**MODERN COLLEGE OF ARTS,SCI. & COMM. PUNE-05.**

**DEPARTMENT OF STATISTICS.**

**M.Sc.( I ) Sem II**

**ST- 28**

**EXPT.NO. 3.**

**TITLE : Lack of fit of the regression model**

**-------------------------------------------------------------------------------------**

1. **Investigate regression model and fit the same also test goodness of fit of the model.**

|  |  |
| --- | --- |
| X | Y |
| 1.0 | 10.84 |
| 1.0 | 9.30 |
| 2.0 | 16.35 |
| 3.3 | 22.88 |
| 3.3 | 24.35 |
| 4.0 | 24.56 |
| 4.0 | 25.86 |
| 4.0 | 29.16 |
| 4.7 | 24.59 |
| 5.0 | 22.25 |
| 5.6 | 25.90 |
| 5.6 | 27.20 |
| 5.6 | 25.61 |
| 6.0 | 25.45 |
| 6.0 | 26.56 |
| 6.5 | 21.03 |
| 6.9 | 21.46 |

1. Perform a thorough analysis of the results including residuals plots.
2. Perform the appropriate test for lack of fit.
3. Find E (M SPE) and E(M SLOF) and comment on lack of fit.

Q.2

On an appropriate measure of the whiteness of rayon(y) the engineers

conducting this experiment wish to minimize this measure. The experiment

results given as follows:

|  |  |
| --- | --- |
| X | Y |
| 75 | 60 |
| 100 | 112 |
| 100 | 136 |
| 125 | 160 |
| 125 | 150 |
| 150 | 152 |
| 175 | 156 |
| 175 | 124 |
| 200 | 124 |
| 200 | 104 |
| 75 | 85 |
| 75 | 88 |
| 100 | 97 |
| 100 | 100 |
| 125 | 142 |
| 125 | 150 |
| 150 | 140 |

1. Perform a thorough analysis of the results including residuals plots.
2. Perform the appropriate test for lack of fit.
3. Find E (M SPE) and E(M SLOF) and comment on lack of fit.

THEORY

A Statistical test that can be used to assess whether or not a model of higher order is necessary

n : Total number of observations

d : Number of distinct predictor values at which Obserrations were taken .

r : n-d : Number of replicates.

k : No of regressor

Divides the sum of squares due to error (SSE) into two sources of variability.

Lack of fit : Variability that cannot be explained by the current model as a result of leaving out one or more important factors (i.e. higher order terms.

Pure error : Variability that cannot be by any model .

Note: Can only be used if there have been multiple responses taken at the same predictor values

E(MSPE) = σ2

E(MSLOF) =

Hypotheses :

Ho: Model is an appropriate fit for the data ( i.e. No lack of fit )

VS

H1: Model is not appropriate fit for the data (i.e. lack of fit)

Error sums of squares SSE = SSLF + SSPE

Lack of fit : SSLF =

Pure error : SSPE =

Where : Sample mean of responses at

Test statistic:

Note: can only be run if k < d . otherwise , scatterplots must be used to gauge lack of fit.

Test criteria :

Reject Ho , Fcal > Ft.V.

Now suppose we fit simple linear regression for the given data. Now we check lack of fit test.

Ho : is an approximate fit for the data .

H1 : is not an approximate fit for the data .

Test Statistic :

If Fcal > then we reject Ho.

As we reject Ho we conclude that the straight line model is not an approximate fit for the data.

Polynomial Model :

Statistical model that models the relationship between the response Y and the predictor X through a kth degree polynomial equation in X .

→ Quadratic

→ Cubic

Now if we get there is a lack of fit in the model. We now add one more variable i.e. X2

Once we add X2 . Find regression model

Now we do the test for significance of X2 variable. (i.e. we test whether X2 is significant in the above model or not. )

We can take decision from P-value i.e.

If P-value > α , we accept Ho.

i.e. X2 is not significant or not necessary.

Test lack of fit for this above model i.e.

If Ho is accepted i.e. there is no lack of fit and given model is best fitted for the given data.

If in lack of fit test Ho is rejected i.e. there is lack of fit in the model then we add one more higher order term in the model i.e.

