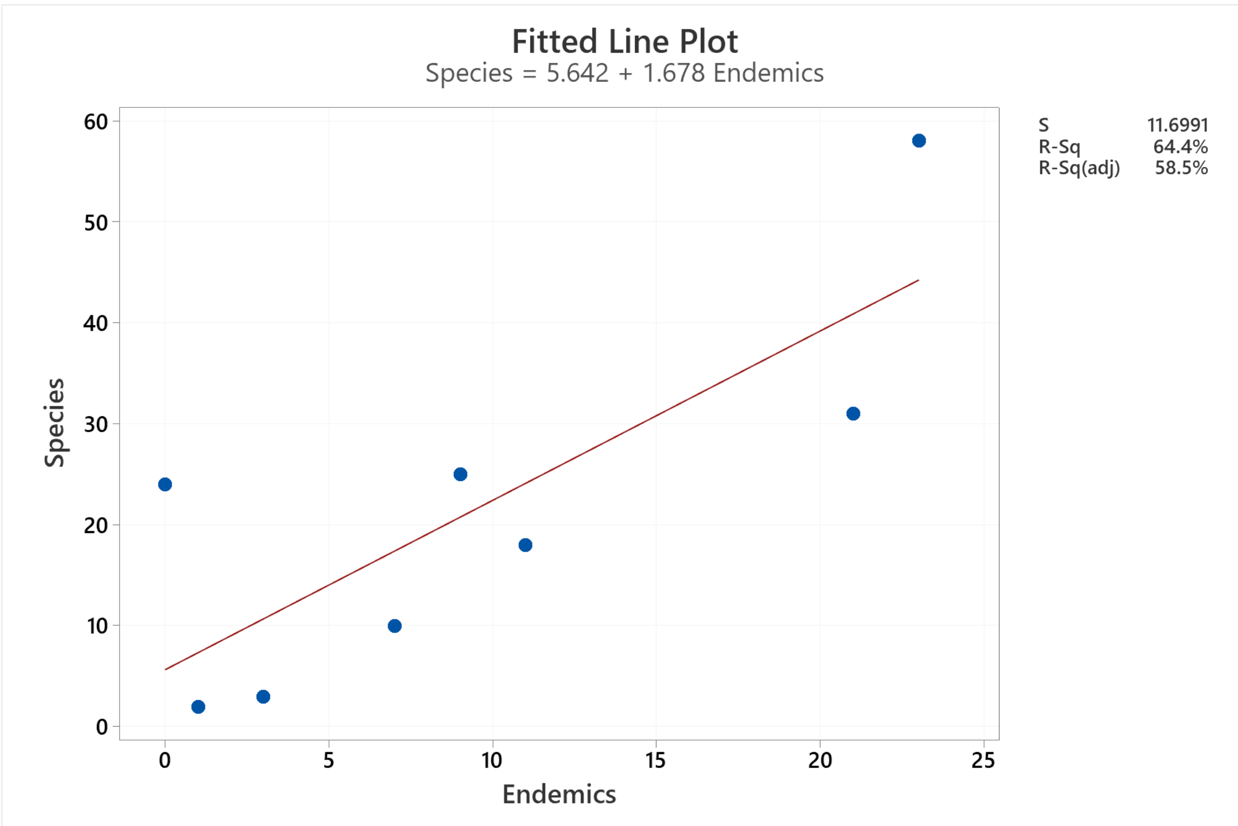
Mrinal Chaudhari **Exam-1** **Stat641**

Date-10/18/2021

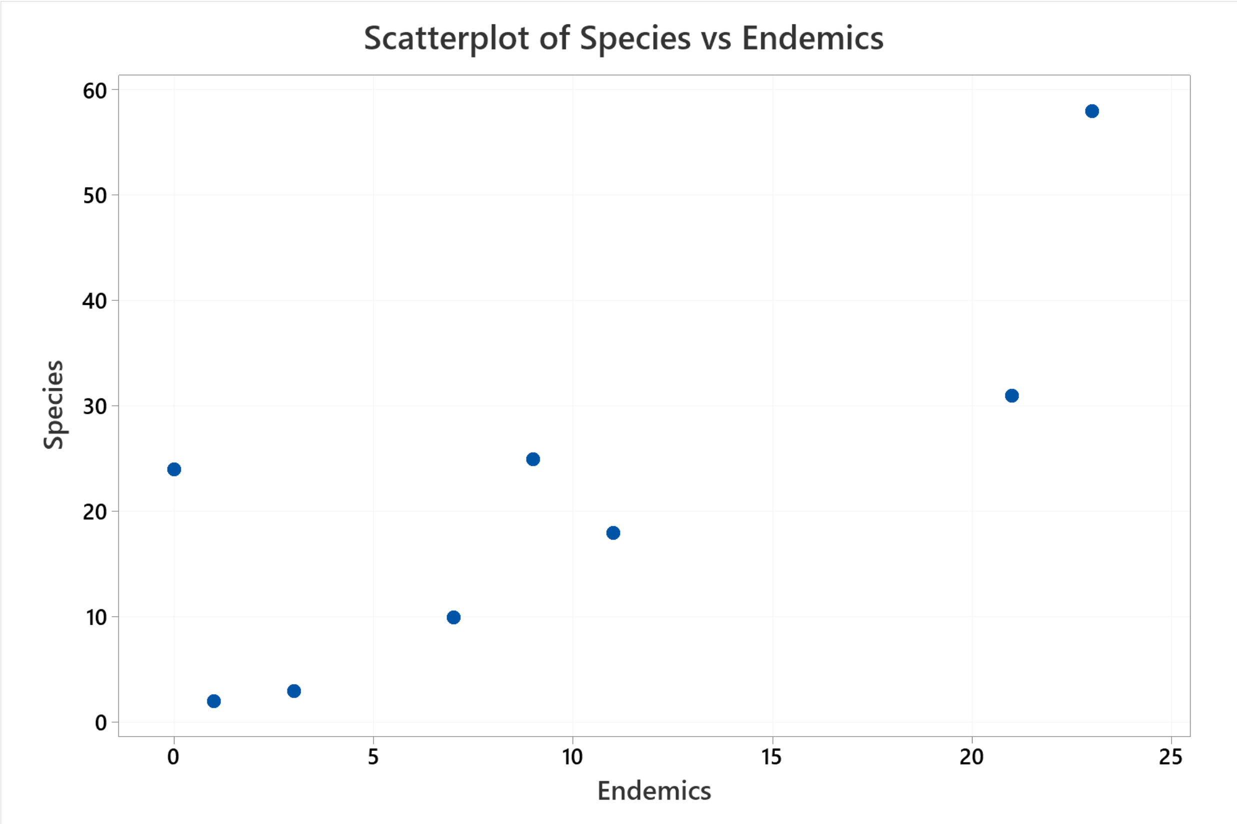
Problem 1:
Solution:

