**Department of Statistics,**

**Modern college of Arts, Science and Commerce, Pune-05**

**M.Sc. I (Statistics) Semester II**

**ST-28 Date:**

**Practical No. 6 Submission date:**

**Title: Multicollinearity and polynomial regression model**

Q.1.The carbonation level of a soft drink beverage is affected by the temperature of the product and the filter operating pressure. Twelve observations were obtained and the resulting data are shown below.

|  |  |  |
| --- | --- | --- |
| Carbonation (y) | Temperature (x1) | Pressure (x2) |
| 2.60 | 31.0 | 21.0 |
| 2.40 | 31.0 | 21.0 |
| 17.32 | 31.5 | 24.0 |
| 15.60 | 31.5 | 24.0 |
| 16.12 | 31.5 | 24.0 |
| 5.36 | 30.5 | 22.0 |
| 6.19 | 31.5 | 22.0 |
| 10.17 | 30.5 | 23.0 |
| 2.62 | 31.0 | 21.5 |
| 2.98 | 30.5 | 21.5 |
| 6.92 | 31.0 | 22.5 |
| 7.06 | 30.5 | 22.5 |

1. Fit a second – order polynomial.
2. Test for significance of regression.
3. Test for interaction term contribute significantly to the model?
4. Do the second- order terms contribute significantly to the model?

Q.2 A finished product is known to lose weight after it is produced. The following data demonstrate this drop in weight.

|  |  |
| --- | --- |
| Time after  Production , t | Weight difference  (in 1/ 16oz), Y |
| 0 | 0.21 |
| 0.5 | -1.46 |
| 1.0 | -3.04 |
| 1.5 | -3.21 |
| 2.0 | -5.04 |
| 2.5 | -5.37 |
| 3.0 | -6.03 |
| 3.5 | -7.21 |
| 4.0 | -7.46 |
| 4.5 | -7.96 |

1. Using orthogonal polynomials, develop a second order fitted equation that represents the loss in weights as a function of time after production.
2. Analyze the residuals from the model and draw concretions about its adequacy.

Q3. Fit a cubic spline regression model with a knot at X=6.5, 13



For the voltage drop data given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs. | time(Xi) | Voltage Drop(yi) | Obs. | time(Xi) | Voltage Drop(yi) |
| 1 | 0 | 8.33 | 21 | 10 | 14.48 |
| 2 | 0.5 | 8.23 | 22 | 10.5 | 14.92 |
| 3 | 1 | 7.17 | 23 | 11 | 14.37 |
| 4 | 1.5 | 7.14 | 24 | 11.5 | 14.63 |
| 5 | 2 | 7.31  7.60 | 25 | 12 | 15.18  14.51 |
| 6 | 2.5 | 7.60 | 26 | 12.5 | 14.51 |
| 7 | 3 | 7.94 | 27 | 13 | 14.34 |
| 8 | 3.5 | 8.30 | 28 | 13.5 | 13.81 |
| 9 | 4 | 8.76 | 29 | 14 | 13.79 |
| 10 | 4.5 | 8.71 | 30 | 14.5 | 13.05 |
| 11 | 5 | 9.71 | 31 | 15 | 13.04 |
| 12 | 5.5 | 10.26 | 32 | 15.5 | 12.60 |
| 13 | 6 | 10.91 | 33 | 16 | 12.05 |
| 14 | 6.5 | 11.67 | 34 | 16.5 | 11.15 |
| 15 | 7 | 11.76 | 35 | 17 | 11.15 |
| 16 | 7.5 | 12.81 | 36 | 17.5 | 10.14 |
| 17 | 8 | 13.30 | 37 | 18 | 10.08 |
| 18 | 8.5 | 13.88 | 38 | 18.5 | 9.78 |
| 19 | 9 | 14.59 | 39 | 19 | 9.80 |
| 20 | 9.5 | 14.05 | 40 | 19.5 | 9.95 |
|  |  |  | 41 | 20 | 9.51 |

