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

**DEPARTMENT OF STATISTICS.**

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

**ST 28**

**EXPT.NO. 01**

**Title : Simple Regression and Regression Diagnostic - I**

Let Y denote the pounds of steam used per month and X denote the average atmospheric temperature in degrees Fahrenheit. The corresponding 25 pair of observations are given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Y** | **X** | **No.** | **Y** | **X** |
| **1** | **10.98** | **35.3** | **2** | **11.13** | **29.7** |
| **3** | **12.51** | **30.8** | **4** | **8.40** | **58.8** |
| **5** | **9.27** | **61.4** | **6** | **8.73** | **71.3** |
| **7** | **6.36** | **74.4** | **8** | **8.50** | **76.7** |
| **9** | **7.82** | **70.7** | **10** | **9.14** | **57.5** |
| **11** | **8.24** | **46.4** | **12** | **12.19** | **28.9** |
| **13** | **11.88** | **28.1** | **14** | **9.57** | **39.1** |
| **15** | **10.94** | **46.8** | **16** | **9.58** | **48.5** |
| **17** | **10.09** | **59.3** | **18** | **8.11** | **70.0** |
| **19** | **6.83** | **70.0** | **20** | **8.88** | **74.5** |
| **21** | **7.68** | **72.1** | **22** | **8.47** | **58.1** |
| **23** | **8.86** | **44.6** | **24** | **10.36** | **33.4** |
| **25** | **11.08** | **28.6** |  |  |  |

1. Plot the scatter diagram of (X,Y).
2. Fit the linear regression line of Y on X .
3. Obtain Y : Predicted value of Y for given X
4. Obtain the residuals ei = Yi – ŷi , Plot them along the fitted straight line.
5. Obtain coefficient of determination R2
6. Obtain confidence interval for parameter.
7. Test the hypothesis regarding the parameter.

(Test of significance of regression, β1=0, β0 = 0)

1. Write the ANOVA table. Obtain the estimate of σ2.

Q2. For the following data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Obs. | x | Y11 | Y12 | Y13 | X’ | Y’ |
| 1 | 10 | 8.04 | 9.14 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8.14 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 8.74 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 8.77 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 9.26 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 8.10 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6.13 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 3.10 | 5.39 | 19 | 12.50 |
| 9 | 12 | 10.84 | 9.13 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7.26 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 4.74 | 5.73 | 8 | 6.89 |

1. Fit the linear regression line of Y11 on X, Y12 on X , Y13 on X and Y’ on X’.
2. Draw the scatter diagram for each pair of (X, Y) with the fitted line and comment on it.

THEORY

A model with a single regressor x that has a relationship with a response y that is a straight line.

Simple linear regression model

𝛽0 and 𝛽1 are regression coefficients.

Residuals (ei) =

Unbiased estimator of σ2

* Confidence Intervals on 𝛽0 , 𝛽1

1. For 𝛽0
2. For 𝛽1

]

Coefficient of determination

R2 =

Test on slope (𝛽1)

To test :

Ho : 𝛽1 = 0 VS H1 : 𝛽1 ≠ 0

(σ is unknown test statistic → Normal distribution

σ is known test statistic → t distribution)

Test statistic :

Under Ho

i.e.

Test criterion :

We reject Ho if || > then reject Ho at α % l.o.s.

P-value criterion

We reject Ho, if p-value < α

Test for intercept (𝛽0)

To test :

Ho : 𝛽0 = 0 VS H1 : 𝛽0 ≠ 0

Test statistic :

Test criterion :

We reject Ho if || > then reject Ho at α % l.o.s.

P-value criterion

We reject Ho, if p-value < α

Q1

1. Plot the scatter diagram of (X,Y)



**Interpretation:** The scatter plot above shows that there is a linear relationship between steam and atmospheric temperature , also they are negatively correlated to each other.