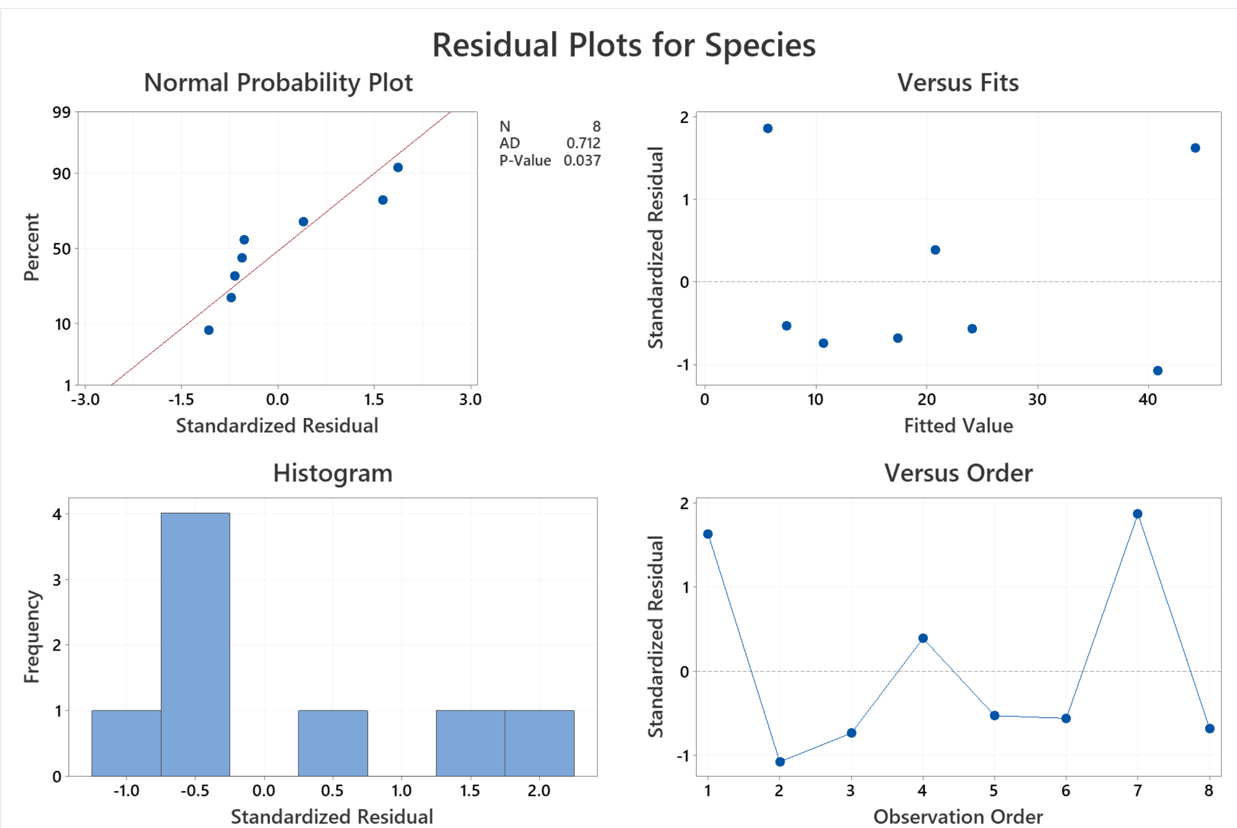
1)Write down the simple linear regression model and corresponding assumptions. Please propose at least one graphical method to check each assumption.

Solution:



Mrinal Chaudhari Exam-1 Stat641





In the above normal probability graph, we can reject the null hypothesis and conclude that the regression is significant. We can also state the residuals are not normally distributed.
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 violated.
The variables follow the straight line and from the above-scattered diagram, we can state that there is a strong statistical relationship between species and Endemics. The model appears to be reasonable.

2)Write down the estimated regression line.
Solution:

Regression Equation

Species = 5.64 + 1.678 Endemics

3)What are the missing standard errors for the intercept and the slope?

