Exercise 4:

1. Regression model shows high variance inflation factor in variable x12 (Relative population potential of hydrocarbons, HC) and x13 (Relative population potential of oxides of nitrogen, NOx), which are 98.64 and 104.98 respectively. - Thees are indication of multicollinearity.

2. Since each variables has its own measure unit, and they are not comparable, using correlation matrix forces each of them contributes the same variability to the total variance. PCA model show that the cumulative percentage shows that first 5 PCs will contributes to 79.40% of total variability. Therefore, the requirement of at least 75% can be achieved using the first 5 PC's.

3. Denote z_i to be the i^{th} PC, then:

Model with the 1st PC (RMSE = 53.07647):

$$y = 940.35850 - 15.58781 z_i$$

Model with the first 2 PC's is (RMSE = 53.25078) :

$$y = 940.35850 - 15.58781 z_1 + 3.29131 z_2$$

Model with the first 3 PC's is (RMSE = 45.11376):

$$y = 940.35850 - 15.58781 z_1 + 3.29131 z_2 + 19.82857 z_3$$

Model with the first 4 PC's (RMSE = 45.40604):

$$y = 940.35850 - 15.58781 z_1 + 3.29131 z_2 + 19.82857 z_3 - 2.70028 z_4$$

Model with the first 5 PC's (RMSE = 45.81701):

$$y = 940.35850 - 15.58781 z_1 + 3.29131 z_2 + 19.82857 z_3 - 2.70028 z_4 + 0.71875 z_5$$

The following table gives information about Root MSE in each of the model:

Number of PC	Root MSE
First PC's	53.07647
First 2 PC's	53.25078
First 3 PC's	45.11376
First 4 PC's	45.40604
First 5 PC's	45.81701

The two models has smallest RMSE is model using the first 4 PC's and model using the first 3 PC's.

4. Regression with C_p screening method shows that the best model based on the first 5 PC's is the model involving PC1 and PC3, which has $C_p = 1.1328$ and R-square = 0.4931. The prediction equation is:

$$y = 940.35850 - 15.58781 z_1 + 19.82857 z_3$$

None of the previous two model is similar to this one. This is because the criteria in each method are different. We try to keep as much as variance as possible in previous model, that is why PC's were added to regression model by order. While C_p is the measure of total variation in the predicted responses, C_p criteria is choosing the model that has the smallest C_p number. And it may happen that we can't achieve both criteria at the same time.

Exercise 4
First 10 observations of data

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	y
1	36	27	71	8.1	3.34	11.4	81.5	3243	8.8	42.6	11.7	21	15	59	59	921.87
2	35	23	72	11.1	3.14	11.0	78.8	4281	3.5	50.7	14.4	8	10	39	57	997.88
3	44	29	74	10.4	3.21	9.8	81.6	4260	0.8	39.4	12.4	6	6	33	54	962.35
4	47	45	79	6.5	3.41	11.1	77.5	3125	27.1	50.2	20.6	18	8	24	56	982.29
5	43	35	77	7.6	3.44	9.6	84.6	6441	24.4	43.7	14.3	43	38	206	55	1071.29
6	53	45	80	7.7	3.45	10.2	66.8	3325	38.5	43.1	25.5	30	32	72	54	1030.38
7	43	30	74	10.9	3.23	12.1	83.9	4679	3.5	49.2	11.3	21	32	62	56	934.70
8	45	30	73	9.3	3.29	10.6	86.0	2140	5.3	40.4	10.5	6	4	4	56	899.53
9	36	24	70	9.0	3.31	10.5	83.2	6582	8.1	42.5	12.6	18	12	37	61	1001.90
10	36	27	72	9.5	3.36	10.7	79.3	4213	6.7	41.0	13.2	12	7	20	59	912.35

Huong Tran - Assignment 3

Exercise 4
Regression model of air pollution

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	174630	11642	9.54	<.0001
Error	44	53681	1220.02049		
Corrected Total	59	228311			

Root MSE	34.92879	R-Square	0.7649
Dependent Mean	940.35850	Adj R-Sq	0.6847
Coeff Var	3.71441		

