Problem -1:

Solution

Source of variation	Sum of Square	Degree of freedom	Mean-Square	F	P-value
Regression	5550.8166	**2**	**2775.4083**	**261.24**	**0.00001**
Residual	**233.726**	**22**	**10.62**		
Total	5784.5426	**24**			

Total=Regression+Residual
5784.5426=5550.8166+Residual
Residual= 233.726

n=25,k=2
$DF_{Regression}$=k=2
$DF_{Residual}$=N-k-1=25-3=22
DF_{Total}=n-1=25-1=24

$MST=SST/DF_{Regression}$
MST(Regression)= 5550.8166/2
MST(Regression)= 2775.4083

$MSE(Residual)=SSE/DF_{Residual}=233.726/24$
MSE(Residual)=10.62

F=MST(Regression)/MSE(Residual)
F= 2775.4083/10.62
F=261.24

p-value=0.00001

Since p-value is less than 0.05 at significance level, we can conclude that the regression model is significant.

Problem-2 :

a. Is there any problem with this model based on the parameter estimates table below? If so, explain the problem and suggest one way to deal with this problem.

Solution:

The parameter estimates(coefficient) are the log odds ratio associated with a one-unit change of the predictor, all other predictors being held constant. A coefficient of model describes the size of the contribution of that predictor whereas a large coefficient indicates that the variable

strongly influences the probability of that outcome, while a near-zero coefficient indicates that variable has little influence on the probability of that outcome. A positive sign indicates that the explanatory variable increases the probability of the outcome, while a negative sign indicates that the variable decreases the probability of that outcome. In the given examples, there are three variables that are positive and three variables are negative that will indicate the variance influence on the response variable.

For example, the coefficient of the rbi is 34.6 , which indicates predictor changes per one-unit. This coefficient is the rate of change in the response per 1 unit change in the log of the predictor.There is a problem in the above model based on parameter estimates We can say that the variables rbi, runs and homers have large positive coefficients that indicates the variable strongly influences the response variable. In such cases we remove such parameters completely from the model.

If so, explain the problem and suggest one way to deal with this problem.

-Removing highly correlated independent variables

- Linearly combining independent variables such as adding them together

- Applying ridge regression.

A standardized parameter estimate removes the unit of measurement of predictor and response variables. They represent the change in standard deviations of the response for 1 standard deviation change of the predictor.

VIF, the variance inflation factor, represents the increase in the variance of the parameter estimate due to correlation (collinearity) between predictors. Collinearity between the predictors can lead to unstable parameter estimates. As a rule of thumb, VIF should be close to the minimum value of 1, indicating no collinearity. When VIF is greater than 5, there is high collinearity between predictors.

b. Following part (a), there are actually 7 more variables available in addition to the 6 variables in part(a) to predict the salary. Use the stepwise selection to choose the best model from ALL the 13 potential predictor variables. What is the best model according to the Mallows CP criterion? Give you reason

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	rbi		rbi	1	0.4345	0.4345	21.5315	28.43	<.0001
2	contract		contract	2	0.1684	0.6029	6.6973	15.27	0.0004
3	ko		ko	3	0.0487	0.6516	3.8308	4.89	0.0336
4	err		err	4	0.0271	0.6787	3.1181	2.87	0.0993

The best model according to Mallow's criterion are

1. Identify subsets of predictors for which the C_p value is near $k+1$
2. The full model always yields $C_p = k+1$, so don't select the full model based on C_p.

3. If all models, except the full model, yield a large C_p not near $k+1$, it suggests some important predictor(s) are missing from the analysis.
4. If a number of models have C_p near $k+1$, choose the model with the smallest C_p value, thereby ensuring that the combination of the bias and the variance is at a minimum.
5. When more than one model has a small value of C_p value near k+*1*, in general, choose the simpler model

For the given example,
we need to compare C_p to the number of parameters ($k+1$)

1. The model containing one predictor contains 2 parameters and its cp value is 21.5315. Since C_p value is greater than k+1, it suggests the model is biased.
2. The model containing two predictors contains 3 parameters and its cp value is 6.69. Since cp value is greater than k+1, it suggests the model is biased.
3. The model containing Three predictors contains 4 parameters and its cp value is 3.83. Since C_p value is smaller than k+1, it suggests the model is unbiased.
4. The model containing the four predictors contains 5 parameters and its cp value is 3.1181. Since C_p value is smaller than k+1, it suggests the model is full model and unbiased.

So In this case, according to Mallows C_p criterion, we will choose the model which contains the maximum difference of [k+1 - C_p] - the best model is the one containing 4 predictors and has a C_p value of 3.1181.

Problem-3:

Continued….

3) Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

where x_i is dummy variable (0 = failure and 1 = success). suppose that the dataset contains n_1 failure & n_2 successes

1) obtain the X^TX matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

∴ the dataset contains n_2 successes,

$\sum x_i = n_2$, $\sum x_i^2 = \sum x_i = n_2$ (as $x_i = 0$ or 1)

∴ Therefore,

$$X^T X = \begin{bmatrix} n & n_2 \\ n_2 & n_2 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & n_2 \\ n_2 & n_2 \end{bmatrix}$$

2) Obtain the $X^T Y$ matrix

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

3) Obtain the least square estimate b

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$= \begin{bmatrix} n & n_2 \\ n_2 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \frac{\begin{bmatrix} n_2 & -n_2 \\ -n_2 & n \end{bmatrix}}{n n_2 - n_2^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

By solving the above matrix equation, we get :-

$$\hat{\beta}_0 = \frac{1}{n_1} \sum y_i - \frac{1}{n_1} \sum x_i y_i$$

$$\hat{\beta}_1 = -\frac{1}{n_1} \sum y_i + \frac{n}{n_1 n_2} \sum x_i y_i$$

Problem-4:

4) The hat matrix plays an important role in diagnostics for regression analysis. Write down the Hat Matrix and show that Hat matrix has the following properties.

a) The hat matrix is symmetric.

for square and invertable matrix A, the inverse and transpose,

$$(A^T)^{-1} = (A^{-1})^T$$

Then we have,

$$[(X^TX)^{-1}]^T = [(X^TX)^T]^{-1} = (X^TX)^{-1}$$

So, X^TX is symmetric matrix. Then,

$$H^T = [X(X^TX)^{-1}X^T]^T$$

$$H^T = X[(X^TX)^{-1}]^TX^T$$

$$H^T = X[(X^TX)^{-1}]^TX^T = X(X^TX)^{-1}X^T$$

$$\therefore H^T = H$$

b) The Hat matrix is idempotent.

for Matrix H-bar, we also have,

$$\bar{H}\,\bar{H} = \bar{H}$$

By definition,

$$\bar{H} = I - H$$

$$\bar{H}\bar{H} = (I-H)(I-H)$$

$$= I - 2H + HH \qquad \because HH = H$$

$$= I - 2H + H$$

$$= I - H$$

$$= \bar{H}$$

$\therefore$ proved

4. c) Show that $0 < h_{ii} < 1$ for all i, and the sum of the h_{ii} values equals p, where $p-1$ is the number of predicting variables in the regression

→ linear regression model,

$$Y_n = X_{n \times p} \beta_p + \varepsilon_{n-1}$$

$$\hat{\varepsilon} = Y - X\hat{\beta} \quad \text{but} \quad X\hat{\beta} = X(X'X)^{-1}X' \cdot Y$$