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	5.64	6.32	(-9.81, 21.10)	0.89	0.406	
Endemics	1.678	0.509	(0.432, 2.924)	3.30	0.016	1.00

The missing standard errors for the intercept and the slope are 6.32 and 0.509

4)Compute a 95% confidence interval for the intercept β_0

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	5.64	6.32	(-9.81, 21.10)	0.89	0.406	
Endemics	1.678	0.509	(0.432, 2.924)	3.30	0.016	1.00

$\beta_0 = 5.640$
p-value for intercept β_0 is 0.406

Since the p-value is greater than 0.05, we failed to reject the null hypothesis. We can conclude that the 95% confidence interval for the intercept is not significant. We can conclude that β_0 is zero when endemics=0

C.I for the intercept β_0 is (-9.81,21.10)

5)Does a 95% confidence interval for the slope β_1 contain 0 or not? Explain without actually computing the interval?

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	5.64	6.32	(-9.81, 21.10)	0.89	0.406	
Endemics	1.678	0.509	(0.432, 2.924)	3.30	0.016	1.00

$\beta_1 = 1.678$

p-value for intercept β_1 is 0.016

Since the p-value is smaller than 0.05, we reject the null hypothesis. We can conclude that the 95% confidence interval for the intercept is significant. We can conclude that β_0 does not equal 0.

6)What is the fitted response value for Endemics = 23? What is the residual?

WORKSHEET 1

Prediction for Species

Regression Equation

Species = 5.64 + 1.678 Endemics

Settings

Variable	Setting
Endemics	23

Prediction

Fit	SE Fit	95% CI	95% PI
44.2404	8.07723	(24.4761, 64.0046)	(9.45377, 79.0269)

The fitted response value for Endemics = 23 is 44.2404

Residual=$e_i = y_i - \hat{y}_i$ (observed_value-fitted_value)=13.7596

7)Given the R output and the fact that there is only 1 predictor variable in the model, compute the F-value for the ANOVA table

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1486.7	64.42%	1486.7	1486.7	10.86	0.016
Endemics	1	1486.7	64.42%	1486.7	1486.7	10.86	0.016
Error	6	821.2	35.58%	821.2	136.9		
Total	7	2307.9	100.00%				

F-value from ANOVA table=10.86

8)Given the R output, test the hypothesis: H_0: $\beta_1 = 2$ vs. H_a: $\beta_1 \neq 2$ at $\alpha = 0.05$.

From the anova table,
p-value=P(F>10.86)
p-value=1-P(F<10.86)
p-value=0.016
Since the p-value is less than the level of significance i.e 0.05, we reject the null hypothesis. We can conclude that there is a linear relationship between the dependent variable endemics and independent variable species.

9)Write down the design matrix X for this data.

Matrix X

1	23
1	21
1	3
1	9
1	1
1	11
1	0
1	7

Problem 2:
Solution:

1) How many observations are used in this analysis?
Solution:
DF=n-1

23=n-1
n=23+1=24
The number of observations used in the analysis is 24

2)Fill in the missing values in the above table.

SS_model=SS_total-SS_error

MS=SS/DF

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	**627.817**	**209.2723**	**68.119**	**0**
Error	**20**	61.44300	**3.07215**		
Corrected Total	23	689.2600			

F=MS_model/MS_error

F=68.119

P-value=0

3)Write down the linear model and the hypothesis for the *F* test. What is your conclusion from this test at the 5% level?

Since $p<0.05$, we reject the null hypothesis. We can conclude that the regression is significant at a 5% significance level.

4)Compute and interpret R^2 for this model

$R^2=SS_R/SS_T$

R^2=627.817/689.26=0.9108

R^2 stated that 91.08% of the variation in the dependent variable is explained by the independent variable.

5)Compute the adjusted R^2 for this model.

Adj R^2=1-(MSE/MST)=1-0.1025= 0.8975

6)What is the estimate of σ?

Estimate of σ=sqrt(MSE)=sqrt(3.07215)

σ=1.753

Problem 3:
Solution:

Continued on next page…..

a:

problem 3.
Solution

show that an equivalent way to perform the test for significance of regression in multiple linear regression is to base the test on R^2 as follows :

To test $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ vs
H_1 : at least one β not zero,
calculate $F_0 = \dfrac{R^2 (n-p)}{k(1-R^2)}$ and,
to reject H_0 if computed value F_0 exceeds $F_{\alpha, k, n-p}$, where $p = k+1$

→ Null hypothesis $H_0 : B_1 = B_2 = \cdots = B_k$
Alternative Hypothesis H_A : at least one $\beta_j \neq 0$

$$T_0 = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$$

$$= \frac{\hat{\beta}_k}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

$$= \frac{S_{xy} / S_{xx}}{\sqrt{\frac{SS_E / (n-p)}{S_{xx}}}}$$

$$= \frac{R\sqrt{n-p}}{K\sqrt{1-R^2}}$$

$$\Rightarrow T_0^2 = \frac{R^2 (n-p)}{K(1-R^2)}$$

$$F_0 = \frac{R^2 (n-p)}{K(1-R^2)}$$

Therefore, the test for significance of regression in multiple linear regression is to base the test on R^2 as to test

$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_K$ Vs

H_1 : at least one $\beta_j \neq 0$

$$F_0 = \frac{R^2 (n-p)}{K(1-R^2)}$$

and reject $H_0 : \beta_K = 0$ if $F_0 > F_{\alpha, 1, n-p}$

b) Suppose that a linear regression model with $k=2$ regressor has been fit to $n=25$ observations and $R^2 = 0.90$. Test for the significance of regression at $\alpha = 0.05$.

