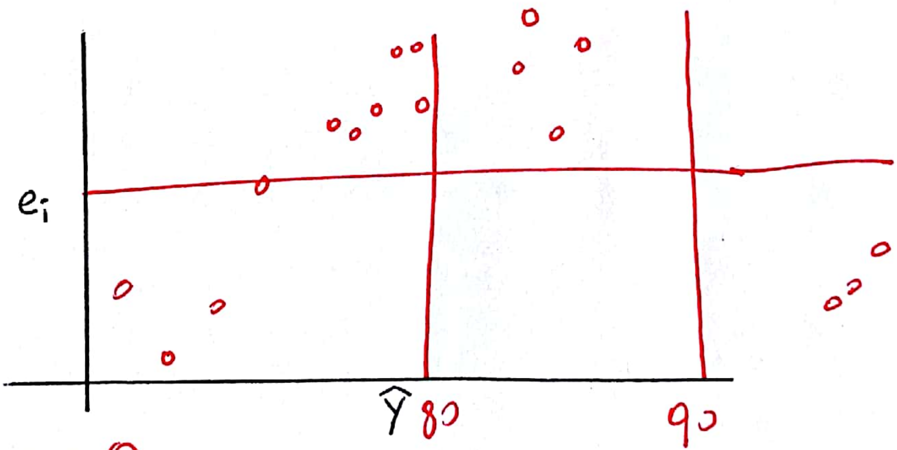
(c) Looking at the plot my estimation for the mean response at $x = 100$ is

$E(y|x = 100) =$ ≈ 85 .

(d) Make a prediction of new observation y at $x = 150$.

$y =$ ≈ 97 .

(e) In the below area give residuals vs fitted values $\hat{y}$ plot (horizontal axis is $\hat{y}$). Then draw two lines onto your plot $\hat{y} = 80$ and $\hat{y} = 90$. Hint use the grids to estimate residuals and note that there are 17 observations. Residuals are defined as $e_i = y_i - \hat{y}_i$.



(f) [T / F] The points x_1, x_2, x_3 are causing multicollinearity.

(g) [T / F] Looking at the scatter diagram, for the smallest few $\hat{y}$ values the residuals are negative, for largest $\hat{y}$ values residuals are also negative, in the middle part residuals are positive therefore residual plot drawn in part (e) will suggest nonlinearity.

(h) [T / F] Removing points x_1, x_2, x_3 together will likely increase R^2.

(i) [T / F] Each of the points x_1, x_2 and x_3 is a leverage point.

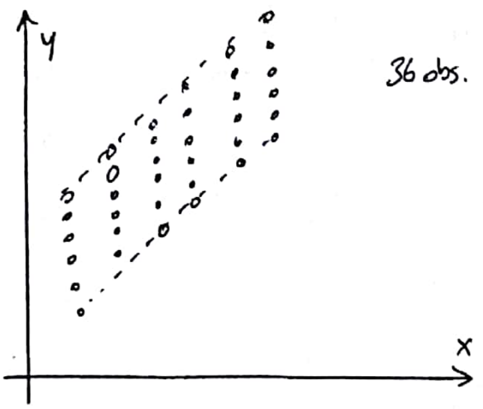
(j) [T / F] The observation at x_1 is a leverage point but not neccesarily an influential point

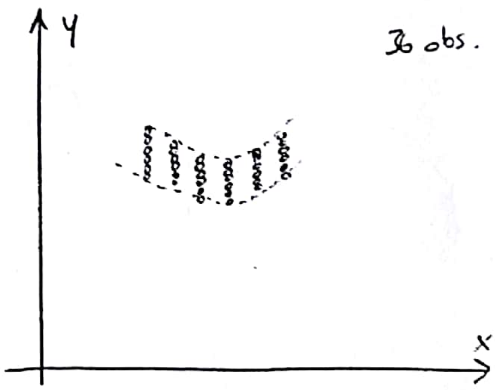
(k) [T / F] The points x_1, x_2 and x_3 are jointly influential.

(4) Suppose that the constant variance assumption holds and σ^2 is known. Can the variance of $y_i - \hat{y}_i$ be different than the variance of $y_j - \hat{y}_j$ for two different observation points i and j?

[Yes / No] Explain: These are residuals e_i and e_j. Their variance depends on location in x-space (h_{ii} and h_{jj}). they can be different.
(This is one reason for studentization of residuals)
See question 8.