& $H = X(X'X)^{-1}X^T$

$$\hat{\varepsilon} = (I - H)Y$$

$$\text{Var}(\hat{\varepsilon}) = \text{Var}\{(1-H)Y\} = \hat{\sigma}^2 (I-H)$$

$$\text{Var}(\hat{\varepsilon}) = \sigma^2 (I-H)$$

for i^{th}, we can write as,

$\sigma_i^2 (1-h_{ii})$ & now for SE,

$\sigma_i^2 (1-h_{ii})$, $\sigma\sqrt{1-h_{ii}}$

Since, we are multiplying σ with $\sqrt{1-h_{ii}}$, we can say that $SE(\hat{\varepsilon}_i) > 0$

but this is possible if $0 < h_{ii} < 1$.
Thus, We can say that $0 < h_{ii} < 1$

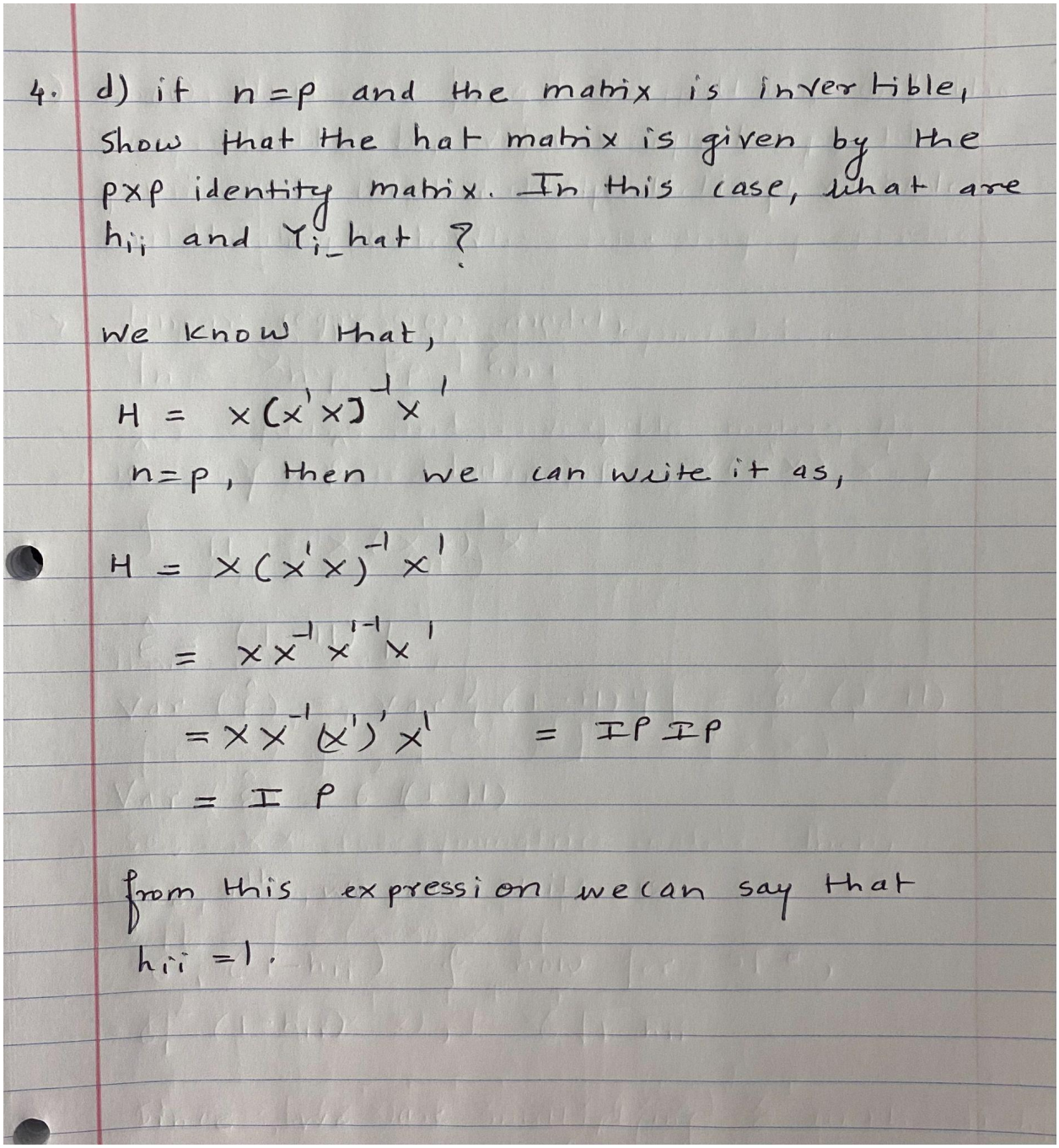
In Order to prove that sum of all i^{th} (h_{ii}) element equals p, We need to calculate the trace of $\hat{H}$.

$\therefore$ Trace $(x\,(x'x)^{-1}x)'$

$=$ Trace $(x'x\,(x'x^{-1}))$

$=$ Trace $(I\ P)$

$\Sigma h_{ii} = P$



4. d) if $n = p$ and the matrix is invertible, show that the hat matrix is given by the $p \times p$ identity matrix. In this case, what are h_{ii} and Y_i_hat ?

We know that,

$$H = X(X'X)^{-1}X'$$

$n = p$, then we can write it as,

$$H = X(X'X)^{-1}X'$$

$$= XX^{-1}X'^{-1}X'$$

$$= XX^{-1}(X')'X' \quad = IpIp$$

$$= I\,p$$

from this expression we can say that $h_{ii} = 1$.

Problem-5:

1. Briefly interpret β0 and construct a 90% confidence interval for β0.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	2273.7	1136.87	61.67	0.000
X1	1	240.1	240.10	13.03	0.004
X2	1	980.1	980.10	53.17	0.000
Error	12	221.2	18.43		
Total	14	2494.9			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	40.40	1.92	21.04	0.000	
X1					
1	9.80	2.72	3.61	0.004	1.33
X2					
1	-19.80	2.72	-7.29	0.000	1.33

From above minitab output,
$\beta 0 = 40.40$
SE coef=1.92
DF_{error} =12
$t_{0.1,12}$=1.356
From This information, we can state that the mean clarity of eco lake is 40.40
Now calculating 90% interval for $\beta 0$
= $\beta 0 \pm t_{0.1,12} * SE(\beta 0)$
=40.40 ± 1.356 * 1.92
=40.40 + 2.603 , 40.40 - 2.603
=(43.00,37.79)

2.Perform a test to determine whether the expected lake clarity is the same for all three ecoregions.
From the above output,
F-value of the regression is 61.67
F-value for the regression is used to determine the clarity for all three ecoregions.
The p-value is equal to zero, This shows that the p-value is less than the significant level. Hence we can conclude that the expected lake clarity is not the same for all three regions

3. Perform a test to determine whether the lake clarity is the same in ecoregion A and B

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	2273.7	1136.87	61.67	0.000
X1	1	240.1	240.10	13.03	0.004
X2	1	980.1	980.10	53.17	0.000
Error	12	221.2	18.43		
Total	14	2494.9			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	40.40	1.92	21.04	0.000	
X1					
1	9.80	2.72	3.61	0.004	1.33
X2					
1	-19.80	2.72	-7.29	0.000	1.33

From the above minitab output of ANOVA table,
P-value of A=0.004 and P-value of B=0.
P-value is less than the significance level i.e 0.05 hence we can state that the lake clarity is not the same in ecoregions A and B.