1. Fit the linear regression line of Y on X .

Regression Equation

y = 13.623 - 0.0798 x

1. Obtain Y : Predicted value of Y for given X

Regression Equation

y = 13.623 - 0.0798 x

|  |  |  |
| --- | --- | --- |
| **sr. no.** | **x** | **FITS1=estimated Y= 13.6229+-0.0798\*x** |
| 1 | 35.3 | 10.80503639 |
| 2 | 30.8 | 11.16426551 |
| 3 | 61.4 | 8.721507499 |
| 4 | 74.4 | 7.683734486 |
| 5 | 70.7 | 7.979100651 |
| 6 | 46.4 | 9.918937899 |
| 7 | 28.1 | 11.37980299 |
| 8 | 46.8 | 9.887006421 |
| 9 | 59.3 | 8.889147755 |
| 10 | 70 | 8.034980736 |
| 11 | 72.1 | 7.86734048 |
| 12 | 44.6 | 10.06262955 |
| 13 | 28.6 | 11.33988864 |
| 14 | 29.7 | 11.25207708 |
| 15 | 58.8 | 8.929062101 |
| 16 | 71.3 | 7.931203435 |
| 17 | 76.7 | 7.500128491 |
| 18 | 57.5 | 9.032839403 |
| 19 | 28.9 | 11.31594003 |
| 20 | 39.1 | 10.50168736 |
| 21 | 48.5 | 9.751297643 |
| 22 | 70 | 8.034980736 |
| 23 | 74.5 | 7.675751616 |
| 24 | 58.1 | 8.984942187 |
| 25 | 33.4 | 10.95671091 |

1. Obtain the residuals ei = Yi – ŷi , Plot them along the fitted straight line.



|  |  |  |
| --- | --- | --- |
| **sr.no.** | **FITS1=estimated Y= 13.6229+-0.0798\*x** | **RESI1** |
| 1 | 10.80503639 | 0.174964 |
| 2 | 11.16426551 | 1.345734 |
| 3 | 8.721507499 | 0.548493 |
| 4 | 7.683734486 | -1.32373 |
| 5 | 7.979100651 | -0.1591 |
| 6 | 9.918937899 | -1.67894 |
| 7 | 11.37980299 | 0.500197 |
| 8 | 9.887006421 | 1.052994 |
| 9 | 8.889147755 | 1.200852 |
| 10 | 8.034980736 | -1.20498 |
| 11 | 7.86734048 | -0.18734 |
| 12 | 10.06262955 | -1.20263 |
| 13 | 11.33988864 | -0.25989 |
| 14 | 11.25207708 | -0.12208 |
| 15 | 8.929062101 | -0.52906 |
| 16 | 7.931203435 | 0.798797 |
| 17 | 7.500128491 | 0.999872 |
| 18 | 9.032839403 | 0.107161 |
| 19 | 11.31594003 | 0.87406 |
| 20 | 10.50168736 | -0.93169 |
| 21 | 9.751297643 | -0.1713 |
| 22 | 8.034980736 | 0.075019 |
| 23 | 7.675751616 | 1.204248 |
| 24 | 8.984942187 | -0.51494 |
| 25 | 10.95671091 | -0.59671 |

**Interpretation:** The horizontal band indicates that model is adequate and there is no trend in the data.

1. Obtain coefficient of determination R2

**Regression Equation**

y = 13.623 - 0.0798 x

**Model Summary**

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

0.890125 71.44% 70.20% 21.4938 66.32%

**Analysis of Variance**

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

Regression 1 45.5924 71.44% 45.5924 45.5924 57.54 0.000

x 1 45.5924 71.44% 45.5924 45.5924 57.54 0.000

Error 23 18.2234 28.56% 18.2234 0.7923

Lack-of-Fit 22 17.4042 27.27% 17.4042 0.7911 0.97 0.680

Pure Error 1 0.8192 1.28% 0.8192 0.8192

Total 24 63.8158 100.00%

**From ANOVA table**

**SSR**= 45.5924

**SST**= 63.8158

**R^2** = SSR/SST = 0.71443749

**Interpretation:** Coefficient of determination is 71.4% . That means Atmospheric temperature is moderately significant and 71.44% variability in the amount of the steam is explained by the proposed model.

28.6% variability are not included in the therefore it is not included in our study .