Solution: $p = k+1 = 2+1 = 3$

$$F_0 = \frac{R^2 (n-p)}{k(1-R^2)}$$

$$= \frac{0.90(25-3)}{2(1-0.90)} = \frac{19.8 \ \cancel{20.7}}{0.2 \ \cancel{0.2}}$$

$$F_0 = 99$$

$$df = n-1 = 25-1 = 24$$

From F-table, the critical value at 0.05 level of significance is 3.44

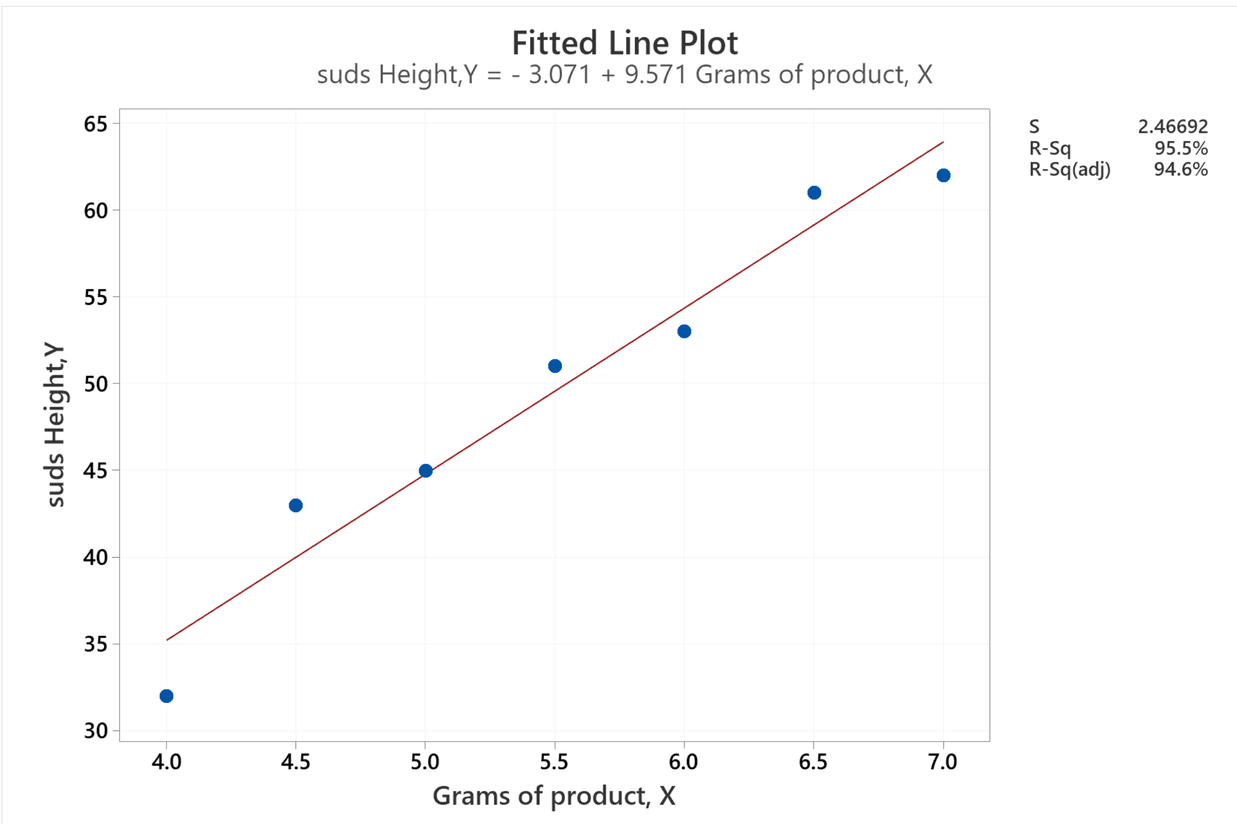
The F calculated value is greater than critical value, we reject the null hypothesis.

Problem 4:
Solution:

a)Determine the best fitting equation

Regression Equation

suds Height,Y = -3.07 + 9.571 Grams of product, X



b)Test the equation for statistical significance.