Q4. The data given below concerning the percentage of conversion on n heptanes

To acetylene and three explanatory variables .These are typical chemical

Process data for which a full quadratic response surface in all three repressors

We wanted to fit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observation | Conversion of n-heptane to Acetylene (%) | Reactor temperature(o C) | Ratio of H2 to n-Heptane | Contact time |
| 1 | 49.0 | 1300 | 7.5 | 0.0120 |
| 2 | 50.2 | 1300 | 9.0 | 0.0120 |
| 3 | 50.5 | 1300 | 11.0 | 0.0115 |
| 4 | 48.5 | 1300 | 13.5 | 0.0130 |
| 5 | 47.5 | 1300 | 17.0 | 0.0135 |
| 6 | 44.5 | 1300 | 23.0 | 0.0120 |
| 7 | 28 | 1200 | 5.3 | 0.0400 |
| 8 | 31.5 | 1200 | 7.5 | 0.0380 |
| 9 | 34.5 | 1200 | 11.0 | 0.0320 |
| 10 | 35 | 1200 | 13.5 | 0.0260 |
| 11 | 38 | 1200 | 17.0 | 0.0340 |
| 12 | 38.5 | 1200 | 23.0 | 0.0410 |
| 13 | 15 | 1100 | 5.3 | 0.0840 |
| 14 | 17 | 1100 | 7.5 | 0.0980 |
| 15 | 20.5 | 1100 | 11.0 | 0.0920 |
| 16 | 29.5 | 1100 | 17.0 | 0.0860 |

1. From the matrix of correlation between the repressors would you suspect that Multicollinearity is present?
2. Calculate Variance inflation factors.
3. Using orthogonal polynomials fit a quadratic surface model

Q.1)

Regression Analysis: y versus x1, x2, x1^2, x2^2, x1x2

Analysis of Variance

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

Regression 5 339.888 67.9775 177.17 0.000

x1 1 0.830 0.8303 2.16 0.192

x2 1 0.033 0.0331 0.09 0.779

x1^2 1 1.033 1.0330 2.69 0.152

x2^2 1 4.878 4.8778 12.71 0.012

x1x2 1 0.847 0.8472 2.21 0.188

Error 6 2.302 0.3837

Lack-of-Fit 3 0.726 0.2420 0.46 0.730

Pure Error 3 1.576 0.5254

Total 11 342.190

Model Summary

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

0.619430 99.33% 98.77% 94.76%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3025 2046 1.48 0.190

x1 -194 132 -1.47 0.192 90910.93

x2 -6.1 20.6 -0.29 0.779 15400.99

x1^2 3.63 2.21 1.64 0.152 97844.72

x2^2 1.154 0.324 3.57 0.012 7759.76

x1x2 -1.332 0.896 -1.49 0.188 37872.36

Regression Equation

y = 3025 - 194 x1 - 6.1 x2 + 3.63 x1^2 + 1.154 x2^2 - 1.332 x1x2

a. Second order polynomial model is

Y=β0 + β1x1 + β2x2 + β11x12 + β22x22 + β12 x1x2 + ε

Regression Equation

y = 3025 - 194 x1 - 6.1 x2 + 3.63 x1^2 + 1.154 x2^2 - 1.332 x1x2

b. Test for significance of regression.

Analysis of Variance

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

Regression 5 339.888 67.9775 177.17 0.000

x1 1 0.830 0.8303 2.16 0.192

x2 1 0.033 0.0331 0.09 0.779

x1^2 1 1.033 1.0330 2.69 0.152

x2^2 1 4.878 4.8778 12.71 0.012

x1x2 1 0.847 0.8472 2.21 0.188

Error 6 2.302 0.3837

Lack-of-Fit 3 0.726 0.2420 0.46 0.730

Pure Error 3 1.576 0.5254

Total 11 342.190

Hypothesis:

