REGRESSION MODELS – FINAL PROJECT

Thanh Doan – Student ID 0159701

DATA SET

Thanh's data

BLE B.9 Pressure Drop Data

x_1	x_2	x_3	x_4	y
2.14	10	0.34	1	28.9
4.14	10	0.34	1	31
8.15	10	0.34	1	26.4
2.14	10	0.34	0.246	27.2
4.14	10	0.34	0.379	26.1
8.15	10	0.34	0.474	23.2
2.14	10	0.34	0.141	19.7
4.14	10	0.34	0.234	22.1
8.15	10	0.34	0.311	22.8
2.14	10	0.34	0.076	29.2
4.14	10	0.34	0.132	23.6
8.15	10	0.34	0.184	23.6
2.14	2.63	0.34	0.679	24.2
4.14	2.63	0.34	0.804	22.1
8.15	2.63	0.34	0.89	20.9
2.14	2.63	0.34	0.514	17.6
4.14	2.63	0.34	0.672	15.7
8.15	2.63	0.34	0.801	15.8
2.14	2.63	0.34	0.346	14
4.14	2.63	0.34	0.506	17.1
8.15	2.63	0.34	0.669	18.3
2.14	2.63	0.34	1	33.8
4.14	2.63	0.34	1	31.7
8.15	2.63	0.34	1	28.1
5.6	1.25	0.34	0.848	18.1
5.6	1.25	0.34	0.737	16.5
5.6	1.25	0.34	0.651	15.4
5.6	1.25	0.34	0.554	15
4.3	2.63	0.34	0.748	19.1
4.3	2.63	0.34	0.682	16.2
4.3	2.63	0.34	0.524	16.3
4.3	2.63	0.34	0.472	15.8
4.3	2.63	0.34	0.398	15.4
5.6	10.1	0.25	0.789	19.2
5.6	10.1	0.25	0.677	8.4
5.6	10.1	0.25	0.59	15
5.6	10.1	0.25	[illegible]	[illegible]

TABLE B.9 (*Continued*)

x_1	x_2	x_3	x_4	y
5.6	10.1	0.34	0.677	21.3
5.6	10.1	0.34	0.59	21.6
5.6	10.1	0.34	0.523	19.8
4.3	10.1	0.34	0.741	21.6
4.3	10.1	0.34	0.617	17.3
4.3	10.1	0.34	0.524	20
4.3	10.1	0.34	0.457	18.6
2.4	10.1	0.34	0.615	22.1
2.4	10.1	0.34	0.473	14.7
2.4	10.1	0.34	0.381	15.8
2.4	10.1	0.34	0.32	13.2
5.6	10.1	0.55	0.789	30.8
5.6	10.1	0.55	0.677	27.5
5.6	10.1	0.55	0.59	25.2
5.6	10.1	0.55	0.523	22.8
2.14	112	0.34	0.68	41.7
4.14	112	0.34	0.803	33.7
8.15	112	0.34	0.889	29.7
2.14	112	0.34	0.514	41.8
4.14	112	0.34	0.672	37.1
8.15	112	0.34	0.801	40.1
2.14	112	0.34	0.306	42.7
4.14	112	0.34	0.506	48.6
8.15	112	0.34	0.668	42.4

y: Dimensionless factor for the pressure drop through a bubble cap
x_1: Superficial fluid velocity of the gas (cm/s)
x_2: Kinematic viscosity
x_3: Mesh opening (cm) ?
x_4: Dimensionless number relating the superficial fluid velocity of the gas to the superficial fluid velocity of the liquid

Source: "A Correlation of Two-Phase Pressure Drops in Screen-plate Bubble Column," by C. H. Liu, M. Kan, and B. H. Chen, *Canadian Journal of Chemical Engineering*, **71**, 460–463.

Question:
Mesh opening? should consider it as an indicator variable?
or it is a numerical measurement?

Objective of the analysis

- Analyzing the Pressure Drop Dataset to study if the response variable y (pressure drop through a bubble cap) is dependent on some or all of the repressors x1, x2, x3, x4.