Analysis of Variance

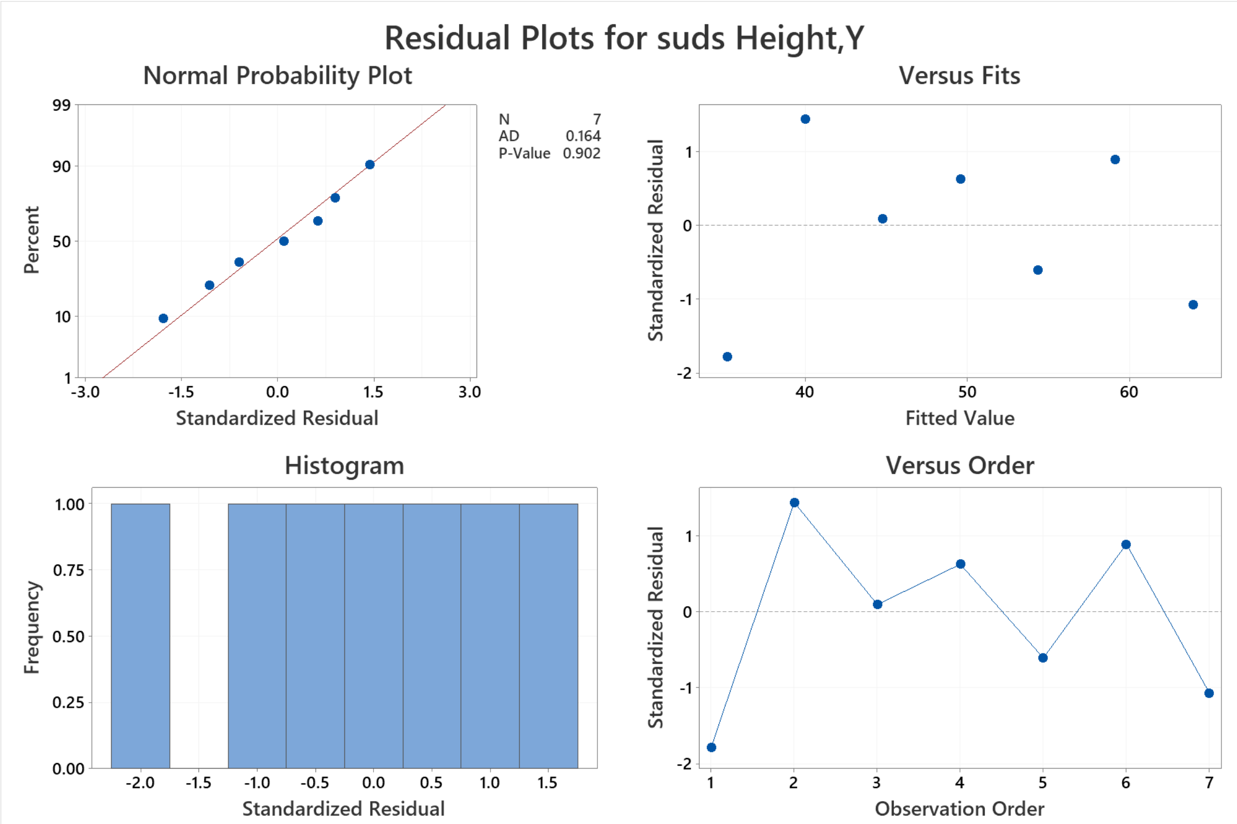
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	641.29	95.47%	641.29	641.286	105.38	0.000
Grams of product, X	1	641.29	95.47%	641.29	641.286	105.38	0.000
Error	5	30.43	4.53%	30.43	6.086		
Total	6	671.71	100.00%				

Since the p-value is smaller than the level of significant i.e 0.05, we reject the null hypothesis. We can conclude that the regression is significant and the height of soap suds in the dishpan is of importance to soap manufacturers.

c)Calculate the residuals and see if there is any evidence suggesting that a more complicated model would be more suitable.

Using Minitab, we got the below output for residual.

C1	C2	C3	C4
Grams of product, X	suds Height,Y	FITS	RESI
4.0	32	35.2143	-3.21429
4.5	43	40.0000	3.00000
5.0	45	44.7857	0.21429
5.5	51	49.5714	1.42857
6.0	53	54.3571	-1.35714
6.5	61	59.1429	1.85714
7.0	62	63.9286	-1.92857



To check whether we need a more complicated model or not, we have these residual graphs to interpret.

The normal probability graph indicates that the residuals are normally distributed and follow the linear relationship.

In the scattered graph we can see that the residuals are randomly scattered and we cannot determine the specific pattern from it.

All this evidence indicates that there is no need for a complicated model to fit this data.

Problem 5:
Solution:

A. Fit a multiple linear regression model relating gasoline mileage Y (miles per gallon) to engine displacement X_1 (cubic inches) and weight X_2.

Solution:

Regression Equation

Y=Miles/gal = 36.53 - 0.0321 X1=Displacement - 0.00199 X2=Weight

B. Construct the analysis of variance table and test for significance of the regression model.

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	966.77	78.12%	966.771	483.386	51.77	0.000
X1=Displacement	1	955.34	77.20%	45.243	45.243	4.85	0.036
X2=Weight	1	11.43	0.92%	11.431	11.431	1.22	0.278
Error	29	270.77	21.88%	270.773	9.337		
Lack-of-Fit	28	270.17	21.83%	270.168	9.649	15.95	0.196
Pure Error	1	0.60	0.05%	0.605	0.605		
Total	31	1237.54	100.00%				

The null hypothesis, H_0:β_1=β_2=0

The alternative hypothesis, H_A: $B_j \neq 0$ for at least one j

Since the p-value is smaller than the level of significance i.e p-value<0.05, we can reject the null hypothesis. The regression is significant and we can state that there is linearity between the dependent and independent variables.

C. What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement and eight?

Solution:

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
3.05565	78.12%	76.61%	340.663	72.47%	168.63	173.01

R-square is 78.12%

It's a coefficient of determination (in this case =0.7812) and it is a measure of the amount of variability in y explained by x. Its value lies between 0 and 1. If the value is greater then the model is good.

In this case, we can conclude that the model is good as R-square is 78%. We can say that 78% of the variation in gasoline milage is explained by the independent variable.

D. Find a 95% confidence interval for the slopes of the regression model and interpret it.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	36.53	2.91	(30.57, 42.48)	12.55	0.000	
X1=Displacement	-0.0321	0.0146	(-0.0620, -0.0023)	-2.20	0.036	9.69
X2=Weight	-0.00199	0.00180	(-0.00568, 0.00169)	-1.11	0.278	9.69

We are 95 % confident that the displacement leads to a decrease in milage between(-0.0620, -0.0023). We can state that there are significant linear relationship presents in the dependent and independent variables.