1. Obtain confidence interval for parameter.

**Coefficients**

Term Coef SE Coef T-Value P-Value VIF

Constant 13.623 0.581 23.43 0.000

X -0.0798 0.0105 -7.59 0.000 1.00

**Inverse Cumulative Distribution Function**

Student’s t distribution with 23 DF

P( X ≤ x ) x

0.975 2.06866

**For β0,**

{β0^ - t n-2, α/2 \* S.E (β0^), β0^ + t n-2, α/2 \* S.E (β0^)}

{13.623 – (2.06866\*0.581), 13.623 + (2.06866\*0.581)}

**{12.42110854, 14.82489}**

**For β1,**

{β1^ - t n-2, α/2 \* S.E (β1^), β1^ + t n-2, α/2 \* S.E (β1^)}

{-0.0798 – (2.06866\*0.0105), -0.0798+ (2.06866\*0.0105)}

**{-0.10152093, -0.05807907}**

1. Test the hypothesis regarding the parameter.

(Test of significance of regression, β1=0, β0 = 0)

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 13.623 0.581 23.43 0.000

X -0.0798 0.0105 -7.59 0.000 1.00

**Inverse Cumulative Distribution Function**

Student’s t distribution with 23 DF

P( X ≤ x ) x

0.975 2.06866

**Hypothesis:**

For β0,

H0: β0 = 0

V/s

H1: β0 ≠ 0

**Test Statistics**

t Cal = β0 ^/ S.E (β0^) = 23.43

Table value = t n-2,α/2 = 2.06866

**Test Procedure**

We reject H0, if | t Cal | > t n-2,α/2 o.w. we accept it

Here | t Cal | = 23.43 > t n-2,α/2 = 2.06866 hence we reject H0.

**P-value criteria**

If p-value < α we reject H0

Here, p-value = 0.000 < 0.05, Hence we reject H0.

**Hypothesis:**

For β1,

Ho: β1= 0

V/s

H1: β1≠ 0

**Test Statistics**

t Cal = β0 ^/ S.E (β0^) = -7.59

Table value = t n-2,α/2 = 2.06866

**Test Procedure**

We reject H0, if | t Cal | > t n-2,α/2 o.w. we accept it

Here | t Cal | = 7.59 > t n-2,α/2 = 2.06866 hence we reject H0.

**P-value criteria**

If p-value < α we reject H0

Here, p-value = 0.000 < 0.05, Hence we reject H0

1. Write the ANOVA table. Obtain the estimate of σ2.

Regression Analysis: Y versus X

**The regression equation is**

Y = 13.62 - 0.07983 X

S = 0.890125 R-Sq = 71.4% R-Sq(adj) = 70.2%

**Analysis of Variance**

Source DF SS MS F P

Regression 1 45.5924 45.5924 57.54 0.000

Error 23 18.2234 0.7923

Total 24 63.8158

**σ2 = SSres / n-2 = 18.2234 / 23 = 0.7923**

**Regression Analysis: Y versus X**

**Analysis of Variance**

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

Regression 1 45.5924 45.5924 57.54 0.000

X 1 45.5924 45.5924 57.54 0.000

Error 23 18.2234 0.7923

Lack-of-Fit 22 17.4042 0.7911 0.97 0.680

Pure Error 1 0.8192 0.8192

Total 24 63.8158

**Model Summary**

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

0.890125 71.44% 70.20% 66.32%

**Q2**

1. Draw the scatter diagram for each pair of (X, Y) with the fitted line and comment on it.



**Interpretation**: The data points are spread in the upward direction. This means there is moderate positive correlation.



**Interpretation**: The data points are spread in the upward direction. This means there is positive and linear correlation.



**Interpretation**: The data points are spread in the upward direction. This means there is perfect positive correlation.



**Interpretation**: Data points are not showing any trend that means there is no correlation

1. Fit the linear regression line of Y11 on X, Y12 on X , Y13 on X and Y’ on X’.

**Y11=3+0.5001x**

**Y12=3.001+0.5000x**

**Y13=3.002+0.4997x**

**Y’=3.002+0.4999X**