Problem-6:

1) Find the maximum likelihood estimates of β0 and β1. State the fitted response function.

Method

Link function	Logit
Rows used	30

Response Information

Variable	Value	Count
Y	1	16 (Event)
	0	14
	Total	30

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -4.81 + 0.1251 X

Coefficients

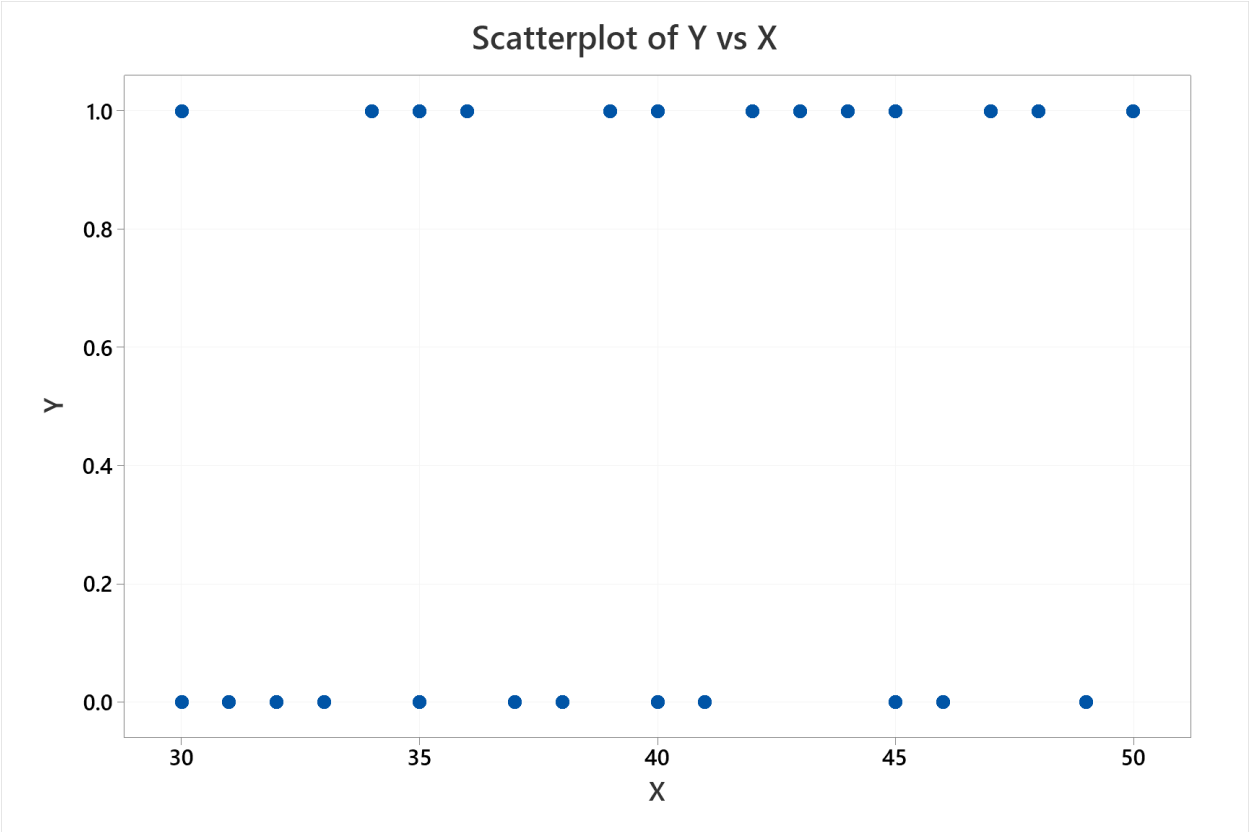
Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-4.81	2.66	(-10.01, 0.40)	-1.81	0.070	
X	0.1251	0.0668	(-0.0058, 0.2559)	1.87	0.061	1.00

The fitted response function is ,

From the above Minitab output, we see that the maximum likelihood estimates of $\beta_0 = -4.81$ and $\beta_1 = 0.125$

The fitted response function is : $\hat{\pi} = [1 + \exp(4.80751 - 0.12508X)]^{-1}$

2)Obtain a scatter plot of the data.



3) Obtain exp(b1) and interpret this number.

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
X	1.1332	(0.9943, 1.2917)

$e^{\beta 1} = e^{0.125} = 1.1332$

This means for every one dollar increase in the annual dues, the odds ratio of getting membership unrenewed (Y=1) versus renewed (Y=0) increases by 1.1332 times.

4) What is the estimated probability that association members will not renew their membership if the dues are increased by $40?

QUESTION-6

Prediction for Y

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -4.81 + 0.1251 X

Settings

Variable	Setting
X	40

Prediction

Fitted Probability	SE Fit	95% CI
0.548749	0.0976329	(0.359587, 0.724798)

The estimated probability that association members will not renew their membership if the dues are increased by $40 is 0.5487

5) Estimate the amount of dues increase for which 75 percent of the members are expected not to renew their association membership.

By this function,

$log(\frac{\pi}{1-\pi}) = \beta 0 + \beta 1 Xi$ we have π=0.75, we need to calculate Xi,

And solving for x_i we get 47.21945

xi can be calculated to be 47.21945. Therefore at least $47.21945 increase in dollars will cause 75% of the members not to renew their memberships.

Problem-7:

1) State the model and assumptions.