We are 95 % confident that the weight leads to a decrease in milage between(-0.00568, 0.00169). We cannot predict that there are significant linear relationship presents in the dependent and independent variables.

E. Find a 95% confidence interval on the mean gasoline mileage if the engine displacement is 275 in^3 and weight 3000 (lbs) and interpret it.

Solution:

EXAM1Q5

Prediction for Y=Miles/gal

Regression Equation

Y=Miles/gal = 36.53 - 0.0321 X1=Displacement - 0.00199 X2=Weight

Settings

Variable	Setting
X1=Displacement	275
X2=Weight	3000

Prediction

Fit	SE Fit	95% CI	95% PI
21.7058	1.07036	(19.5167, 23.8950)	(15.0840, 28.3277)

We are 95 % confident that the mean gasoline mileage if the engine displacement is 275 in^3 and weight 3000 (lbs) is (19.5167, 23.8950).

F. Suppose that we wish to predict the gasoline mileage obtained from a car with a 275 in^3 engine and 3000 (lbs) weight. Give a point estimate, Y^, of mileage. Find a 95% prediction interval on the mileage and interpret it.

Solution:

Prediction

Fit	SE Fit	95% CI	95% PI
21.7058	1.07036	(19.5167, 23.8950)	(15.0840, 28.3277)

The 95% prediction interval on mileage is (15.0840, 28.3277) when the gasoline mileage is obtained from a car with a 275 in^3 engine and 3000 (lbs) weight.

The range of the prediction interval is somewhat wider than the confidence interval

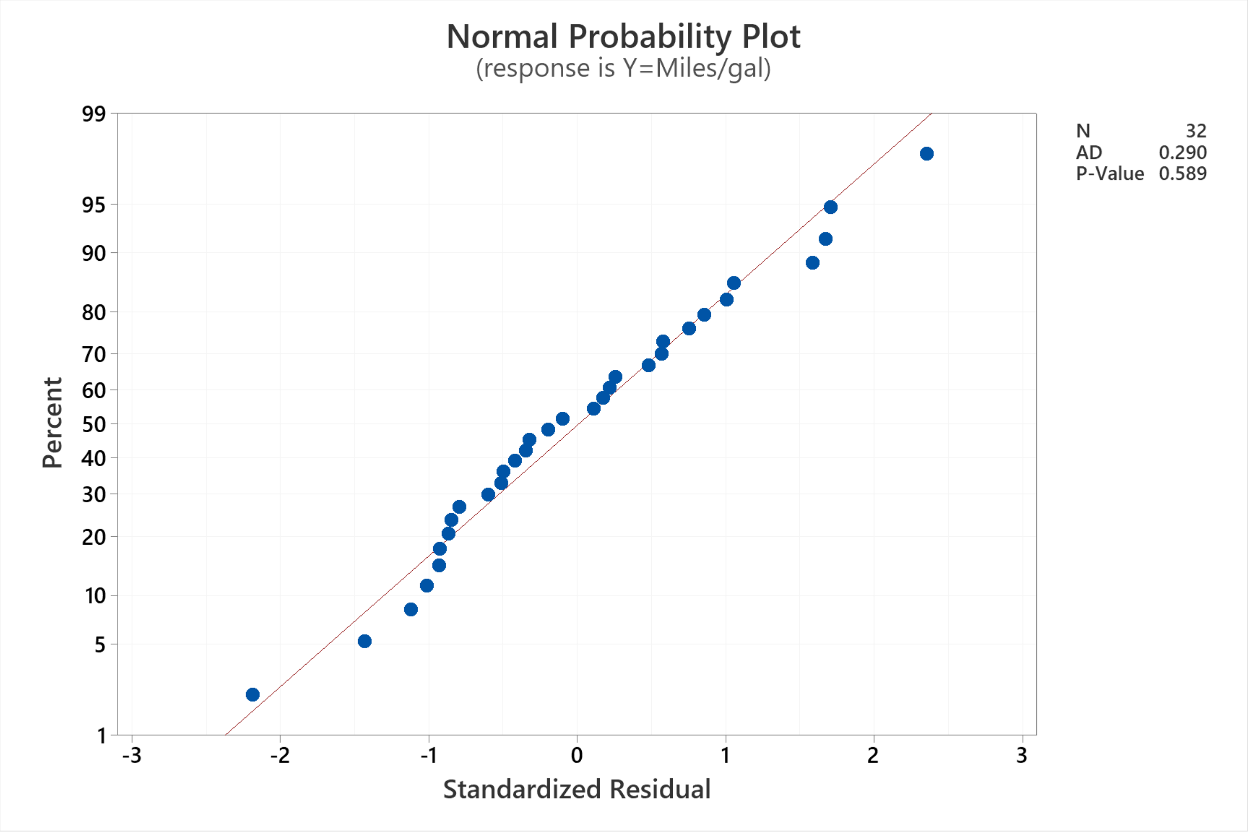
G. Compare the two intervals obtained in parts (e) and (f). Explain the difference between them. Which one is wider, and why?