1. Investigate regression model and fit the same also test goodness of fit of the model.

|  |  |
| --- | --- |
| X | Y |
| 1.0 | 10.84 |
| 1.0 | 9.30 |
| 2.0 | 16.35 |
| 3.3 | 22.88 |
| 3.3 | 24.35 |
| 4.0 | 24.56 |
| 4.0 | 25.86 |
| 4.0 | 29.16 |
| 4.7 | 24.59 |
| 5.0 | 22.25 |
| 5.6 | 25.90 |
| 5.6 | 27.20 |
| 5.6 | 25.61 |
| 6.0 | 25.45 |
| 6.0 | 26.56 |
| 6.5 | 21.03 |
| 6.9 | 21.46 |

1. Perform a thorough analysis of the results including residuals plots.

As there is only one regressor in the regression model we first fit the simple linear regression model



From the scatter plot it can be seen that there is no linear relationship between the predictor and response variable.



From the above normal probability plot it can be seen that the observations come from the normal distribution.



A curved plot as above indicates violation of linearity assumption.

From the above graphs it can be seen that simple linear regression model is not appropriately fit to the given data.

Regression Analysis: Y versus X

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 237.48 237.479 14.24 0.002

X 1 237.48 237.479 14.24 0.002

Error 15 250.13 16.676

Lack-of-Fit 8 234.57 29.321 13.19 0.001

Pure Error 7 15.56 2.223

Total 16 487.61

Model Summary

S R-sq R-sq(adj) R-sq(pred)

4.08358 48.70% 45.28% 27.53%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 13.21 2.66 4.96 0.000

X 2.130 0.565 3.77 0.002 1.00

Regression Equation

Y = 13.21 + 2.130 X

1. Perform the appropriate test for lack of fit.

Model Summary

S R-sq R-sq(adj) PRESS R-sq(pred)

4.08358 48.70% 45.28% 353.384 27.53%

Coefficients

Term Coef SE Coef 95% CI T-Value P-Value VIF

Constant 13.21 2.66 ( 7.53, 18.89) 4.96 0.000

X 2.130 0.565 (0.927, 3.334) 3.77 0.002 1.00

Regression Equation

Y = 13.21 + 2.130 X

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 1 237.48 48.70% 237.48 237.479 14.24 0.002

X 1 237.48 48.70% 237.48 237.479 14.24 0.002

Error 15 250.13 51.30% 250.13 16.676

Lack-of-Fit 8 234.57 48.11% 234.57 29.321 13.19 0.001

Pure Error 7 15.56 3.19% 15.56 2.223

Total 16 487.61 100.00%

To test the significance of simple linear regression model.

We have fitted a simple linear regression to the given data

Now we check lack of fit test

To test,

Ho : Y = 0 + 1 X + is an appropriate fit for the data

Vs

H1 : Y = 0 + 1 X + is not an appropriate fit for the data

Test Statistics –

F =

Where ,

n = Total number of observations = 17

d = Number of distinct predictor values at which observations were taken = 10

k = Number of regressors = 1

Thus d-k-1 = 8 and n-d = 17-10 = 7

Also from ANOVA table degrees of freedom of lack of fit = d-k-1 = 8 and degrees of freedom of pure error = 7

From ANOVA table :

SSLF = 234.57

SSPE = 15.56

Thus,

MSLF = = = 29.32125

Also from the ANOVA table mean square of lack of fit = 29.321

MSPE = = = 2.2228

Also from the ANOVA table mean square of pure error = 2.223

Therefore,

F = = = = 13.191132

Also from ANOVA table

F calculated = 13.19

F- table:

Inverse Cumulative Distribution Function

F distribution with 8 DF in numerator and 7 DF in denominator

P( X ≤ x ) x

0.95 3.72573

F – Table value = 3.72573

Thus F – Calculated > F – Table. Therefore, Reject Ho at 5% L.O.S.

Therefore, Y = 0 + 1 X + is not an appropriate fit for the data.

Simple linear regression model is not appropriate as we can observe that there have been multiple responses corresponding to the same predictor values.

As there is lack of fit in the model we now add one more higher order variable to the model.