Regression Equation

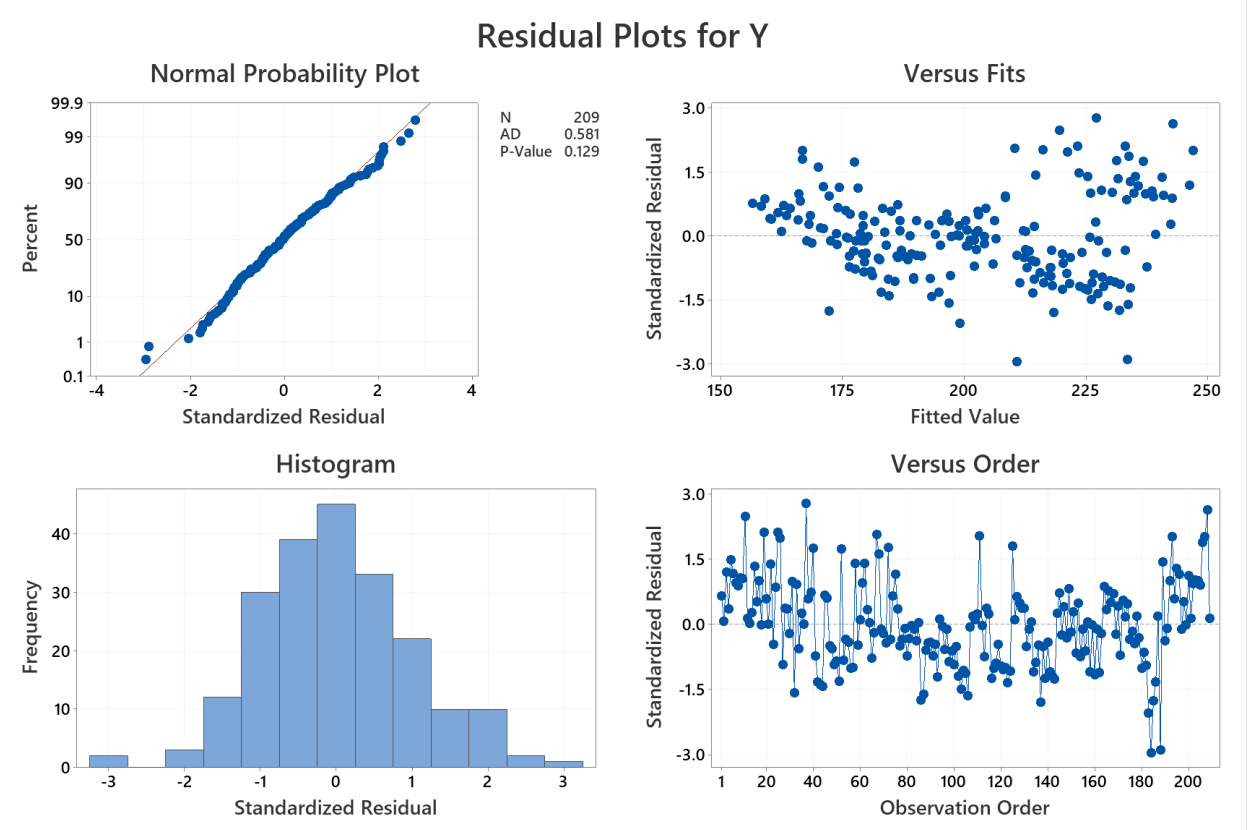
Y = 183.57 - 3.808 X1 + 1.741 X2 + 40.33 X3 - 32.72 X4 + 4.28 X5

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	183.57	5.22	35.16	0.000	
X1	-3.808	0.748	-5.09	0.000	1.01
X2	1.741	0.375	4.64	0.000	1.04
X3	40.33	3.46	11.67	0.000	1.05
X4	-32.72	9.58	-3.41	0.001	1.04
X5	4.28	3.60	1.19	0.236	1.03

Analysis of Variance

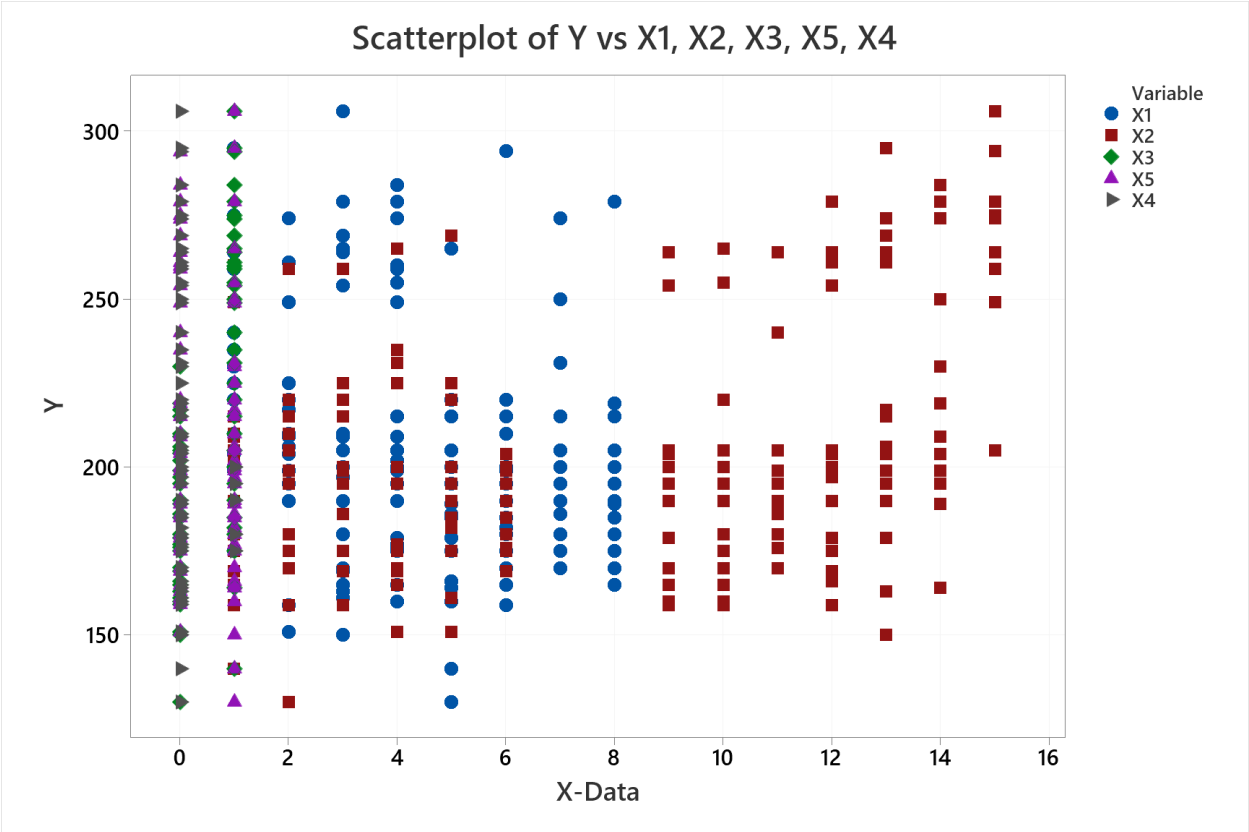
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	118091	23618.2	39.69	0.000
X1	1	15411	15411.1	25.90	0.000
X2	1	12833	12833.3	21.57	0.000
X3	1	80996	80996.2	136.11	0.000
X4	1	6938	6938.1	11.66	0.001
X5	1	840	839.7	1.41	0.236
Error	203	120802	595.1		
Lack-of-Fit	197	118839	603.2	1.84	0.224
Pure Error	6	1963	327.2		
Total	208	238893			