Solution:

Comparing results obtained from parts 5. e and 5. f, we can state that the prediction intervals are wider compared to the confidence interval. There is a difference between CI and PI, CI is used when the given predicted setting is lies between the range, and PI is used when we can expect the future data in that range. Population mean and variance makes prediction interval wider and gives detailed information about the data point

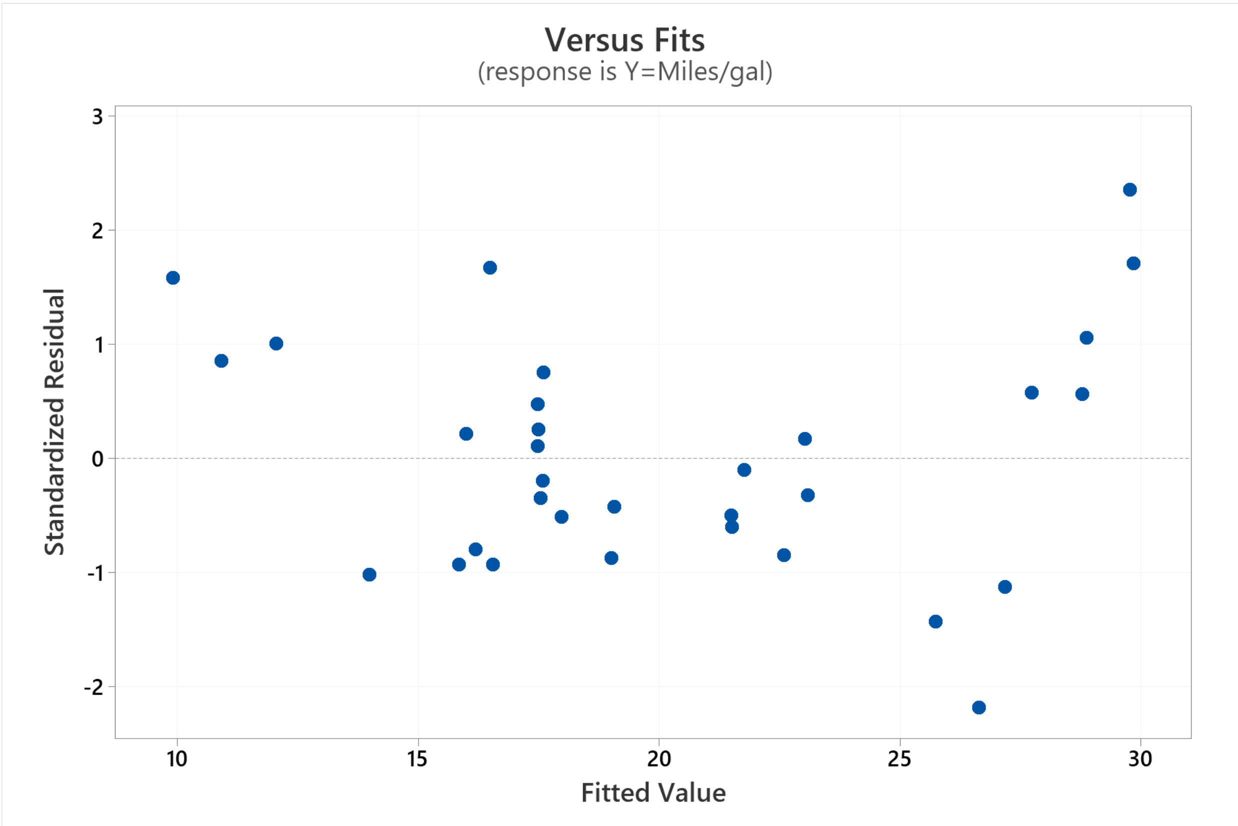
H. Construct the following residual plots and comment on model adequacy:

(1) Normal probability plot



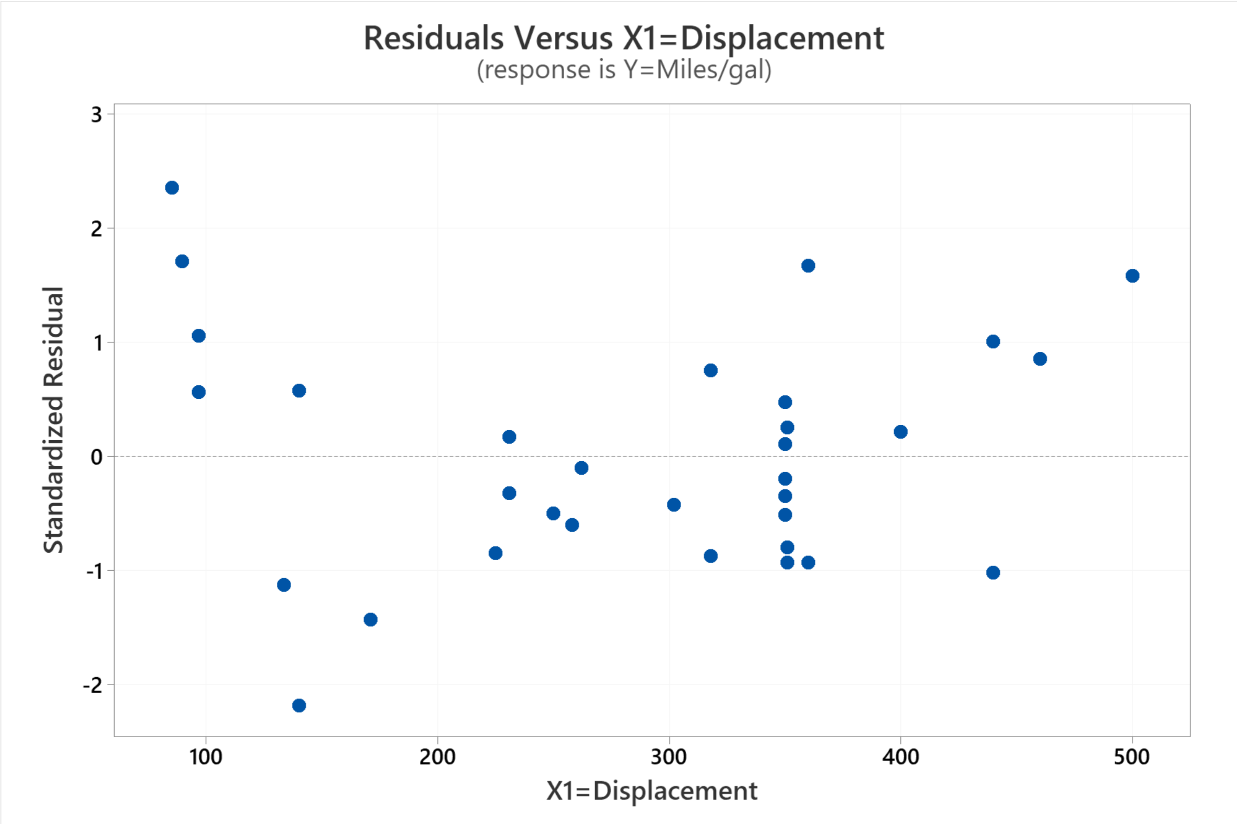
In this graph, the residuals are normally distributed and follow a straight line.