Assumption

- All variables (y, x1, x2, x3, and x4) are numerical measurements. None of the repressors are categorical/indicator variables

1. First I compute the variance/covariance of all study variables as well as Pearson Correlation to have some understanding about the variability of each variable and the pair-wise correlation between them. Though pair-wise correlation only tells us part of the story.

```
> var(PressureDrop)
             x1            x2            x3           x4           y
x1  3.960418429    0.17059648  0.0064181914 0.1445918614  -0.7893144
x2  0.170596483 1402.57076375 -0.0956332099 0.4607139979 255.4941198
x3  0.006418191   -0.09563321  0.0033620307 0.0003297938   0.1004120
x4  0.144591861    0.46071400  0.0003297938 0.0541005637   0.3429128
y  -0.789314384  255.49411978  0.1004119513 0.3429127975  76.1066552
> cor(PressureDrop)
             x1           x2          x3         x4           y
x1  1.000000000  0.002288954  0.05562132 0.31237223 -0.04546405
x2  0.002288954  1.000000000 -0.04403982 0.05288934  0.78200081
x3  0.055621320 -0.044039817  1.00000000 0.02445348  0.19850582
x4  0.312372229  0.052889337  0.02445348 1.00000000  0.16899408
y  -0.045464054  0.782000806  0.19850582 0.16899408  1.00000000
>
```

- **Comment:** The variance of x2 is very high. Most of the value of x2 are under 10 but then the last 9 data points (observation 54-62) has value 112 that makes the variability of x2 high.

- **Comment:** The variance of x3 is very low, almost zero because most of data points has value x3=0.34. Some other values are different but not much different

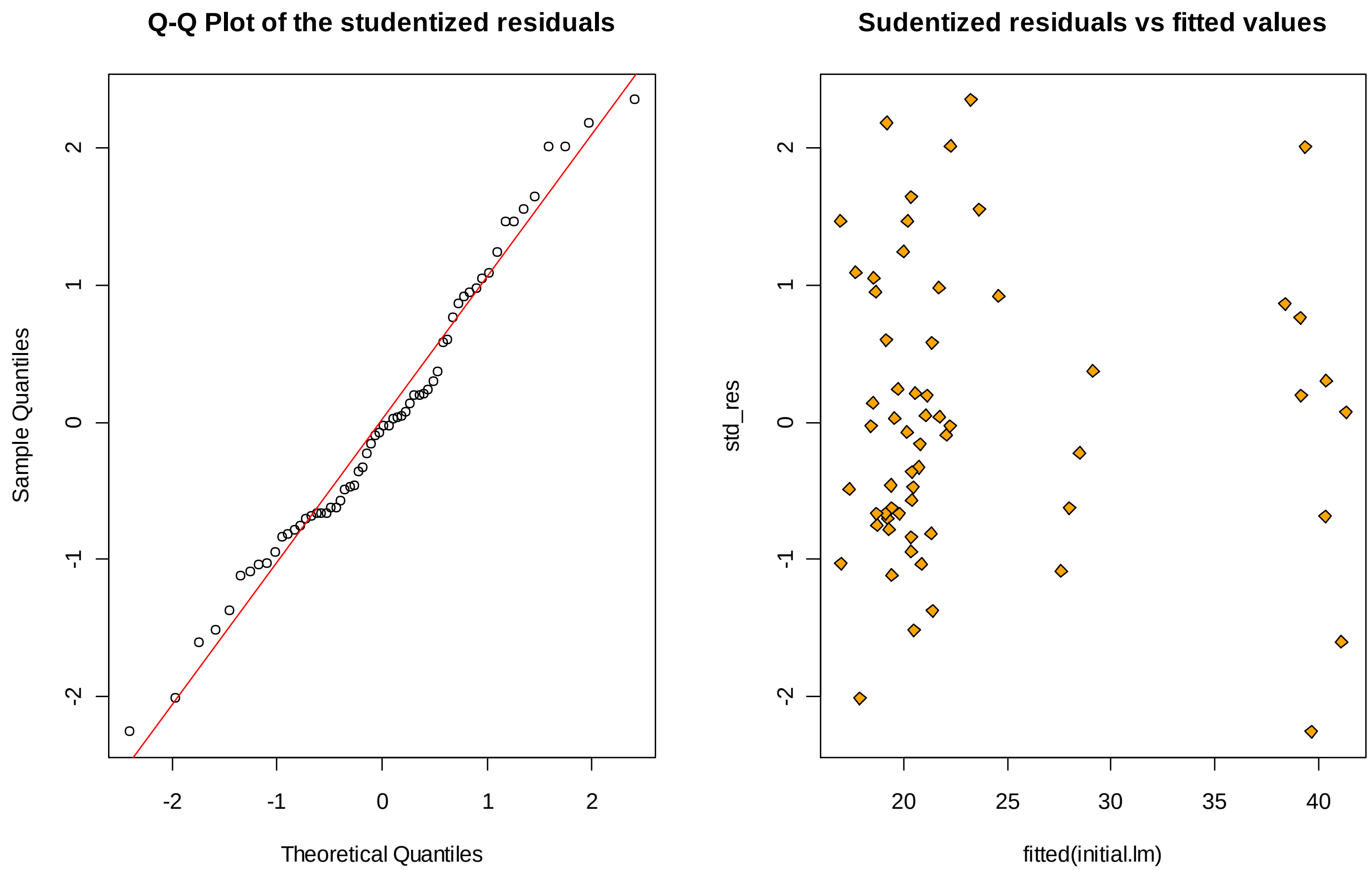
- **Comment:** Among all repressors, x2 has highest correlation with the response variable y. Variables x1 and x2 are almost uncorrelated (Pearson correlation coefficient is 0.002) . Similarly x1 and y are almost uncorrelated (Pearson correlation coefficient is 0.045).