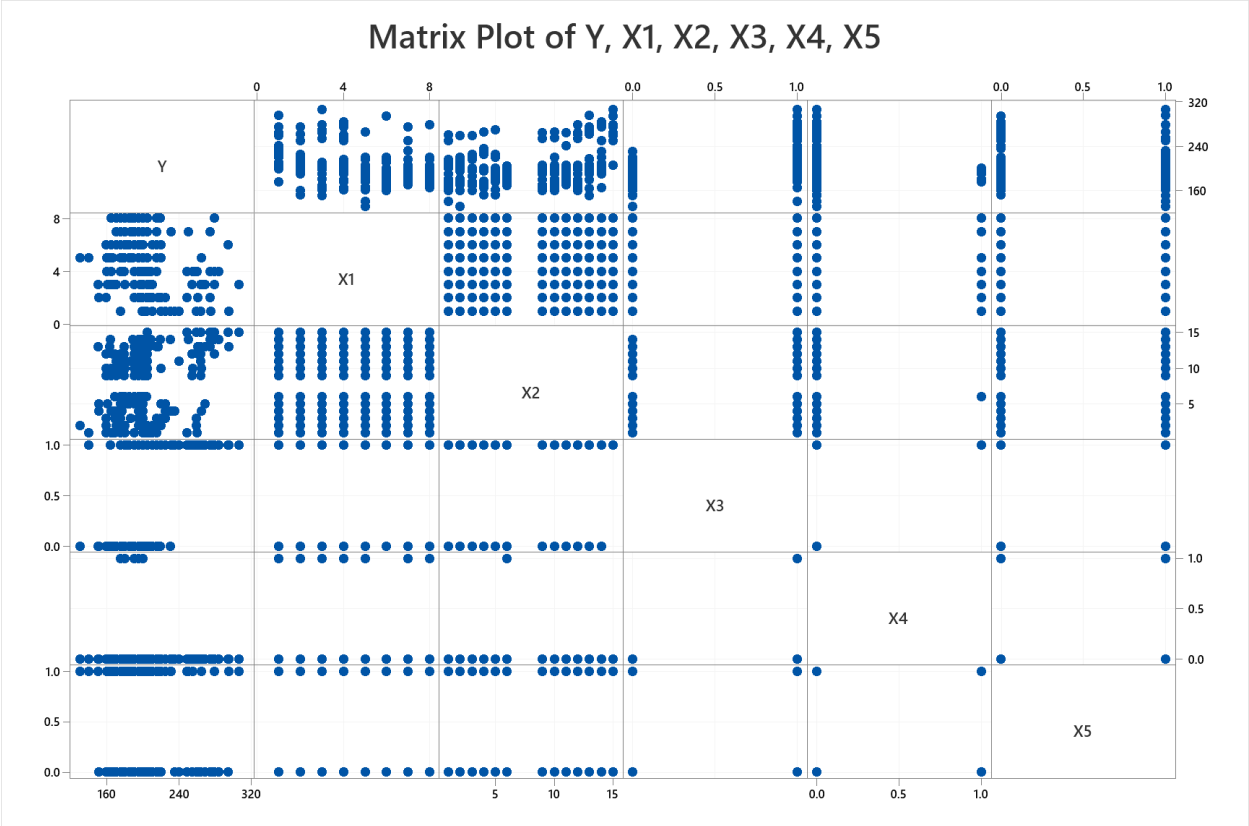


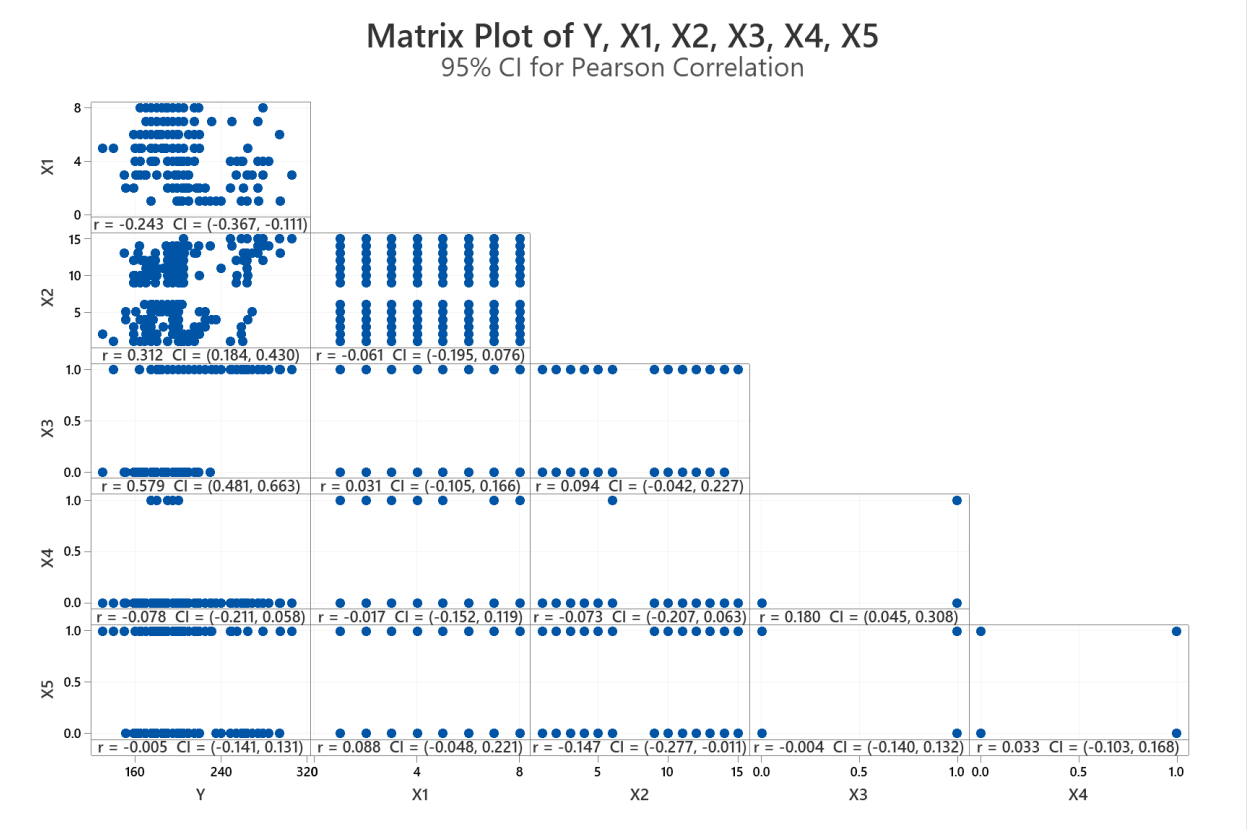
Following are the assumptions for the above tables and plots:

A normal probability graph is used to test the normality assumption. Normal probability graph residuals should approximately follow the same line. There does not seem to be any problem with the normality assumption.
In the verses plot, we can state that the points are randomly scattered and we cannot see any reasonable pattern so the regression is the better choice and constant variance assumptions are not violated.
We can state that there is a strong statistical relationship between Response variable and predictors. The model appears to be reasonable.

2)Use scatter plot matrices and correlation matrices to examine the data for linear relationships.


Scatterplot of Y vs X1, X2, X3, X5, X4
Variable
X1
X2
X3
X5
X4
Y
X-Data
300
250
200
150
0
2
4
6
8
10
12
14
16


Matrix Plot of Y, X1, X2, X3, X4, X5
Y
X1
X2
X3
X4
X5



From the above graph, we can conclude that there is a linear co-relationship between the predictors and the strength of the relationship.

We also observe a strong linear relationship between the sales prices and the predictors of the data.

3)Test for a regression relationship. Evaluate the usefulness of the model.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	118091	23618.2	39.69	0.000
X1	1	15411	15411.1	25.90	0.000
X2	1	12833	12833.3	21.57	0.000
X3	1	80996	80996.2	136.11	0.000
X4	1	6938	6938.1	11.66	0.001
X5	1	840	839.7	1.41	0.236
Error	203	120802	595.1		
Lack-of-Fit	197	118839	603.2	1.84	0.224
Pure Error	6	1963	327.2		
Total	208	238893			

Since p-value is smaller than 0.05, we reject the null hypothesis. We can state that the regression is significant.

4) Test the individual parameters.