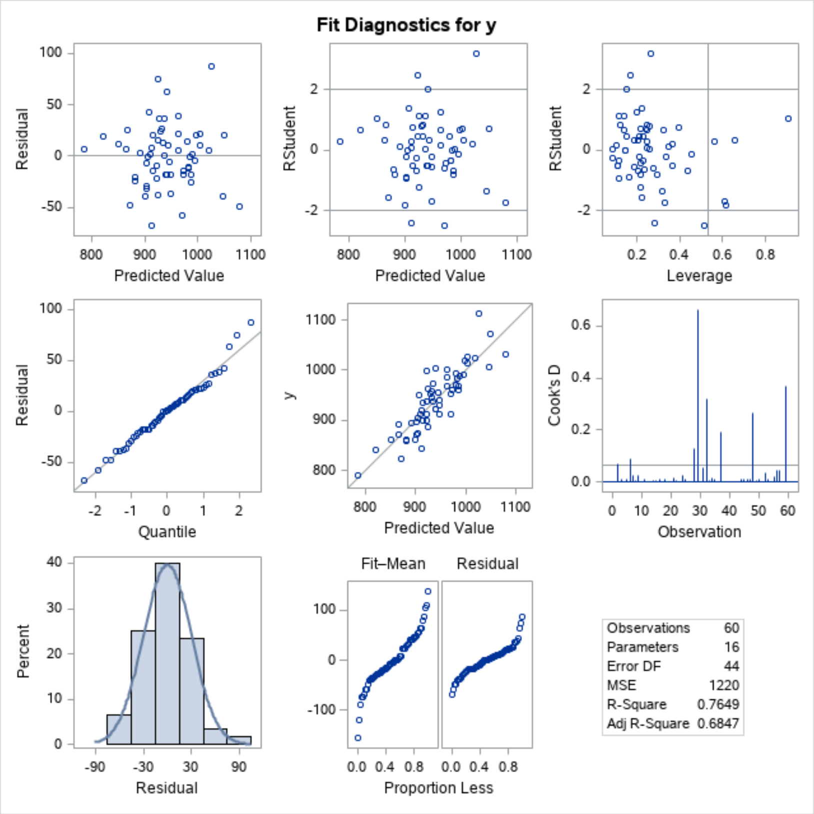
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1763.99793	437.33031	4.03	0.0002	0
x1	1	1.90536	0.92374	2.06	0.0451	4.11389
x2	1	-1.93762	1.10839	-1.75	0.0874	6.14355
x3	1	-3.10040	1.90167	-1.63	0.1102	3.96777
x4	1	-9.06517	8.48622	-1.07	0.2912	7.47004
x5	1	-106.83103	69.78007	-1.53	0.1329	4.30762
x6	1	-17.15689	11.86012	-1.45	0.1551	4.86054
x7	1	-0.65111	1.76777	-0.37	0.7144	3.99478
x8	1	0.00360	0.00403	0.89	0.3761	1.65828
x9	1	4.45958	1.32721	3.36	0.0016	6.77960
x10	1	-0.18715	1.66169	-0.11	0.9108	2.84158
x11	1	-0.16741	3.22730	-0.05	0.9589	8.71707
x12	1	-0.67216	0.49102	-1.37	0.1780	98.63993
x13	1	1.34010	1.00559	1.33	0.1895	104.98240
x14	1	0.08626	0.14752	0.58	0.5617	4.22893
x15	1	0.10674	1.16943	0.09	0.9277	1.90709