Fitting a Quadratic model by adding higher order variable X2

Now the regression model is as follows

Regression Analysis: Y versus X, X2

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 440.40 220.201 65.30 0.000

X 1 299.26 299.259 88.74 0.000

X2 1 202.92 202.924 60.18 0.000

Error 14 47.21 3.372

Lack-of-Fit 7 31.65 4.521 2.03 0.185

Pure Error 7 15.56 2.223

Total 16 487.61

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.83634 90.32% 88.93% 86.93%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 0.18 2.06 0.08 0.934

X 10.82 1.15 9.42 0.000 20.47

X2 -1.124 0.145 -7.76 0.000 20.47

Regression Equation

Y = 0.18 + 10.82 X - 1.124 X2

Fits and Diagnostics for Unusual Observations

Obs Y Fit Resid Std Resid

8 29.160 25.473 3.687 2.15 R

10 22.250 26.179 -3.929 -2.25 R

R Large residual

Regression Equation

Y = 0.18 + 10.82 X - 1.124 X^2

Check whether x2 is significant

To test,

HO  : β2 = 0 Vs H1 : β2 ≠ 0

Test Statistics

From Coefficients table we observe that

t – Calculated = = = -7.7517

Inverse Cumulative Distribution Function

Student’s t distribution with 14 DF

P( X ≤ x ) x

0.95 1.76131

t- table = 1.76131

Here, |t-calculated| > t-table

Also, p-value = 0.000 < α =0.05

Therefore, we reject HO at 5% L.O.S.

Therefore x2 is significant for the model

Testing of the significance of the Quadratic model:

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 440.40 90.32% 440.40 220.201 65.30 0.000

X 1 237.48 48.70% 299.26 299.259 88.74 0.000

X^2 1 202.92 41.62% 202.92 202.924 60.18 0.000

Error 14 47.21 9.68% 47.21 3.372

Lack-of-Fit 7 31.65 6.49% 31.65 4.521 2.03 0.185

Pure Error 7 15.56 3.19% 15.56 2.223

Total 16 487.61 100.00%

To test,

Ho : Y = 0 + 1 X + β2 X2 + is an appropriate fit for the data

Vs

H1 : Y = 0 + 1 X + β2 X2 + is not an appropriate fit for the data

Test Statistics –

F =

Where ,

n = Total number of observations = 17

d = Number of distinct predictor values at which observations were taken = 10

k = Number of regressors = 2

Thus d-k-1 = 7 and n-d = 17-10 = 7

Also from ANOVA table degrees of freedom of lack of fit = d-k-1 = 7 and degrees of freedom of pure error = 7

From ANOVA table :

SSLF = 31.65

SSPE = 15.56

Thus,

MSLF = = = 4.5214

Also from the ANOVA table mean square of lack of fit = 4.521

MSPE = = = 2.2228

Also from the ANOVA table mean square of pure error = 2.223

Therefore,

F = = = = 2.0341

Also from ANOVA table

F calculated = 2.03

F- table :

Inverse Cumulative Distribution Function

F distribution with 7 DF in numerator and 7 DF in denominator

P( X ≤ x ) x

0.95 3.78704

F – Table value = 3.78704

Thus F – Calculated < F – Table . Therefore, Accept Ho at 5% L.O.S.

Therefore, Y = 0 + 1 X + β2 X2 + is an appropriate fit for the data.

Therefore, current model is best fit for the data.



From the above normal probability plot it can be seen that the observations come from the normal distribution.



The horizontal band indicates that the data points are evenly distributed which indicates that the model is adequate.

1. Find E (M SPE) and E(M SLOF) and comment on lack of fit.

E (M SPE) =

From ANOVA table

E (M SPE) = = 2.223

Also,

E(M SLOF) = = 4.521

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 1 237.48 48.70% 237.48 237.479 14.24 0.002

X 1 237.48 48.70% 237.48 237.479 14.24 0.002

Error 15 250.13 51.30% 250.13 16.676

Lack-of-Fit 8234.5748.11% 234.57 29.321 13.19 0.001

Pure Error 715.56 3.19% 15.56 2.223

Total 16 487.61 100.00%

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 440.40 90.32% 440.40 220.201 65.30 0.000

X 1 237.48 48.70% 299.26 299.259 88.74 0.000

X^2 1 202.92 41.62% 202.92 202.924 60.18 0.000

Error 14 47.21 9.68% 47.21 3.372

Lack-of-Fit731.656.49% 31.65 4.521 2.03 0.185

Pure Error715.56 3.19% 15.56 2.223

Total 16 487.61 100.00%

Conclusion for the lack of fit :

* The degrees of freedom of the lack of fit decreases by 1 when we fit the Quadratic model as it gets added to the higher order regressor.
* As we fit the Polynomial regression model to the data the lack of fit decreases and the adequacy of the model increases but the pure error remains the same and cannot be removed the model as pure error means variability that cannot be explained by any regression model.
* As the lack of fit decreases after fitting the Quadratic model the contribution of the regression model in to the given data increases from 48.7% to 90.32 %.
* The strength of the regression model can also be observed from the coefficient of determination R2 which increases from 0.487 to 0.889.