Regression Equation

$$Y = 183.57 - 3.808\ X1 + 1.741\ X2 + 40.33\ X3 - 32.72\ X4 + 4.28\ X5$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	183.57	5.22	35.16	0.000	
X1	-3.808	0.748	-5.09	0.000	1.01
X2	1.741	0.375	4.64	0.000	1.04
X3	40.33	3.46	11.67	0.000	1.05
X4	-32.72	9.58	-3.41	0.001	1.04
X5	4.28	3.60	1.19	0.236	1.03

A: H0: β1 = 0
T-value for β1= -5.09
P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significance. We can say that X1 is a good predictor for the model.

B: H0: β2 = 0
T-value for β2= 4.64
P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significance. We can say that X2 is a good predictor for the model.

C: H0: β3 = 0
T-value for β3= 11.67
P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significance. We can say that X3 is a good predictor for the model.

D: H0: β4 = 0
T-value for β4= -3.41
P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significance. We can say that X4 is a good predictor for the model.

E: H0: β5 = 0
T-value for β5= 1.19
P-value decision: Since p-value >0.05, we failed to reject the null hypothesis therefore the regression is not significant at 5% of the level of significance. We can say that X5 is not a good predictor for the model.

5) Is there multicollinearity in the full model? Explain.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	183.57	5.22	35.16	0.000	
X1	-3.808	0.748	-5.09	0.000	1.01
X2	1.741	0.375	4.64	0.000	1.04
X3	40.33	3.46	11.67	0.000	1.05
X4	-32.72	9.58	-3.41	0.001	1.04
X5	4.28	3.60	1.19	0.236	1.03

From the above coefficient table, we can conclude that the p-value is approximately 0 and VIF is less than 5 and closer to 1. We can conclude that there exists no multicollinearity in the predicted variables.

6) Use stepwise and best subsets techniques to reduce the model.

Stepwise Selection of Terms

Candidate terms: X1, X2, X3, X4, X5

	-----Step 1----		-----Step 2----		-----Step 3----		-----Step 4----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	181.05		198.25		184.01		185.19	
X3	39.16	0.000	39.71	0.000	38.12	0.000	40.33	0.000
X1			-3.895	0.000	-3.664	0.000	-3.736	0.000
X2					1.797	0.000	1.679	0.000
X4							-32.45	0.001
S	27.7012		26.3040		25.0337		24.4189	
R-sq	33.51%		40.34%		46.22%		49.08%	
R-sq(adj)	33.19%		39.76%		45.44%		48.08%	
Mallows' Cp	61.93		36.52		14.89		5.41	
AICc	1985.60		1965.03		1945.43		1936.13	
BIC	1995.51		1978.21		1961.84		1955.77	

α to enter = 0.15, α to remove = 0.15

For this example, the stepwise procedure added a variable in each step. The process stopped when there were no variables it could add or remove from the model. The final columns display the model that the procedure produced.

There are four independent variables in our model: X3,X1,X2,X4. This model has an r-squared of 49.08% and the highest adjusted R-square.

Response is Y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	X 1	X 2	X 3	X 4	X 5
1	33.5	33.2	32.2	61.9	27.701			X		
1	9.7	9.3	7.8	157.3	32.274		X			
2	40.3	39.8	38.7	36.5	26.304	X		X		
2	40.2	39.6	38.4	37.0	26.333		X	X		
3	46.2	45.4	44.1	14.9	25.034	X	X	X		
3	44.0	43.2	42.3	23.9	25.549	X		X	X	
4	49.1	48.1	46.9	5.4	24.419	X	X	X	X	
4	46.5	45.5	43.9	15.7	25.023	X	X	X		X
5	49.4	48.2	46.8	6.0	24.394	X	X	X	X	X

The model that has a high adjusted R-Squared, a small standard error and a mallows Cp close to the number of variables. The highlighted model appears to be the best model because its cp value(5.4) is near to k+1(4+1=5) value and it has high R-square.

7) Is there multicollinearity in the reduced model? Explain.

Regression Equation

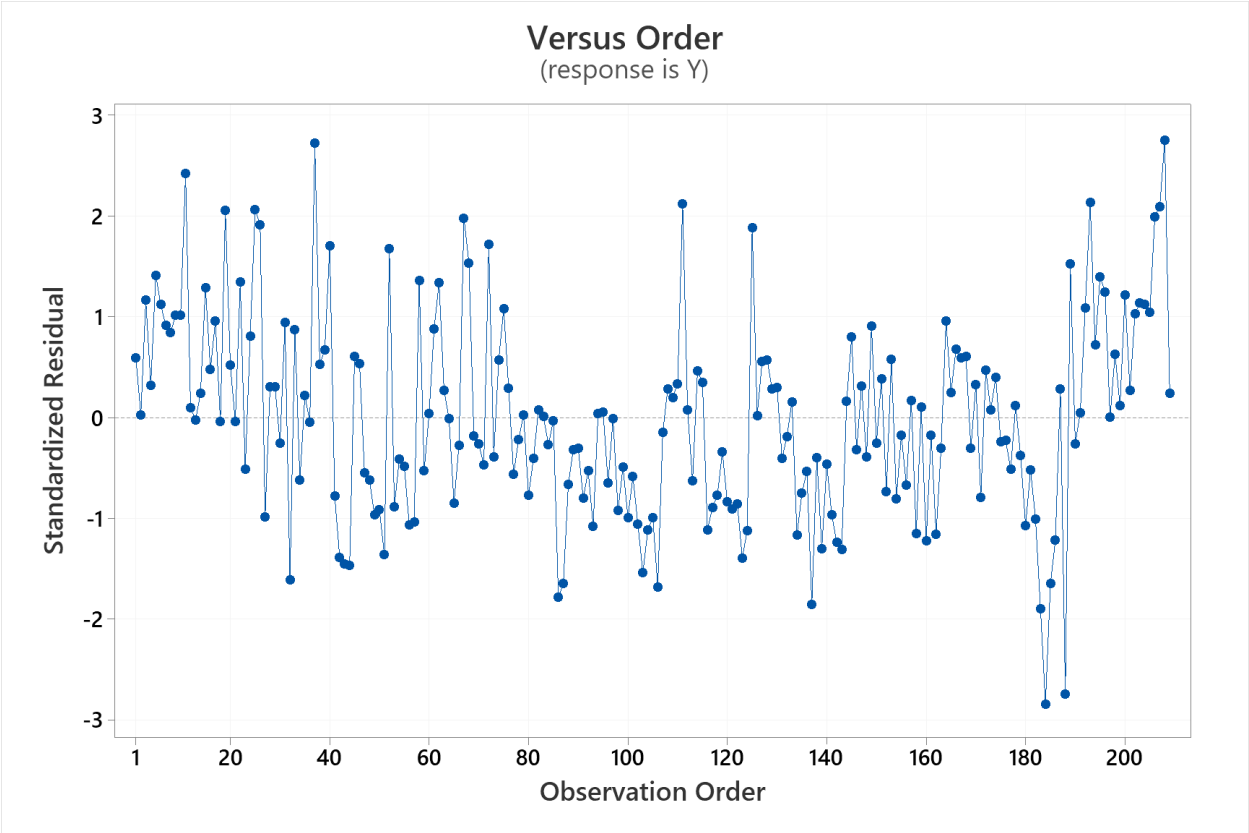
Y = 185.19 - 3.736 X1 + 1.679 X2 + 40.33 X3 - 32.45 X4