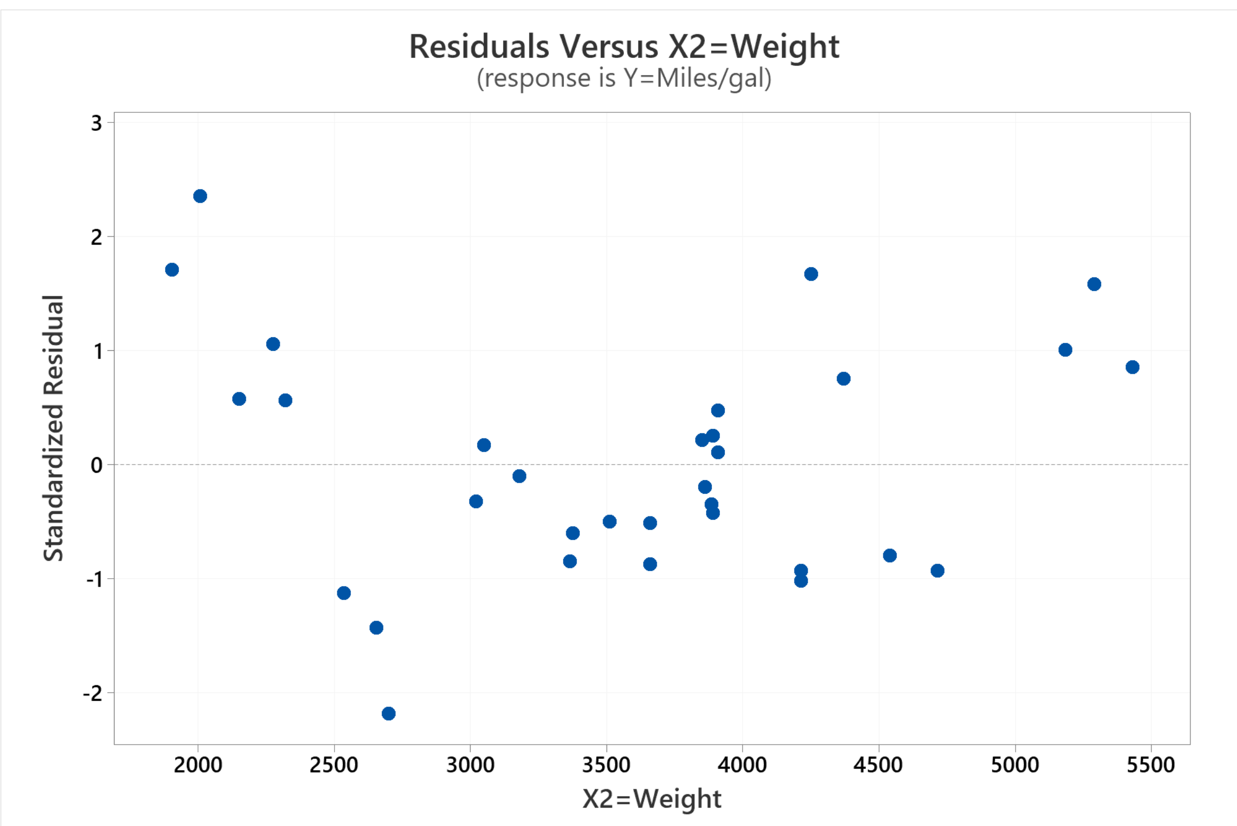
(2) Plot residuals against



In this graph, the residuals are scattered and we can say that there is no clear pattern is present in it.

(3) Plot residuals against X_i


Residuals Versus X1=Displacement
(response is Y=Miles/gal)
Standardized Residual
3
2
1
0
-1
-2
100
200
300
400
500
X1=Displacement


Residuals Versus X2=Weight
(response is Y=Miles/gal)
Standardized Residual
3
2
1
0
-1
-2
2000
2500
3000
3500
4000
4500
5000
5500
X2=Weight

These residuals versus graphs are randomly scattered and independent of each other.