1. On an appropriate measure of the whiteness of rayon(y) the engineers conducting this experiment wish to minimize this measure. The experiment results given as follows:

|  |  |
| --- | --- |
| X | Y |
| 75 | 60 |
| 100 | 112 |
| 100 | 136 |
| 125 | 160 |
| 125 | 150 |
| 150 | 152 |
| 175 | 156 |
| 175 | 124 |
| 200 | 124 |
| 200 | 104 |
| 75 | 85 |
| 75 | 88 |
| 100 | 97 |
| 100 | 100 |
| 125 | 142 |
| 125 | 150 |
| 150 | 140 |

1. Perform a thorough analysis of the results including residuals plots.

As there is only one regressor in the regression model we first fit the simple linear regression model



From the scatter plot it can be seen that there is no linear relationship between the predictor and response variable



From the above normal probability plot it can be seen that the observations come from the normal distribution.



A curved plot as above indicates violation of linearity assumption.

From the above graphs it can be seen that simple linear regression model is not an appropriate fit to the given data.

1. Perform the appropriate test for lack of fit.

Model Summary

S R-sq R-sq(adj) PRESS R-sq(pred)

26.5723 23.78% 18.70% 14474.7 0.00%

Coefficients

Term Coef SE Coef 95% CI T-Value P-Value VIF

Constant 77.9 21.5 ( 32.0, 123.8) 3.62 0.003

X 0.348 0.161 (0.005, 0.690) 2.16 0.047 1.00

Regression Equation

Y = 77.9 + 0.348 X

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 1 3305 23.78% 3305 3304.6 4.68 0.047

X 1 3305 23.78% 3305 3304.6 4.68 0.047

Error 15 10591 76.22% 10591 706.1

Lack-of-Fit 4 8229 59.22% 8229 2057.2 9.58 0.001

Pure Error 11 2362 17.00% 2362 214.8

Total 16 13896 100.00%

To test the significance of simple linear regression model

We have fitted a simple linear regression to the given data

Now we check lack of fit test

To test,

Ho : Y = 0 + 1 X + is an appropriate fit for the data

Vs

H1 : Y = 0 + 1 X + is not an appropriate fit for the data

Test Statistics –

F =

Where ,

n = Total number of observations = 17

d = Number of distinct predictor values at which observations were taken = 6

k = Number of regressors = 1

Thus d-k-1 = 4 and n-d = 17-6 = 11

Also from ANOVA table degrees of freedom of lack of fit = d-k-1 = 8 and degrees of freedom of pure error = 7

From ANOVA table :

SSLF = 8229

SSPE = 2362

Thus,

MSLF = = = 2057.25

Also from the ANOVA table mean square of lack of fit = 2057.2

MSPE = = = 214.72727

Also from the ANOVA table mean square of pure error = 214.8

Therefore,

F = = = = 9.58075

Also from ANOVA table

F calculated = 9.58

F- table :

Inverse Cumulative Distribution Function

F distribution with 4 DF in numerator and 11 DF in denominator

P( X ≤ x ) x

0.95 3.35669

F – Table value = 3.35669

Thus F – Calculated > F – Table . Therefore, Reject Ho at 5% L.O.S.

Therefore, Y = 0 + 1 X + is not an appropriate fit for the data.

Simple linear regression model is not appropriate as we can observe that there have been multiple responses corresponding to the same predictor values.

As there is lack of fit in the model we now add one more higher order variable to the model.

Fitting a Quadratic model by adding higher order variable X2

Now the regression model is as follows

Regression Analysis: Y versus X, X2

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 11071.6 5535.8 27.44 0.000

X 1 9061.0 9061.0 44.92 0.000

X2 1 7767.1 7767.1 38.50 0.000

Error 14 2824.2 201.7

Lack-of-Fit 3 461.8 153.9 0.72 0.562

Pure Error 11 2362.4 214.8

Total 16 13895.9

Model Summary

S R-sq R-sq(adj) R-sq(pred)

14.2032 79.68% 76.77% 69.93%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -160.0 40.0 -4.00 0.001

X 4.192 0.625 6.70 0.000 53.04

X2 -0.01412 0.00228 -6.20 0.000 53.04

Regression Equation

Y = -160.0 + 4.192 X - 0.01412 X2

Regression Equation

Y = -160.0 + 4.192 X - 0.01412 X^2

Check whether x2 is significant

To test,

HO  : β2 = 0 Vs H1 : β2 ≠ 0

Test Statistics

From Coefficients table we observe that

t – Calculated = = = -6.192

Inverse Cumulative Distribution Function

Student’s t distribution with 14 DF

P( X ≤ x ) x

0.95 1.76131

t- table = 1.76131

Here, |t-calculated| > t-table

Also, p-value = 0.000 < α =0.05

Therefore, we reject HO at 5% L.O.S.