2. Fit and study an initial (full) multiple regression model
   This initial model is used to study the residual analysis and understand **if** at least one regressor is important.

```
> initial.lm= lm(y ~ x1+x2+x3+x4, data=PressureDrop);
> summary(initial.lm);

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = PressureDrop)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.9958 -3.3092 -0.2419  3.3924 10.5668 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.89453    4.32508   1.363  0.17828    
x1          -0.47790    0.34002  -1.406  0.16530    
x2           0.18271    0.01718  10.633 3.78e-15 ***
x3          35.40284   11.09960   3.190  0.00232 ** 
x4           5.84391    2.90978   2.008  0.04935 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.014 on 57 degrees of freedom
Multiple R-squared: 0.6914,     Adjusted R-squared: 0.6697 
F-statistic: 31.92 on 4 and 57 DF,  p-value: 5.818e-14
```

- **Comment:** The overall F test (F-statistic = 31.92) indicates that one or more regressors are important. The t-test in individual coefficients indicate that only x1 (superficial fluid velocity of the gas) is not important. We will analyze this further because this individual coefficient and its corresponding t-test only tell part of the story (and sometimes misleading) due to the nature of partial effect and if there is multi-collinearity or interaction in the data.

3. Plot of the residuals versus the predicted values and the normal probability plot of the residuals



- **Comment:** The studentized residuals are quite spreading evenly around zero. Most residuals are less than 3 std. deviation from the mean zero. There is no serious departure from normality.

4. Since the number of regressors are small (4), we can use all-possible-regressions approach to fit all regression equations involving one candidate regressor, two candidate regressors, and so on using both adjusted R square and Mallows Cp statistics criteria to find the best subset.

```
                         The SAS System          16:16 Wednesday, May 4

                        The REG Procedure
                         Model: MODEL1
                     Dependent Variable: y

                      C(p) Selection Method

          Number of Observations Read          62
          Number of Observations Used          62
```

Number in Model	C(p)	R-Square	Variables in Model
3	4.9755	0.6807	x2 x3 x4
4	5.0000	0.6914	x1 x2 x3 x4
2	5.7063	0.6659	x2 x3
3	7.0336	0.6695	x1 x2 x3
2	12.7306	0.6279	x2 x4
3	13.1733	0.6363	x1 x2 x4
1	13.7477	0.6115	x2
2	15.3353	0.6138	x1 x2
3	116.0593	0.0792	x1 x3 x4
2	116.4343	0.0664	x3 x4

Comment: Out of all possible regression models, three models below have similar small Cp statistics and similar adjusted R square values.

