1. For the linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon$

(1) (3 pts) Show that $(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$

(2) (3 pts) Show that $(\hat{y}_i - \bar{y}_i) = \hat{\beta}_1(x_i - \bar{x})$

(3) (6 pts) Show that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

2. In this question we shall make the following assumptions:

(i) Y is related to x by the simple linear regression model $Y_i = \beta x_i + \epsilon$, (i=1,2,...,n)

(ii) The errors $e_1, e_2, \ldots, e_n$ are independent of each other

(iii) The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2

(iv) The errors are normally distributed with a mean of 0 and variance σ^2 (especially when the sample size is small), i.e., $e \mid X \sim N(0,\sigma^2)$

(1) (4 pts) Find the least squares estimate of $\hat{\beta}$

(2) (4 pts) Show that $Var(\hat{\beta}|X) = \dfrac{\sigma^2}{\sum_1^n x_i^2}$

(3) (4 pts) Show that $\hat{\beta}|X \sim N(\beta, \dfrac{\sigma^2}{\sum_1^n x_i^2})$

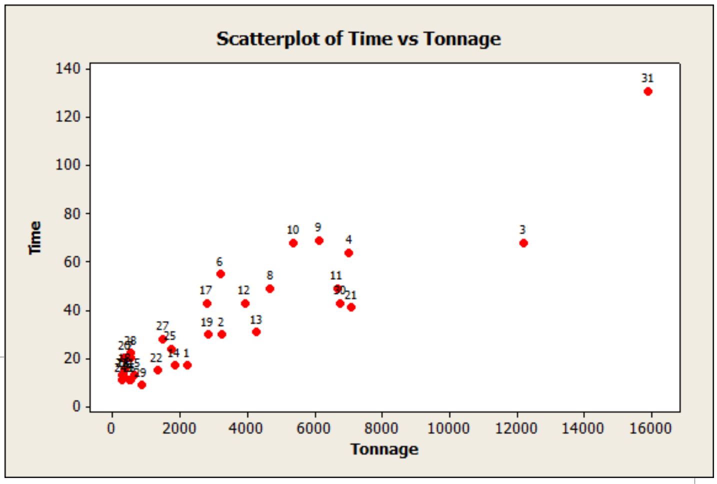
3. (<u>Canadian port</u>) The Canadian port on the Great Lakes wish to estimate the relationship between the volume of a ship's cargo and the time required to load and unload this cargo. It is envisaged that this relationship will be used for planning purposes as well as for making comparisons with the productivity of other ports. Records of the tonnage loaded and unloaded as well as the time spent in port by 31 liquid-carrying vessels that used the port over the most recent summer are available.

Tonnage	2213	3256	12203	7021	529	3192	547	4682	6112	5375	6666
Time	17	30	68	64	11	55	20	49	69	68	49
Tonnage	3930	4263	…	…	329	2790	353	2829	363	7084	
Time	43	31	…	…	13	43	15	30	20	41	
Tonnage	1328	294	268	1732	507	1486	536	851	6760	15900	
Time	15	13	11	24	11	28	22	9	43	131	

where:

$\sum Time_i = 1073$, $\sum Time_i^2 = 57705$, $\sum Tonnage_i = 105911$, $\sum Tonnage_i^2 = 767795783$,

And $\sum Tonnage_i * Time_i = 6311833$



1. Using the scatterplot of **Time vs Tonnage**:

(ii) (4 pts) Interpret the relationship between the two variables? what is the value of the sample correlation coefficient?

(2 pts) Scatter plot Interpretation:

(2 pts) Sample correlation Coefficient:

(i) (2 pts) Is there any bad/good leverage points? Explain.

Answer:

3. (6 pts) Fit the model: $Time = \beta_0 + \beta_1 Tonnage + \epsilon$ Show formulas and your calculations

4. (2 pts) What is the residual when tonnage=268.

5. (3 pts) What is the amount of explained variation in the time that can be explained by the tonnage?

6. (8 pts) Find the 95% C.I. for the slope and Interpret its meaning in this problem.

7. (8 pts) Test the significance of the model? Use $\alpha = 0.01$.

Points distribution
(2) H0: vs. H1:
(2) Test statistic:
(1) Decision rule:
(1) critical value(s):
(2) Decision and Conclusion:

8. (7 pts) State the model assumptions? Does the straight line regression model seem to fit the data well? If not, list any weaknesses apparent in model.

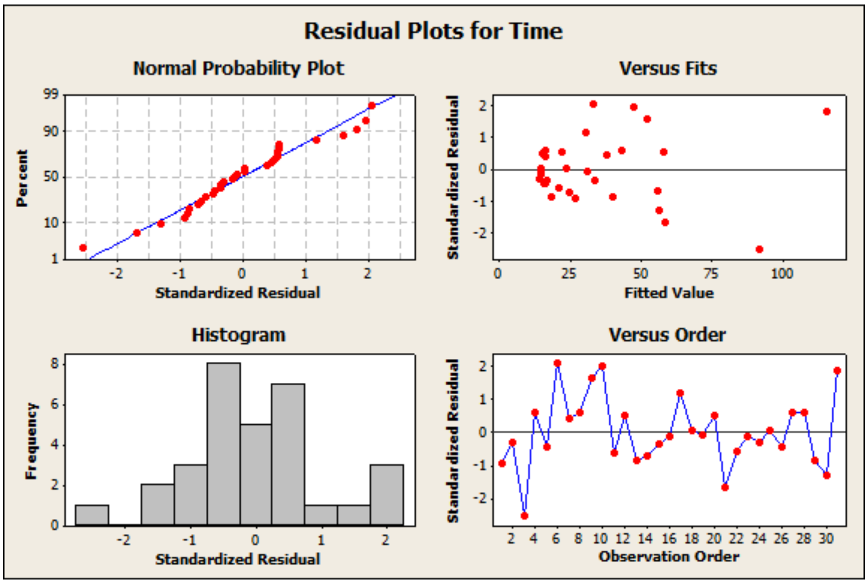
You must explain which graph(s) will be used to examine each assumption.