Huong Tran - Assignment 3

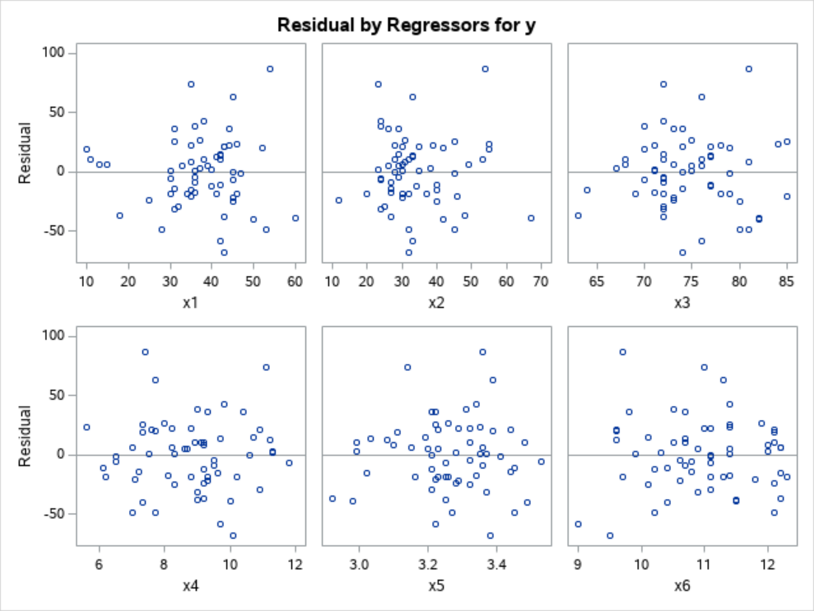
Exercise 4
Regression model of air pollution