- The initial **full** model: y ~ x1+ x2+ x3 + x4
- The first reduced model: y ~ x2+ x3 + x4
- The second reduce model: y ~ x2+ x3

5. Perform validation of 3 regression models suggested at step 4.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = PressureDrop)

Coefficients:
(Intercept)           x1           x2           x3           x4
     5.8945      -0.4779       0.1827      35.4028       5.8439

> PRESS(initial.lm);
[1] 1762.856
>
> x2x3x4.lm;

Call:
lm(formula = y ~ x2 + x3 + x4, data = PressureDrop)

Coefficients:
(Intercept)           x2           x3           x4
     4.6406       0.1830      34.6243       4.5688

> PRESS(x2x3x4.lm);
[1] 1750.517
>
> x2x3.lm;

Call:
lm(formula = y ~ x2 + x3, data = PressureDrop)

Coefficients:
(Intercept)           x2           x3
     7.1897       0.1846      35.1162

> PRESS(x2x3.lm);
[1] 1711.096
```

Comment:

- Here we use PRESS statistic to assess the predictive power of 3 suggested models. This is one form of data splitting in a sense that we use n-1 data points to estimate/train the model and use observation (i) to assess the model. The smaller the PRESS statistic the better.

- In the 3 model above… the last (most reduced) model
  - x2x3.lm = lm(y ~ x2+x3, data=PressureDrop);

  has smallest PRESS statistic.

- This reduced model, y ~ x2+x3, also has small Cp value while it has equivalent adjusted R square value comparing to the initial full model.

6. Perform multi-collinearity diagnostics of 3 regression models suggested at step 4.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	5.89453	4.32508	1.36	0.1783	0