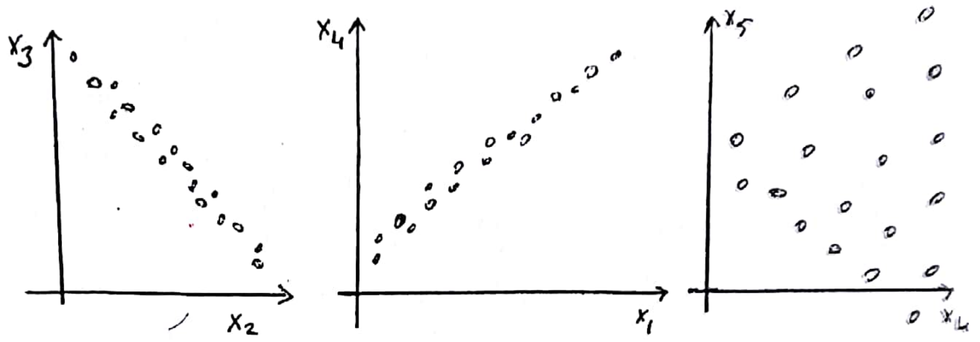
(5) Consider the following two sets of observations where two simple linear models fit to each data set and assume that both models gave the same SS_{Res}. Which of the below do you expect higher test statistic F_0 for lack of fit test.





[*Left* / *Right* / *None*] Explain: Right seems to have lower SS_{PE} (and therefore higher SS_{LOF} as SS_{Res} is constant)
This implies higher test statistic: $F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}$

(6) Below are regressor vs regressor plots



Suppose that you did a regression analysis with model $y = \beta_0 + \beta_1 x_3 + \beta_2 x_4 + \varepsilon$ and you are looking forward to add a new regressor to improve your model. Answer the following :

(a) [T / F] There is a strong linear relationship between x_1 and x_4 with a positive slope. Adding x_1 into the model might increase multicollinearity, on the other hand, the negative slope relationship between x_2 and x_3 implies that adding x_2 might decrease multicollinearity issues.

(b) [T / F] Adding x_5 to the model will worsen constant variance violations.

(c) If I want to choose only one regressor and add it to the model in order to improve it,
I would add: ____ x_5 ____. Because ____ both x_1 and x_2 ____
____ seems ~~[illegible]~~ to have a linear relationship ____
____ with x_3 and x_4 which are already in the model.

(7) Answer the following questions about h_{ii} and hat matrix H:
(a) [T / F] h_{ii} takes values between 0 and 1
(b) Draw a circle around the item number/numbers if it is correct. Given only H matrix, by studying various properties of it (such as diagonal

values, eigenvalues, singular values etc) we can determine whether an observation is

(i) an influential point

(ii) an outlier

(iii) a leverage point

(8) [T / F] One of the advantages of studentized residuals is that their variance does not depend on the x value of the observation point as opposed to the standardized residuals d_i.

(9) Suppose that observation-1 is (x_1, y_1) and it is an outlier but not a leverage point. Observation-2 is (x_2, y_2) is a pure leverage point.

(a) If we delete one of these observation points and repeat simple regression fitting, which deletion will more likely cause a bigger change in MS_{Res}

[*Observation* $-$ 1 / *Observation* $-$ 2 / *Not comparable*]

(b) If we delete one of these observation points and repeat simple regression fitting, which deletion will more likely cause a bigger change in R^2

[*Observation* $-$ 1 / *Observation* $-$ 2 / *Not comparable*]

(c) Suppose now there was an error in one of the measurements y_1 or y_2, and it needs to be changed to its correct value. After correction a new fitting is done. Which one will more likely cause a bigger change in the estimated coefficients $\hat{\beta}$

[*Error in* y_1 / *Error in* y_2 / *Not comparable*]

(10) Explain in one or two sentences. If you want to test contribution of regressors x_2 and x_3 in the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

then you can do ....

A joint F test to compare full model with deleted model

We can also fit both models and compare their statistics and visually compare their various diagnostic plots

(11) Suppose that for an application the following is expected to be a good model

$$y = 1 + (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)^k$$

How would you approach to this problem and solve using linear regression?