The REG Procedure
Model: MODEL1
Dependent Variable: y

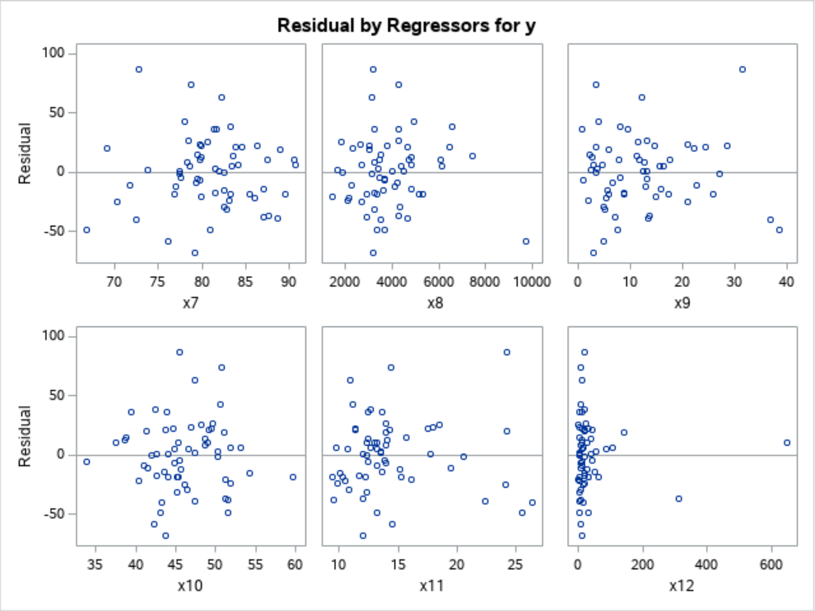


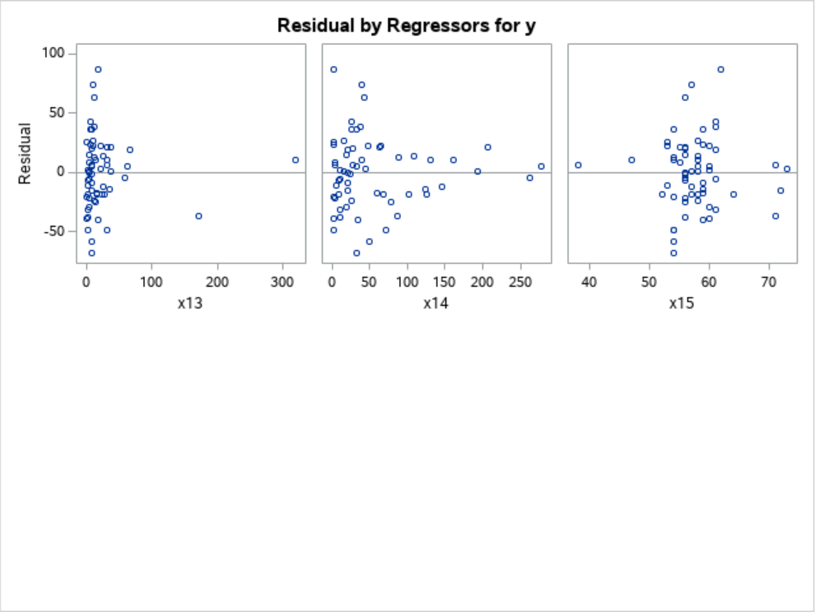
Fit Diagnostics for y

Observations	60
Parameters	16
Error DF	44
MSE	1220
R-Square	0.7649
Adj R-Square	0.6847



Residual by Regressors for y





Huong Tran - Assignment 3

Exercise 4
PCA - Air pollution

The PRINCOMP Procedure

Observations	60
Variables	15

Simple Statistics

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x
Mean	37.36666667	33.98333333	74.58333333	8.798333333	3.263166667	10.97333333	80.91333333	3876.050000	11.87000000	46.08166667	14.37333333	37.85000000	22.65000000	53.76666667	57.66666
StD	9.98467753	10.16889852	4.76317679	1.464551955	0.135252327	0.84529940	5.14137312	1454.102361	8.92114798	4.61304310	4.16009561	91.97767323	46.33328964	63.39046784	5.36993

Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
x1	1.0000	0.0922	0.5033	0.1011	0.2634	-.4904	-.4908	-.0035	0.4132	-.2973	0.5066	-.5318	-.4873	-.1069	-.0773
x2	0.0922	1.0000	0.3463	-.3981	-.2092	0.1163	0.0149	-.1001	0.4538	0.2380	0.5653	0.3508	0.3210	-.1078	0.0679
x3	0.5033	0.3463	1.0000	-.4340	0.2623	-.2385	-.4150	-.0610	0.5753	-.0214	0.6193	-.3565	-.3377	-.0993	-.4528
x4	0.1011	-.3981	-.4340	1.0000	-.5091	-.1389	0.0650	0.1620	-.6378	-.1177	-.3098	-.0205	-.0021	0.0172	0.1124
x5	0.2634	-.2092	0.2623	-.5091	1.0000	-.3951	-.4106	-.1843	0.4194	-.4257	0.2599	-.3882	-.3584	-.0041	-.1357
x6	-.4904	0.1163	-.2385	-.1389	-.3951	1.0000	0.5522	-.2439	-.2088	0.7032	-.4033	0.2868	0.2244	-.2343	0.1765
x7	-.4908	0.0149	-.4150	0.0650	-.4106	0.5522	1.0000	0.1819	-.4103	0.3387	-.6807	0.3868	0.3483	0.1180	0.1219