x1	1	-0.47790	0.34002	-1.41	0.1653	1.11114
x2	1	0.18271	0.01718	10.63	<.0001	1.00504
x3	1	35.40284	11.09960	3.19	0.0023	1.00517
x4	1	5.84391	2.90978	2.01	0.0494	1.11159

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	4.11226	1.00000
2	0.68331	2.45319
3	0.10184	6.35455
4	0.08992	6.76253
5	0.01267	18.01831

Collinearity Diagnostics

Number	Proportion of Variation				
	Intercept	x1	x2	x3	x4
1	0.00123	0.00700	0.01647	0.00149	0.00624
2	0.00060624	0.00577	0.97162	0.00089677	0.00352

Comment:

- For the initial full model… there is no problem with collinearity in the data
- From VIF analysis… all variance inflation factors are around 1.0 and 1.1
- From Eigensystem analysis… all condition indices are less than 20 so there is no problem with collinearity in the set of 4 regressors

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.64065	4.26751	1.09	0.2813	0
x2	1	0.18302	0.01733	10.56	<.0001	1.00488
x3	1	34.62435	11.17861	3.10	0.0030	1.00267
x4	1	4.56878	2.78788	1.64	0.1067	1.00353

Collinearity Diagnostics

			Proportion of Variation			
Number	Eigenvalue	Condition Index	Intercept	x2	x3	x4
1	3.23620	1.00000	0.00206	0.02989	0.00239	0.01106
2	0.65975	2.21477	0.00142	0.96327	0.00192	0.00832
3	0.09127	5.95459	0.02917	2.212183E-8	0.06146	0.93612
4	0.01278	15.91567	0.96735	0.00684	0.93422	0.04449

Comment:

- For the first reduced model y ~ x2+x3+x4, there is no problem with collinearity in the data
- From VIF analysis… all variance inflation factors are around 1.0
- From Eigensystem analysis… all condition indices are less than 16 so there is no problem with collinearity in the set of x2,x3,x4 regressors