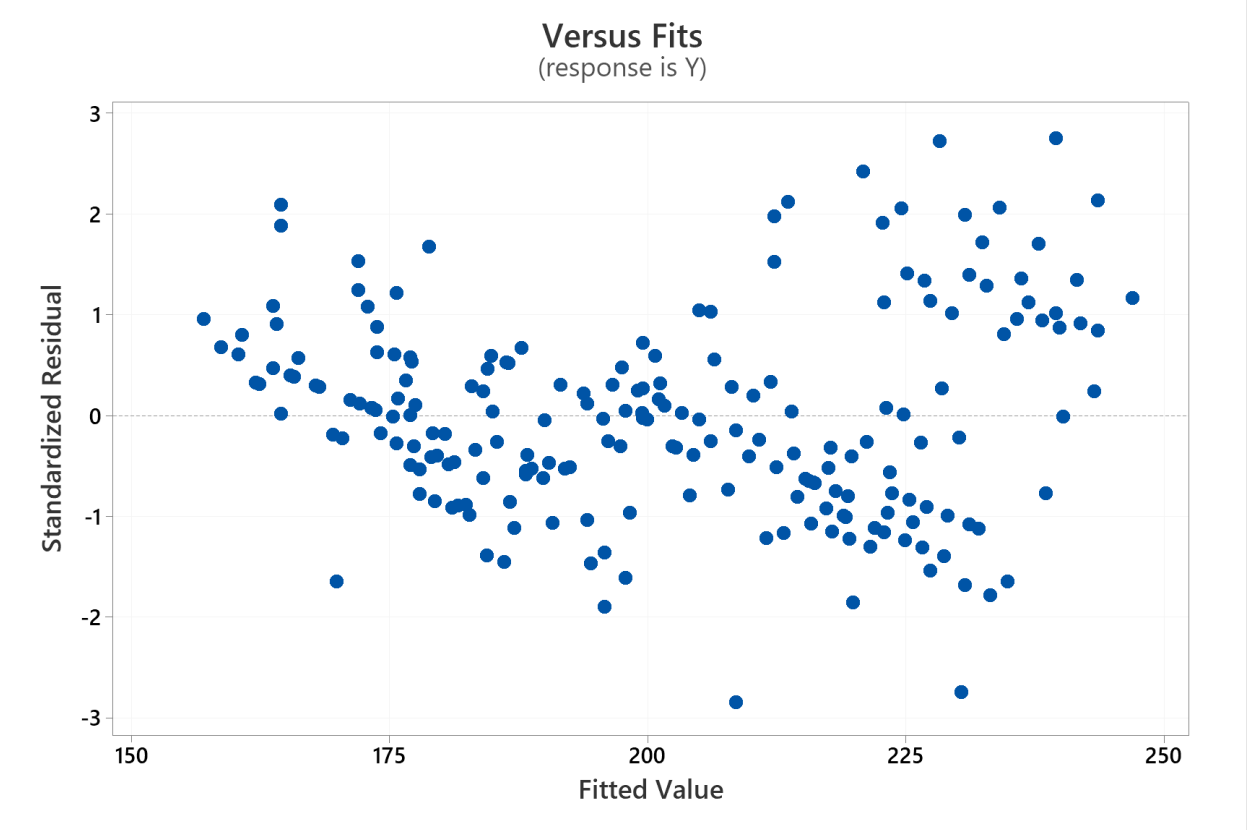
Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	185.19	5.04	(175.25, 195.14)	36.72	0.000	
X1	-3.736	0.747	(-5.208, -2.264)	-5.00	0.000	1.01
X2	1.679	0.372	(0.946, 2.412)	4.52	0.000	1.02
X3	40.33	3.46	(33.51, 47.16)	11.66	0.000	1.05
X4	-32.45	9.59	(-51.35, -13.54)	-3.38	0.001	1.04

From the above coefficient table, we can conclude that the p-value is approximately 0 and VIF is less than 5 and equal to 1. We can conclude that there exists no multicollinearity in the predicted variables.

8) Using the reduced model, examine the data for outliers and influential observations.





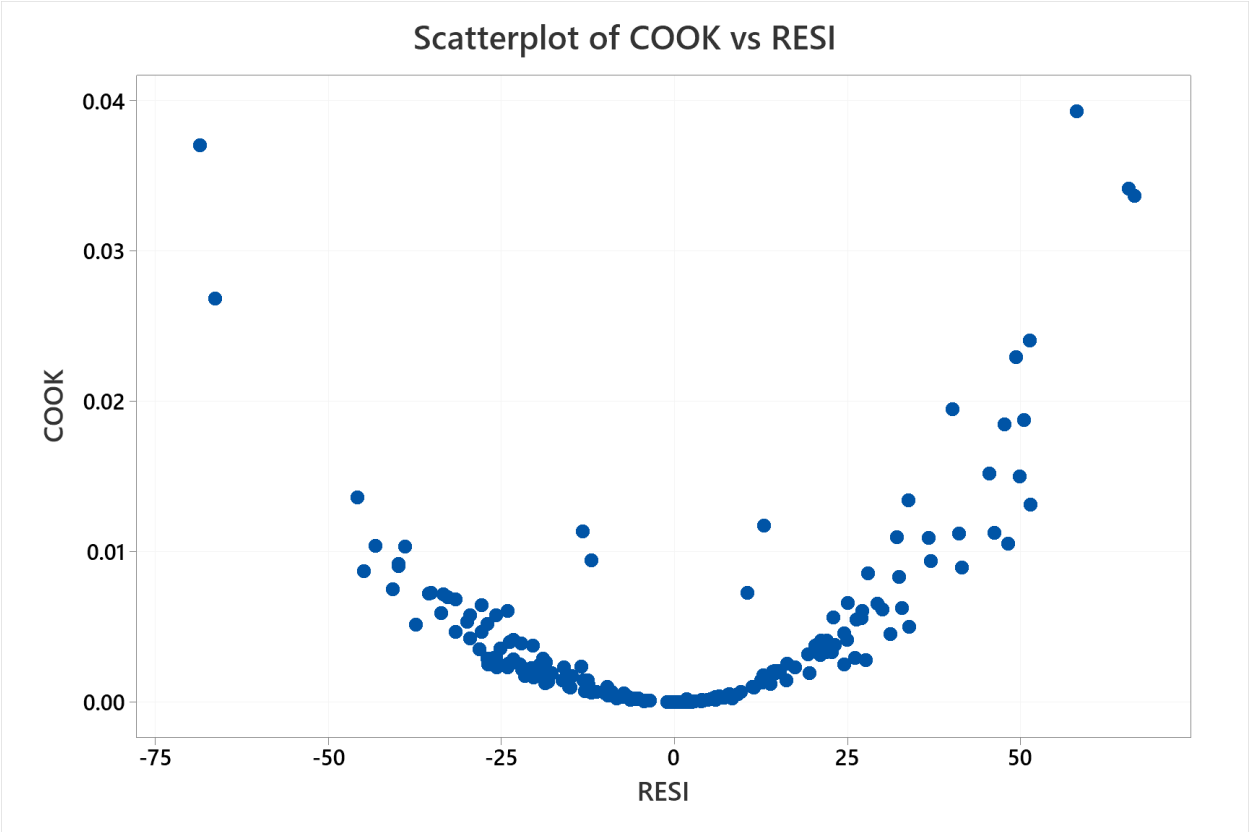
We can see that in residual vs fitted graph, there is some pattern that indicates a non linear association in the data.