For βi,

H0: βi= 0

V/s H1: βi≠ 0 for at least one i where i=1, 2, 3… k k=No. of regressors in the model

Test Statistics

From Anova table, SSR=339.888 SSE=2.302 k=5 p=k+1=6 n-p=12-6=6

F Cal = (SSR/k)/ (SSE/ (n-p)) = (339.888/5)/ (2.302/6)) = 67.9776/ 0.383666= 177.17

F Cal= 177.17

Inverse Cumulative Distribution Function

F distribution with 5 DF in numerator and 6 DF in denominator

P( X ≤ x ) x

0.95 4.38737

F tabulated = 4.38737

Test Procedure

We reject H0, if | F Cal | > F k, n-k-1, α/2  o.w. we accept it

Here | F Cal | = 177.17 > F k, n-k-1, α/2 = 4.38737 hence we reject H0.

I.e. there is at least one regressor variable is significant in the model.

P-value criteria

If p-value < α we reject H0

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

c. Test for interaction term contribute significantly to the model?

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3025 2046 1.48 0.190

x1 -194 132 -1.47 0.192 90910.93

x2 -6.1 20.6 -0.29 0.779 15400.99

x1^2 3.63 2.21 1.64 0.152 97844.72

x2^2 1.154 0.324 3.57 0.012 7759.76

x1x2 -1.332 0.896 -1.49 0.188 37872.36

Hypothesis

H0: βij = 0 v/s H1: βij ≠0

i.e H0: β12 = 0 v/s H1: β12 ≠0

Test Statistics

t Cal = β12 / S.E (β12) = -1.332/ 0.896 = -1.48660

Test Procedure

Inverse Cumulative Distribution Function

Student’s t distribution with 6 DF

P( X ≤ x ) x

0.975 2.44691

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

Here | t Cal | = 1.148660 < t n-p, α/2 = t 6, 0.975 = 2.44691 hence we accept H0. I.e. accepting H0 implies that interaction term does not contribute significantly to the model.

P-value criteria

If p-value < α we reject H0

Here, p-value = 0.188 > 0.05, Hence we accept H0.

d. Do the second- order terms contribute significantly to the model?

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3025 2046 1.48 0.190

x1 -194 132 -1.47 0.192 90910.93

x2 -6.1 20.6 -0.29 0.779 15400.99

x1^2 3.63 2.21 1.64 0.152 97844.72

x2^2 1.154 0.324 3.57 0.012 7759.76

x1x2 -1.332 0.896 -1.49 0.188 37872.36

Hypothesis

H0: βii = 0 v/s H1: βii ≠0

i)

H0: β11= 0 v/s H1: β11 ≠0

Test Statistics

t Cal = β11 / S.E (β11) = 3.63/2.21 = 1.6425

Test Procedure

Inverse Cumulative Distribution Function

Student’s t distribution with 6 DF

P( X ≤ x ) x

0.975 2.44691

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

Here | t Cal | = 1.6425 < t n-p, α/2 = t 6, 0.975 = 2.44691 hence we accept H0. I.e. accepting H0 implies that second order them (i.e. x12 does not contribute significantly to the model.

P-value criteria

If p-value < α we reject H0

Here, p-value = 0.152 > 0.05, Hence we accept H0.

ii)

H0: β22= 0 v/s H1: β22 ≠0

Test Statistics

t Cal = β22 / S.E (β22) = 1.154/0.324= 3.57

Test Procedure

Inverse Cumulative Distribution Function

Student’s t distribution with 6 DF

P( X ≤ x ) x

0.975 2.44691

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

Here | t Cal | = 3.57 > t n-p, α/2 = t 6, 0.975 = 2.44691 hence we reject H0. I.e. rejecting H0 implies that second order them (i.e. x22 contribute significantly to the model.

P-value criteria

If p-value < α we reject H0

Here, p-value = 0.012 < 0.05, Hence we reject H0.