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	7.18971	4.03031	1.78	0.0796	0
x2	1	0.18456	0.01755	10.52	<.0001	1.00194
x3	1	35.11616	11.33308	3.10	0.0030	1.00194

Collinearity Diagnostics

			Proportion of Variation		
Number	Eigenvalue	Condition Index	Intercept	x2	x3
1	2.36405	1.00000	0.00434	0.06734	0.00438
2	0.62266	1.94851	0.00429	0.92372	0.00489
3	0.01329	13.33953	0.99137	0.00894	0.99072

Comment:

- For the second reduced model y ~ x2+x3, there is no problem with collinearity in the data
- From VIF analysis… all variance inflation factors are around 1.0
- From Eigensystem analysis… all condition indices are less than 14. There is no problem with collinearity in the set of x2, x3 regressors

Comments

- After analysis in step 4, 5, 6 it is suggested that x2 and x3 are main explanatory variables of the response variable y (pressure drop through a bubble cap).

- In terms of adjusted R^2 , both x2 and x3 collectively has the same explain power of the variability of y as all x1, x2, x3, x4 because their adjusted R^2 values are similar

- Similarly in term of Cp statistic, the reduced model with only x2 and x3 in the regression equation is adequate.

- In terms of PRESS statistic, the reduced model with only x2 and x3 in the regression equation give the best PRESS statistic.

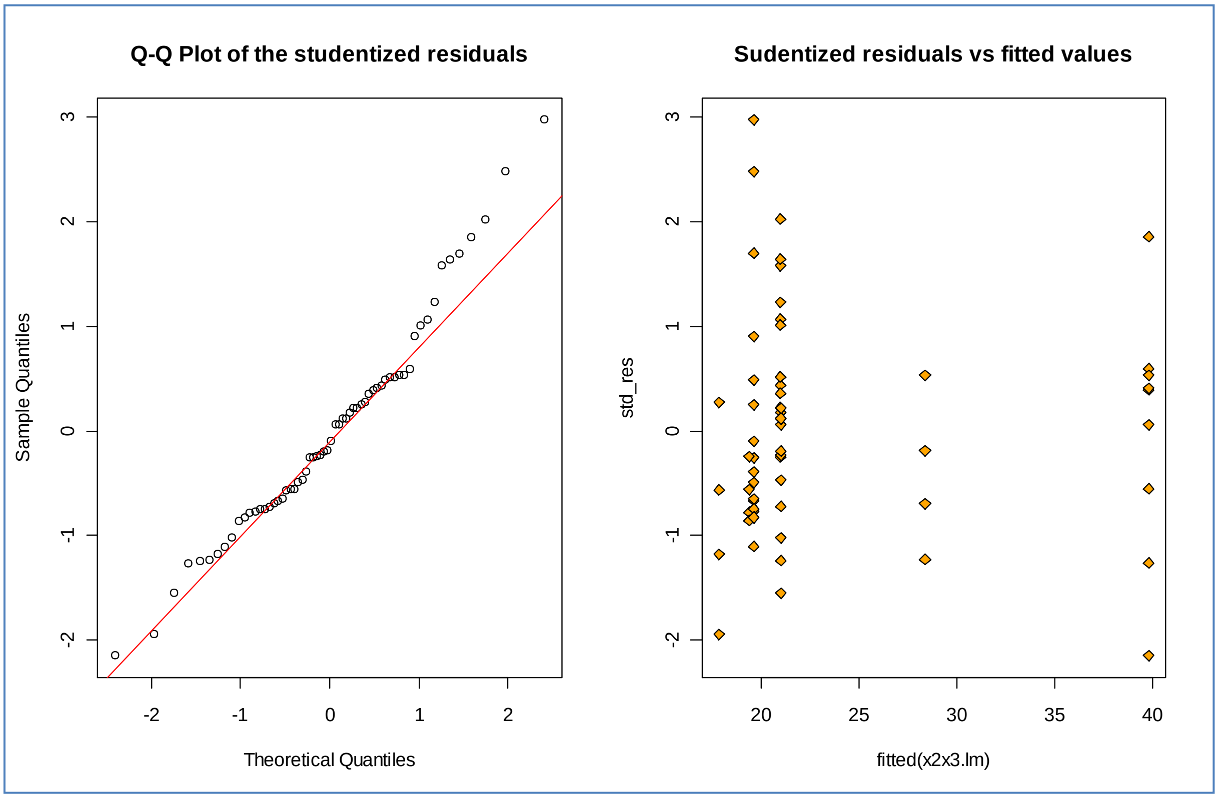
- If I have more time, I would two third of the data (randomly) for estimating the regression model and use 1/3 of the 62 data points to compute the root means square error (RMSE) or sum of square error to validate the model. However the PRESS statistic is also can be considered as a method to validate the model because it is equivalent to leave-one-out cross validation.