Correlation Matrix

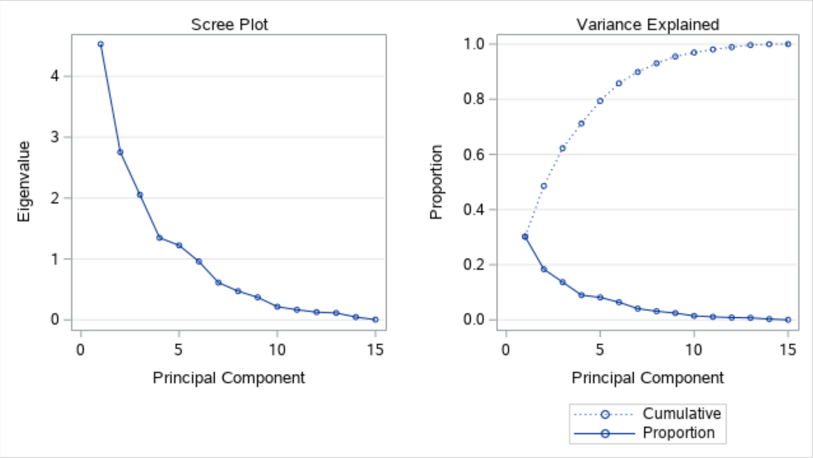
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
x8	-.0035	-.1001	-.0610	0.1620	-.1843	-.2439	0.1819	1.0000	-.0057	-.0318	-.1629	0.1203	0.1653	0.4321	-.1250
x9	0.4132	0.4538	0.5753	-.6378	0.4194	-.2088	-.4103	-.0057	1.0000	-.0044	0.7049	-.0259	0.0184	0.1593	-.1180
x10	-.2973	0.2380	-.0214	-.1177	-.4257	0.7032	0.3387	-.0318	-.0044	1.0000	-.1852	0.2037	0.1600	-.0685	0.0607
x11	0.5066	0.5653	0.6193	-.3098	0.2599	-.4033	-.6807	-.1629	0.7049	-.1852	1.0000	-.1298	-.1025	-.0965	-.1522
x12	-.5318	0.3508	-.3565	-.0205	-.3882	0.2868	0.3868	0.1203	-.0259	0.2037	-.1298	1.0000	0.9838	0.2823	-.0202
x13	-.4873	0.3210	-.3377	-.0021	-.3584	0.2244	0.3483	0.1653	0.0184	0.1600	-.1025	0.9838	1.0000	0.4094	-.0459
x14	-.1069	-.1078	-.0993	0.0172	-.0041	-.2343	0.1180	0.4321	0.1593	-.0685	-.0965	0.2823	0.4094	1.0000	-.1026
x15	-.0773	0.0679	-.4528	0.1124	-.1357	0.1765	0.1219	-.1250	-.1180	0.0607	-.1522	-.0202	-.0459	-.1026	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.52839160	1.77355006	0.3019	0.3019
2	2.75484154	0.70037750	0.1837	0.4855
3	2.05446404	0.70607446	0.1370	0.6225
4	1.34838958	0.12516962	0.0899	0.7124
5	1.22321996	0.26277598	0.0815	0.7940
6	0.96044398	0.34770243	0.0640	0.8580
7	0.61274155	0.14072983	0.0408	0.8988
8	0.47201172	0.10115870	0.0315	0.9303
9	0.37085302	0.15445834	0.0247	0.9550
10	0.21639468	0.05004428	0.0144	0.9695
11	0.16635040	0.03934529	0.0111	0.9805
12	0.12700511	0.01301833	0.0085	0.9890
13	0.11398677	0.06794703	0.0076	0.9966
14	0.04603974	0.04117345	0.0031	0.9997
15	0.00486629		0.0003	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15
x1	-.345479	-.102644	0.026814	0.332836	0.122322	0.182749	-.012230	0.486269	0.511519	0.043197	0.116337	0.176019	-.304682	-.269586	0.010002
x2	-.065253	0.482160	-.106010	0.328810	-.085158	0.078125	-.361931	0.068805	-.233566	0.388214	-.061821	0.241011	-.204280	0.431137	0.006663
x3	-.344486	0.195414	-.078102	0.024804	0.398216	-.115198	-.091622	0.126786	-.282331	-.398096	0.537503	0.099787	0.313506	0.068848	0.005121
x4	0.162984	-.364872	0.156177	0.520266	0.035112	-.139829	0.209995	0.124547	0.128086	0.067300	-.043288	0.054673	0.475266	0.459541	0.044591
x5	-.297274	-.065986	0.031979	-.559719	-.230784	0.026087	-.035724	0.018059	0.286594	0.363021	0.151021	0.378662	0.350476	0.184447	0.021337
x6	0.286505	0.172856	-.429393	-.110144	0.135099	0.030269	0.151159	0.077909	0.199693	-.417767	-.350462	0.541477	0.015632	0.102208	0.048169
x7	0.360761	0.050430	-.054928	-.148351	0.193103	0.162125	-.495214	0.485936	0.018302	0.133974	-.149770	-.243123	0.399638	-.193275	-.021203
x8	0.071507	-.021398	0.440225	0.056126	0.417692	0.388432	-.313595	-.528251	0.206562	-.066861	-.042823	0.210383	0.038831	0.000093	0.001479
x9	-.302012	0.368878	0.061182	-.102458	-.017797	0.272246	0.158490	0.045448	0.314995	-.240960	-.240320	-.527822	0.082845	0.395119	0.015576
x10	0.196455	0.240228	-.333329	0.045635	0.383862	0.190599	0.469359	-.170488	0.118339	0.483502	0.269232	-.123898	0.091626	-.112737	-.027259
x11	-.357860	0.268327	0.023721	0.279044	-.133031	-.071306	0.095532	-.160367	-.103327	0.073579	-.420737	0.115959	0.453653	-.501399	0.011177
x12	0.282771	0.367140	0.239428	0.052767	-.216823	-.224512	-.016119	-.012455	0.249251	-.070375	0.265488	-.007130	0.027843	-.119851	0.688878
x13	0.264969	0.363391	0.310218	0.043253	-.202457	-.188680	0.069696	0.057385	0.211929	-.096720	0.201296	0.083349	0.045310	-.062805	-.712151
x14	0.067505	0.095172	0.519432	-.177784	0.101689	0.301013	0.434739	0.377590	-.421833	0.050491	-.114927	0.204871	-.070947	-.002959	0.108311
x15	0.113307	-.082162	-.191181	0.180357	-.520579	0.675848	0.000957	-.052651	-.117371	-.203896	0.300821	0.040852	0.167503	-.071816	-.007743