Q.2)

Orthogonal polynomial model is,

Y=α0 P0 (xi) + α1P1 (xi) + α2P2 (xi) + ε

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time after X | (x-xbar/d) | P1(x)=ƛ1\*(x-xbar/d) | (x-xbar/d)^2-(n^2-1)/12 | P1(x)=ƛ2\*(x-xbar/d)^2-(n^2-1)/12 |
| 0 | -4.5 | -45 | 12 | 12 |
| 0.5 | -3.5 | -35 | 4 | 4 |
| 1 | -2.5 | -25 | -2 | -2 |
| 1.5 | -1.5 | -15 | -6 | -6 |
| 2 | -0.5 | -5 | -8 | -8 |
| 2.5 | 0.5 | 5 | -8 | -8 |
| 3 | 1.5 | 15 | -6 | -6 |
| 3.5 | 2.5 | 25 | -2 | -2 |
| 4 | 3.5 | 35 | 4 | 4 |
| 4.5 | 4.5 | 45 | 12 | 12 |

|  |  |
| --- | --- |
| X bar | 2.25 |
| d | 0.5 |
| (n^2-1)/12 | 8.25 |
| ƛ1 | 10 |
| ƛ2 | 1 |

X=

|  |  |  |
| --- | --- | --- |
| P0(x) | P1(x) | P2(x) |
| 1 | -45 | 12 |
| 1 | -35 | 4 |
| 1 | -25 | -2 |
| 1 | -15 | -6 |
| 1 | -5 | -8 |
| 1 | 5 | -8 |
| 1 | 15 | -6 |
| 1 | 25 | -2 |
| 1 | 35 | 4 |
| 1 | 45 | 12 |

Y=

|  |
| --- |
| 0.21 |
| -1.46 |
| -3.04 |
| -3.21 |
| -5.04 |
| -5.37 |
| -6.03 |
| -7.21 |
| -7.46 |
| -7.96 |

Minitab Outout

α= (X’X)-1 X’Y

Data Display

Matrix M8

-4.65700

-0.08798

0.05784

α= [α0 α1 α2]’

α= [-4.65700 -0.08798 0.05784]’

Therefore,

Orthogonal polynomial model is,

Y=-4.65700 P0 (xi) -0.08798P1 (xi) + 0.05784P2 (xi)

b. Analyze the residuals from the model and draw concretions about its adequacy.

Regression Analysis: Y\_1 versus x1, x2

Analysis of Variance

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

Regression 2 65.6281 32.8140 282.34 0.000

x1 1 63.8616 63.8616 549.48 0.000

x2 1 1.7665 1.7665 15.20 0.006

Error 7 0.8135 0.1162

Total 9 66.4416

Model Summary

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

0.340912 98.78% 98.43% 97.71%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -4.657 0.108 -43.20 0.000

x1 -0.08798 0.00375 -23.44 0.000 1.00

x2 0.0578 0.0148 3.90 0.006 1.00

Regression Equation

Y\_1 = -4.657 - 0.08798 x1 + 0.0578 x2

R2 adj= 98.43%

Interpretation: it shows that 98.43% of the variation in Y is explained by regressor X1, X2. It shows that 98.43% of the data fit the regression model.

Q.3)