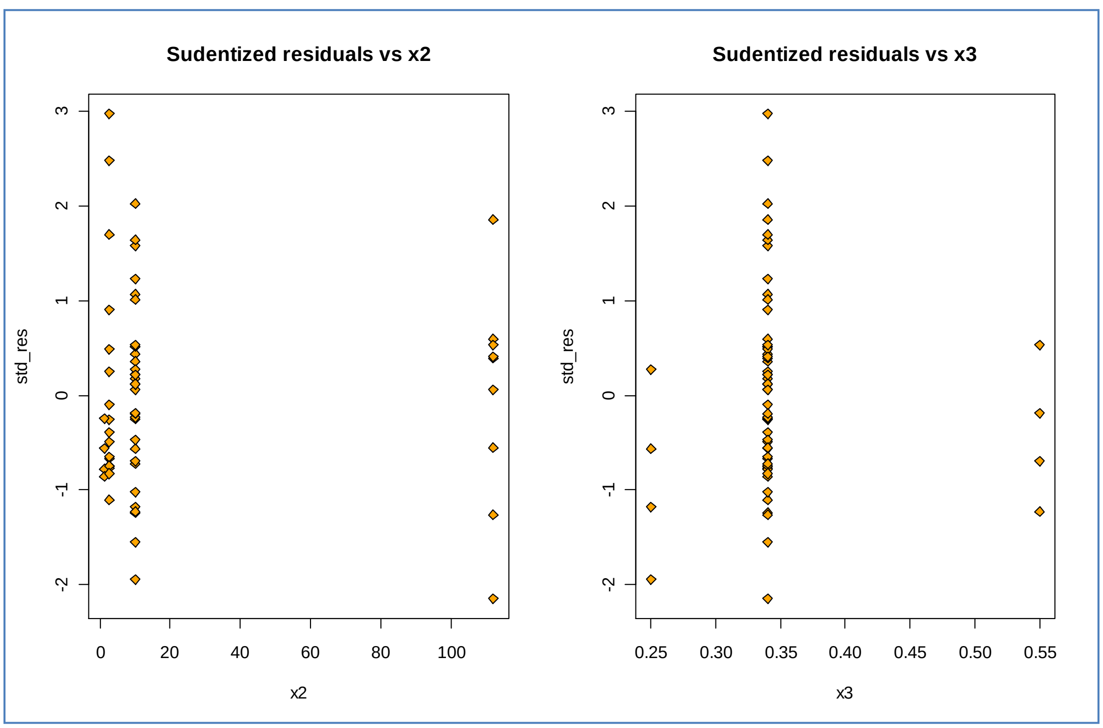
- In term of multi-collinearity issue, there is no problem in any model, including the full model

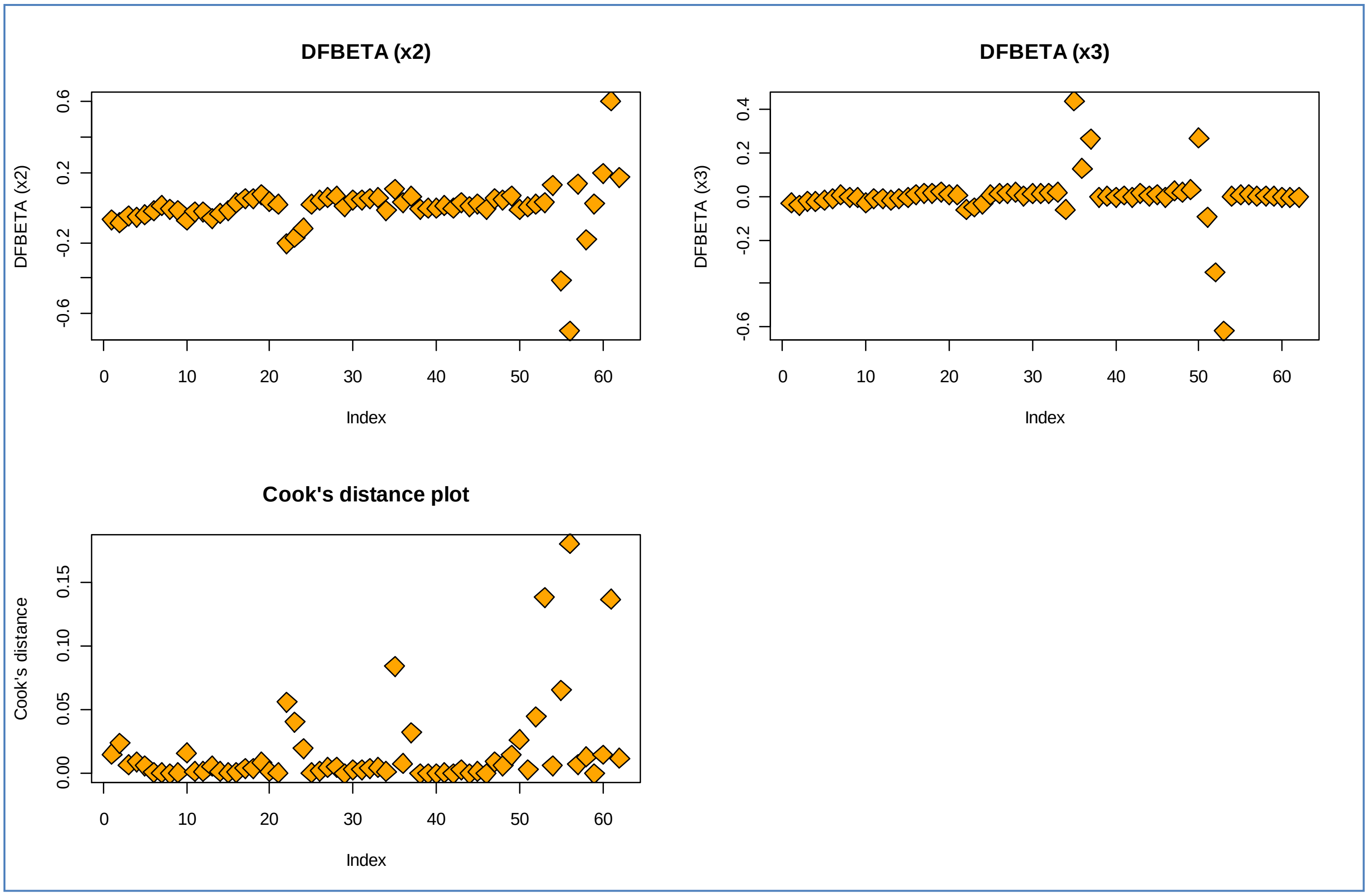
With the above analysis I would suggest the reduced model with only x2 and x3 in the regression equation.

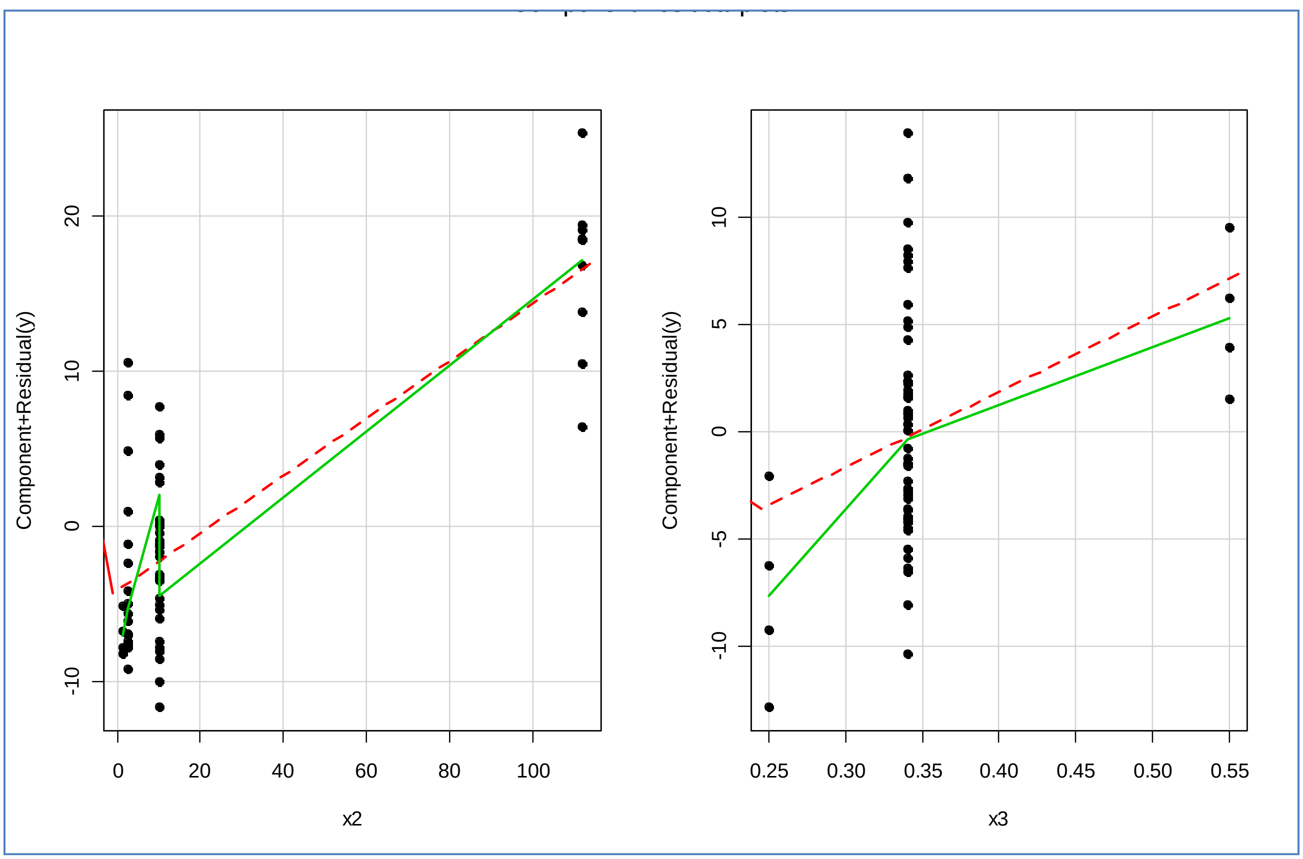
```
Call:
lm(formula = y ~ x2 + x3, data = PressureDrop)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0994  -3.6236  -0.6911   2.4722  14.1854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.18971    4.03031   1.784  0.07958 .  
x2           0.18456    0.01755  10.518 3.74e-15 ***
x3          35.11616   11.33308   3.099  0.00298 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.127 on 59 degrees of freedom
Multiple R-squared: 0.6659,     Adjusted R-squared: 0.6546 
F-statistic:  58.8 on 2 and 59 DF,  p-value: 9.007e-15
```

7. Perform further analysis on the reduced model (only x2, x3 in the equation) to see if transformation or weighted regression is needed






DFBETA (x2)
DFBETA (x2)
Index
DFBETA (x3)
DFBETA (x3)
Index
Cook's distance plot
Cook's distance
Index


Component+Residual(y)
x2
Component+Residual(y)
x3

Comments

- From the cook distance plot there is no issue as the cook distance is less than 0.2 for all points
- From leverage analysis through the **H** matrix (see attached R/SPLUSs code) there is no serious problem.
- From the measure of influence statistics DFBETAS and DFFITS there is no serious influence point