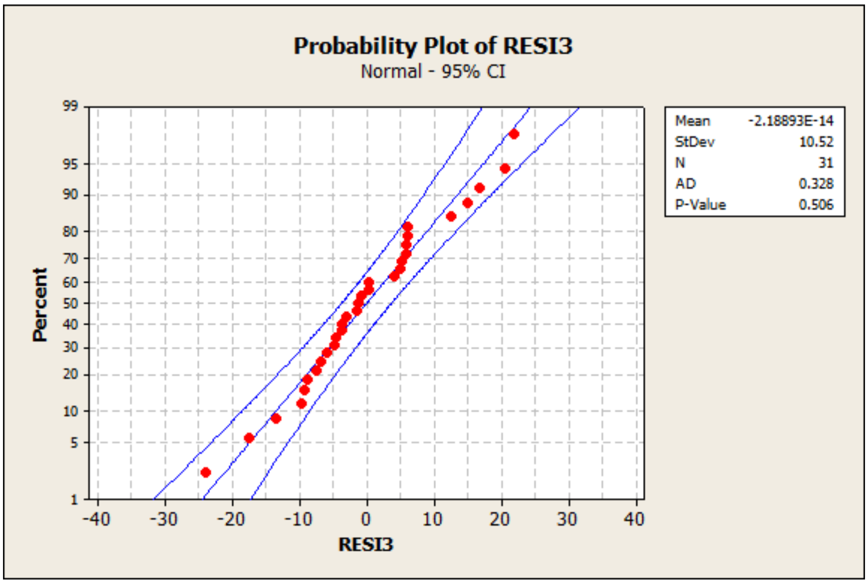
9. (3 pts) Explain how to use the scatterplot of HI vs Tonnage in this problem?

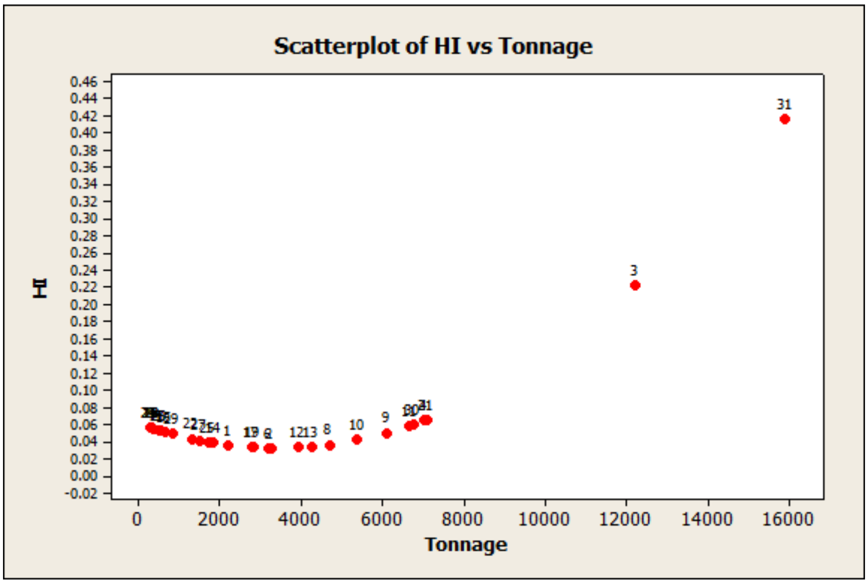
10. (7 pts) Suppose the model was considered:

(I) (5 pts) Use the model to compute a 95% prediction interval for Time when Tonnage = 10,000.

(II) (2 pts) Would the interval be valid? Give a reason to support your answer.


Residual Plots for Time
Normal Probability Plot
Percent
Standardized Residual
Versus Fits
Standardized Residual
Fitted Value
Histogram
Frequency
Standardized Residual
Versus Order
Standardized Residual
Observation Order


Probability Plot of RESI3
Normal - 95% CI
Percent
RESI3
Mean -2.18893E-14
StDev 10.52
N 31
AD 0.328
P-Value 0.506


Scatterplot of HI vs Tonnage
HI
Tonnage
31
3

4. *Regarding the problem* (Canadian port) *in Q3,*

Suppose another model fitted to the data was :

$$Y = X\beta + \epsilon \quad ,where\ Y = LN(Time)\ and\ X = Tonnage^{0.25}$$

Given*:* Partial minitab outputs *(last page) and*

$$X'X = \begin{pmatrix} 31 & 212.398515 \\ 212.398515 & 1578.97687 \end{pmatrix} \quad and \quad X'Y = \begin{pmatrix} 102.49278 \\ 740.47587 \end{pmatrix}$$

1. (6 pts) *Find* $\hat{\beta}$ *?*

2. (8 pts) Estimate $Var(\hat{\beta}_0)$, $Var(\hat{\beta}_1)$ *and* $Cov(\hat{\beta}_0, \hat{\beta}_1)$.

3. (4 pts) compute the coefficient of determination and interpret its value.

4. (4 pts) Is this model an improvement over model in part one in terms of predicting Time? If so, please describe all the ways in which it is an improvement.

5. (4 pts) ***<u>Someone suggestion</u>***: *"to improve the model in Q3, remove the observations 3 and 31 from the data then fit the linear model and do the transformation if needed".*

Do you agree with this suggestion? Explain.

Part Two: Partial computer outputs

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	11.820	— —	— — —	0.000
Residual Error	29	2.670	— —		
Total	30	14.489