Therefore x2 is significant for the model

Testing of the significance of the Quadratic model:

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 11071.6 79.68% 11071.6 5535.8 27.44 0.000

X 1 3304.6 23.78% 9061.0 9061.0 44.92 0.000

X^2 1 7767.1 55.89% 7767.1 7767.1 38.50 0.000

Error 14 2824.2 20.32% 2824.2 201.7

Lack-of-Fit 3 461.8 3.32% 461.8 153.9 0.72 0.562

Pure Error 11 2362.4 17.00% 2362.4 214.8

Total 16 13895.9 100.00%

To test,

Ho : Y = 0 + 1 X + β2 X2 + is an appropriate fit for the data

Vs

H1 : Y = 0 + 1 X + β2 X2 + is not an appropriate fit for the data

Test Statistics –

F =

Where ,

n = Total number of observations = 17

d = Number of distinct predictor values at which observations were taken = 6

k = Number of regressors = 2

Thus d-k-1 = 3 and n-d = 17-6 = 11

Also from ANOVA table degrees of freedom of lack of fit = d-k-1 = 3 and degrees of freedom of pure error = 11

From ANOVA table :

SSLF = 461.8

SSPE = 2362.4

Thus,

MSLF = = = 153.933

Also from the ANOVA table mean square of lack of fit = 153.93

MSPE = = = 214.7636

Also from the ANOVA table mean square of pure error = 214.8

Therefore,

F = = = = 0.716755

Also from ANOVA table

F calculated = 0.72

F- table :

Inverse Cumulative Distribution Function

F distribution with 3 DF in numerator and 11 DF in denominator

P( X ≤ x ) x

0.95 3.58743

F – Table value = 3.58743

Thus F – Calculated < F – Table . Therefore, Accept Ho at 5% L.O.S.

Therefore, Y = 0 + 1 X + β2 X2 + is an appropriate fit for the data.

Therefore, current model is best fit for the data.



From the above normal probability plot it can be seen that the observations come from the normal distribution.



The horizontal band indicates that the data points are evenly distributed which indicates that the model is adequate.

1. Find E (M SPE) and E(M SLOF) and comment on lack of fit.

E (M SPE) = = 214.72727

From ANOVA table

E (M SPE) = = 214.8

Also,

E(M SLOF) = = 2057.25

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 1 3305 23.78% 3305 3304.6 4.68 0.047

X 1 3305 23.78% 3305 3304.6 4.68 0.047

Error 15 10591 76.22% 10591 706.1

Lack-of-Fit4822959.22% 8229 2057.2 9.58 0.001

Pure Error112362 17.00% 2362 214.8

Total 16 13896 100.00%

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 11071.6 79.68% 11071.6 5535.8 27.44 0.000

X 1 3304.6 23.78% 9061.0 9061.0 44.92 0.000

X^2 1 7767.1 55.89% 7767.1 7767.1 38.50 0.000

Error 14 2824.2 20.32% 2824.2 201.7

Lack-of-Fit3461.83.32% 461.8 153.9 0.72 0.562

Pure Error11 2362.4 17.00% 2362.4 214.8

Total 16 13895.9 100.00%

The estimate of ,

E (M SPE) = 153.9

E(M SLOF) = 214.8

Conclusion for the lack of fit :

* The degrees of freedom of the lack of fit decreases by 1 when we fit the Quadratic model as it gets added to the higher order regressor.
* As we fit the Polynomial regression model to the data the lack of fit decreases and the adequacy of the model increases but the pure error remains the same and cannot be removed the model as pure error means variability that cannot be explained by any regression model.
* As the lack of fit decreases after fitting the Quadratic model the contribution of the regression model in to the given data increases from 23.78% to 79.68 %.
* The strength of the regression model can also be observed from the coefficient of determination R2 which increases from 0.187 to 0.768.