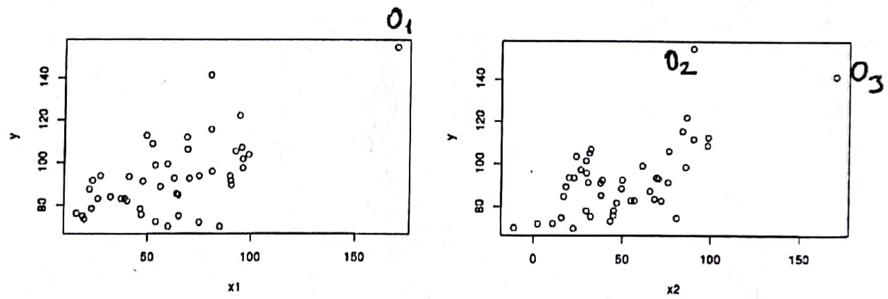
$$(Y-1)^{\frac{1}{k}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}$$

$$Y_{new} \leftarrow (Y-1)^{1/k} \quad \text{transformation}$$

(12) You did a multiple linear regression fitting on a given data, and you are asked "What is your estimation of σ^2 and how good is it?". How would you answer and which quantities would you present, explain :

$\hat{\sigma}^2$ and confidence interval on $\hat{\sigma}^2$ (or variance of $\hat{\sigma}^2$) as an estimator)

(13) Consider the below graphs of the data for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



From the above two plots which of the following we can deduce :

(a) [T / F] o_2 is an outlier → not necessarily, note this is X_2 vs Y (see final exam Rmd file)

(b) [T / F] o_1 and o_3 are leverage points

(c) [T / F] From the first plot we see that constant variance assumption is violated (cannot be deduced, see below)

(d) [T / F] The point $(x_1, x_2) = (100, 150)$ in x-space is a hidden extrapolation point. (cannot deduce from given graph) (see final exam Rmd)

for part c: Recall from HW that we've seen such behaviour can be fixed adding another regressor. (This plot is X_1 vs Y and model has another regressor)

(14) Suppose that matrix X consists of columns X_1, X_2, X_3 and X_4 each of which denote the n levels of corresponding regressors variables. Below are the eigenvalues and eigenvectors of matrix $(X'X)$,

$$\lambda_1 = 0.000001, \quad \lambda_2 = 0.2, \quad \lambda_3 = 8, \quad \lambda_4 = 9$$

$$\vec{v}_1 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix}, \quad \vec{v}_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

(a) Is there a multicollinearity problem?
[*Yes* / *No* / *Maybe*]

(b) If you choose yes, then which regressors have near-linear dependence?
Regressors: ________ X_2 and X_3 ____________

$$X\vec{v}_1 = \lambda_1 \vec{v}_1 \approx 0 \Rightarrow 0 \cdot X_1 + 2 \cdot X_2 + 1 \cdot X_3 + 0 \cdot X_4 \approx 0$$

(15) In general do we expect that "higher VIF_j implies larger confidence interval of β_j" and vice versa?
[*Yes* / *No*] Explain __ In general yes, these two are related as both are proportional to the var$(\hat{\beta}_j)$ (everything else assumed to be same)

(16) Fill in the blanks with one of the following

LSE, MLE, Loess, Weighted LSE, Kernel Regression, Cubic splines, Ridge Regression, Polynomial regression, smaller variance, larger variance, smaller expectation, larger expectation, smaller VIF, larger VIF, smaller bias, larger bias, variance, expectation, variance, VIF, bias.

The method of LSE always gives an unbiased estimator of $\vec{\beta}$. Instead, one can give up from this constraint and develop a method which gives a biased estimator with a ____ smaller variance ___. One such method is Ridge regression which tries to provide a biased estimator with a smaller value of bias squared plus Variance.

(17) Consider multiple linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

along with the observation data which consists of X_1, X_2 and Y values. Answer the following.

(a) [T / F] It is possible to determine leverage points just looking at the X_1 and X_2 values

(b) [T / F] In order to calculate j^{th} Cook's Distance D_j we only need X_1 and X_2 values and observation y_j

(c) [T / F] DFFITS determines influential points using only Y values of data table

(18) Considering piecewise fitting, what is the effect of increasing number of knot points on SS_{Res} and R^2? Will the fitting improve in general and is it desirable to increase knot points always?

Explain.

Might improve fitting, decrease SS_{Res}, increase R^2.

However ① might complicate the model,
② ~~...~~ can be oscillatory fitting.
③ explanation value of the model can get worse
④ new predictions can be more unstable.

thus might not be desirable.