For the voltage drop data given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs. | time(Xi) | Voltage Drop(yi) | Obs. | time(Xi) | Voltage Drop(yi) |
| 1 | 0 | 8.33 | 21 | 10 | 14.48 |
| 2 | 0.5 | 8.23 | 22 | 10.5 | 14.92 |
| 3 | 1 | 7.17 | 23 | 11 | 14.37 |
| 4 | 1.5 | 7.14 | 24 | 11.5 | 14.63 |
| 5 | 2 | 7.31  7.60 | 25 | 12 | 15.18  14.51 |
| 6 | 2.5 | 7.60 | 26 | 12.5 | 14.51 |
| 7 | 3 | 7.94 | 27 | 13 | 14.34 |
| 8 | 3.5 | 8.30 | 28 | 13.5 | 13.81 |
| 9 | 4 | 8.76 | 29 | 14 | 13.79 |
| 10 | 4.5 | 8.71 | 30 | 14.5 | 13.05 |
| 11 | 5 | 9.71 | 31 | 15 | 13.04 |
| 12 | 5.5 | 10.26 | 32 | 15.5 | 12.60 |
| 13 | 6 | 10.91 | 33 | 16 | 12.05 |
| 14 | 6.5 | 11.67 | 34 | 16.5 | 11.15 |
| 15 | 7 | 11.76 | 35 | 17 | 11.15 |
| 16 | 7.5 | 12.81 | 36 | 17.5 | 10.14 |
| 17 | 8 | 13.30 | 37 | 18 | 10.08 |
| 18 | 8.5 | 13.88 | 38 | 18.5 | 9.78 |
| 19 | 9 | 14.59 | 39 | 19 | 9.80 |
| 20 | 9.5 | 14.05 | 40 | 19.5 | 9.95 |
|  |  |  | 41 | 20 | 9.51 |

Regression Analysis: Y versus X, X^2, X^3, (X-6.5)^3, (X-13)^3

The following terms cannot be estimated and were removed:

(X-6.5)^3, (X-13)^3

Analysis of Variance

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

Regression 3 230.444 76.8148 88.14 0.000

X 1 7.873 7.8729 9.03 0.005

X^2 1 1.346 1.3456 1.54 0.222

X^3 1 11.788 11.7882 13.53 0.001

Error 37 32.244 0.8715

Total 40 262.689

Model Summary

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

0.933523 87.73% 86.73% 82.40%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 6.491 0.534 12.16 0.000

X 0.703 0.234 3.01 0.005 90.13

X^2 0.0340 0.0274 1.24 0.222 528.13

X^3 -0.003307 0.000899 -3.68 0.001 209.89

Regression Equation

Y = 6.491 + 0.703 X + 0.0340 X^2 - 0.003307 X^3



Q.4)

a. From the matrix of correlation between the repressors would you suspect that Multicollinearity is present?

Data Display

Matrix CORR1

1.00000 0.22363 -0.95820

0.22363 1.00000 -0.24023

-0.95820 -0.24023 1.00000

Correlation between x1 and x2 = 0.22363 < 0.75, we conclude that there is no multicollinearity.

Correlation between x1 and x3 = 0.95820 > 0.75, we conclude that there is multicollinearity.

b. Calculate Variance inflation factors.

Regression Analysis: Conversion of n- versus Reactor temperat, Ratio of H2 to n, Contact time

Analysis of Variance

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

Regression 3 1953.42 651.140 45.88 0.000

Reactor temperature(o C) 1 128.34 128.340 9.04 0.011

Ratio of H2 to n-Heptane 1 54.89 54.895 3.87 0.073

Contact time 1 0.44 0.441 0.03 0.863

Error 12 170.29 14.191

Total 15 2123.71

Model Summary

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

3.76707 91.98% 89.98% 84.16%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -121.3 55.4 -2.19 0.049

Reactor temperature(o C) 0.1269 0.0422 3.01 0.011 12.23

Ratio of H2 to n-Heptane 0.348 0.177 1.97 0.073 1.06

Contact time -19 108 -0.18 0.863 12.32

Regression Equation

Conversion of n-heptane to Acet = -121.3 + 0.1269 Reactor temperature(o C)

+ 0.348 Ratio of H2 to n-Heptane - 19 Contact time

For X1 variable i.e. Reactor temperature (0C) VIF is 12.23 it means X1 variable has very high correlation with other variables.

For X2 variable i.e. Ratio of H2 to n-Heptane VIF is 1.06 it means X1 variable is not correlated with other variables.

For X3 variable i.e. Contact time VIF is 12.32 it means X3 variable has very high correlation with other variables.