- From the added-variable plot (in attached R code, but not shown here) and the component-residual plot (shown above) … it is suggestive that term X3 is not perfectly entered the model linearly.
- However, after different trying to transform y and/or x3 the resulted transformation does not give a better model in terms of adjusted R square as well as PRESS statistic.

<u>Recommended model</u>

- **y,** the response variable, is the pressure drop through a bubble cap
- x2 - Kinematic viscosity
- x3 - Mesh opening

```
Call:
lm(formula = y ~ x2 + x3, data = PressureDrop)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0994  -3.6236  -0.6911   2.4722  14.1854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.18971    4.03031   1.784  0.07958 .  
x2           0.18456    0.01755  10.518 3.74e-15 ***
x3          35.11616   11.33308   3.099  0.00298 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.127 on 59 degrees of freedom
Multiple R-squared: 0.6659,     Adjusted R-squared: 0.6546 
F-statistic:  58.8 on 2 and 59 DF,  p-value: 9.007e-15
```

- I don't recommend 2 or three models in this analysis because this model has smaller subset of repressors, while having similar adjusted R square value and Cp statistic value.
- This model also has more predict power because its PRESS statistic is smaller comparing to two other models where x4 and x1 is included model.
- The conclusion of my analysis is similar to the author in the following article http://onlinelibrary.wiley.com/doi/10.1002/cjce.5450710317/abstract

  In terms of concluding that x2 and x3 are only explanatory variables.