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
11	279.00	220.83	4.40	(212.16, 229.50)	58.17	2.42	2.45	0.032432
19	274.00	224.57	3.98	(216.72, 232.41)	49.43	2.05	2.07	0.026556
25	284.00	234.09	3.22	(227.75, 240.43)	49.91	2.06	2.08	0.017338
37	294.00	228.30	3.67	(221.07, 235.53)	65.70	2.72	2.77	0.022548
79	200.00	199.42	9.55	(180.59, 218.25)	0.58	0.03	0.03	0.152947
85	195.00	195.68	9.39	(177.18, 214.19)	-0.68	-0.03	-0.03	0.147740
92	180.00	191.95	9.28	(173.65, 210.24)	-11.95	-0.53	-0.53	0.144402
101	175.00	188.21	9.23	(170.01, 206.41)	-13.21	-0.58	-0.58	0.142933
111	265.00	213.57	2.93	(207.79, 219.34)	51.43	2.12	2.14	0.014393
114	195.00	184.47	9.24	(166.25, 202.70)	10.53	0.47	0.46	0.143334
153	190.00	177.00	9.45	(158.37, 195.63)	13.00	0.58	0.58	0.149743
173	175.00	173.27	9.64	(154.27, 192.27)	1.73	0.08	0.08	0.155751
184	140.00	208.53	3.66	(201.31, 215.75)	-68.53	-2.84	-2.89	0.022484
188	164.00	230.36	3.23	(223.98, 236.74)	-66.36	-2.74	-2.79	0.017546
193	295.00	243.62	3.92	(235.89, 251.35)	51.38	2.13	2.15	0.025777
207	215.00	164.46	3.54	(157.48, 171.43)	50.54	2.09	2.11	0.020988
208	306.00	239.51	3.60	(232.41, 246.61)	66.49	2.75	2.80	0.021741

Obs	Cook's D	DFITS	
11	0.04	0.448797	R
19	0.02	0.341610	R
25	0.01	0.276076	R
37	0.03	0.420001	R
79	0.00	0.010975	X
85	0.00	-0.012568	X
92	0.01	-0.216897	X
101	0.01	-0.238251	X
111	0.01	0.258632	R
114	0.01	0.190130	X
153	0.01	0.241851	X
173	0.00	0.033105	X
184	0.04	-0.438159	R
188	0.03	-0.372417	R
193	0.02	0.349806	R
207	0.02	0.308872	R
208	0.03	0.417227	R

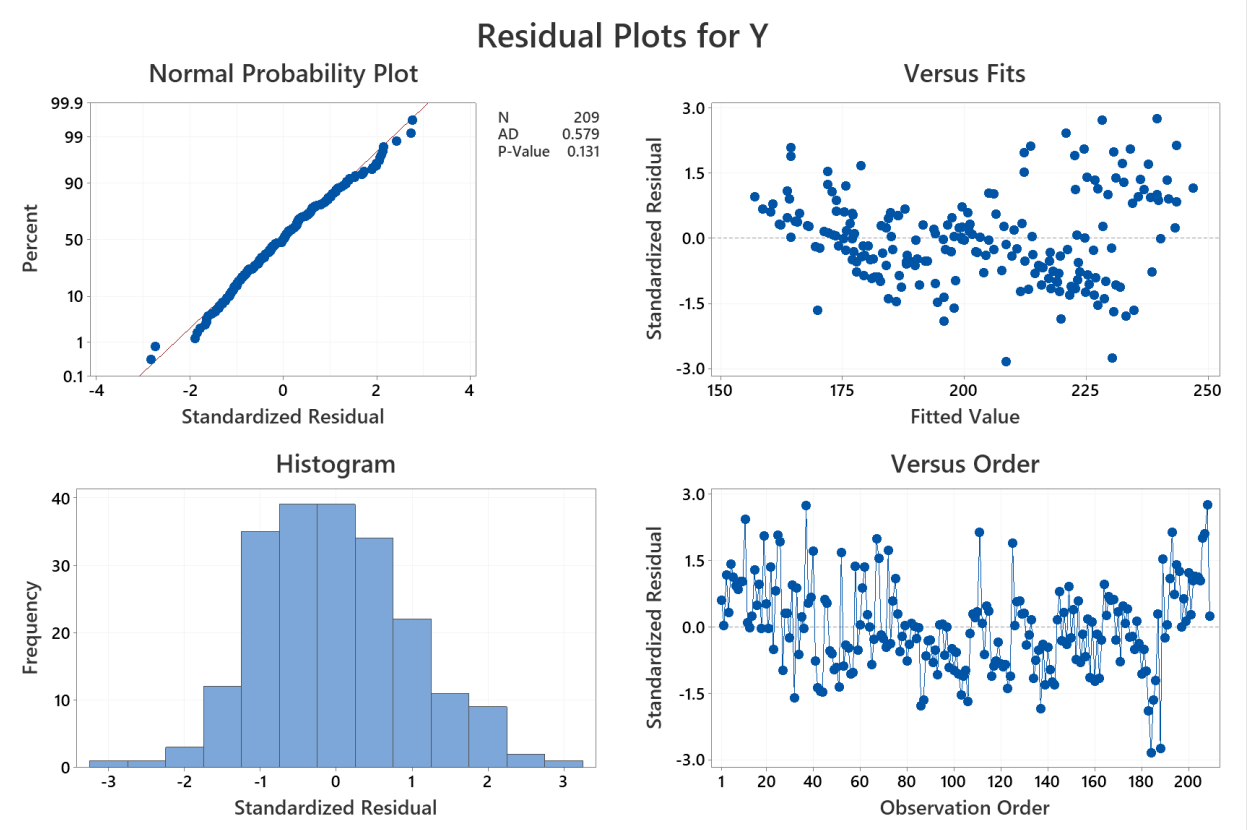
R Large residual
X Unusual X



Looking at the above figure, we can state that the outlier presents in the data.

In the above graph and fits observations, we can see that, observations **11, 184, 188, 193 and 208** have high leverage points and most influential points according to cook's distance

9) Check the assumptions for the reduced model.



In this above normality graph, the residuals are normally distributed and follow a straight line.

In the residual graph, the residuals are scattered and we can say that there is no clear pattern present in it. These residual graphs are randomly scattered and independent of each other and we cannot see any reasonable pattern so the regression is the better choice and constant variance assumptions are not violated.

Looking at the versus order plot, there is no trend seen to be followed and our independence assumption also holds true.

The variables follow the straight line and from the above-scattered diagram, we can state that there is a strong statistical relationship between predictor and response variable. The model appears to be reasonable.

10) If there is a problem with the reduced model, propose a solution. (You don't have to do anything, just make suggestions.)

Multicollinearity is not present in the reduced model as stated earlier. The problem with the reduced model is outliers and high influential points according to cook's distance.Influential point and outlier greatly affects the slope of the regression. To solve this problem we have various solutions.

1. *We can trim the dataset and replace outliers with the nearest good data as opposed to truncating them completely from the dataset.*

2. *Replace outliers with the mean and median for that variable to avoid missing data points.*
3. *Using the training data, find the best hyperplane and remove points which are far away from that line and retrain the model.*