Huong Tran - Assignment 3

Exercise 4
Regression model by first PC

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	64918	64918	23.04	<.0001
Error	58	163392	2817.11180		
Corrected Total	59	228311			

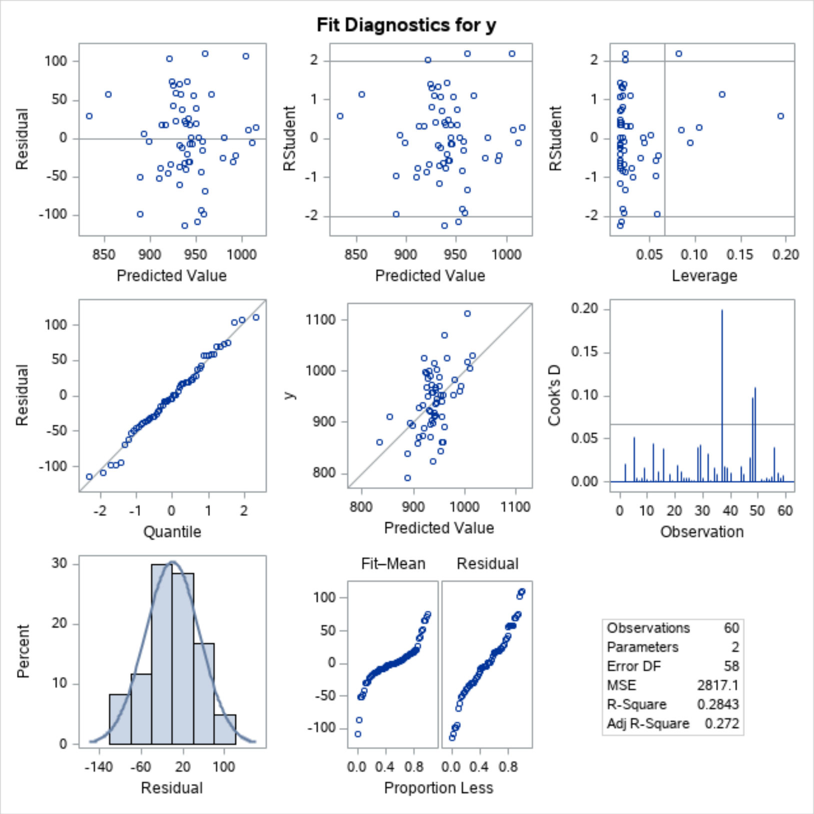
Root MSE	53.07647	R-Square	0.2843
Dependent Mean	940.35850	Adj R-Sq	0.2720
Coeff Var	5.64428		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	940.35850	6.85214	137.24	<.0001
Prin1	1	-15.58781	3.24716	-4.80	<.0001

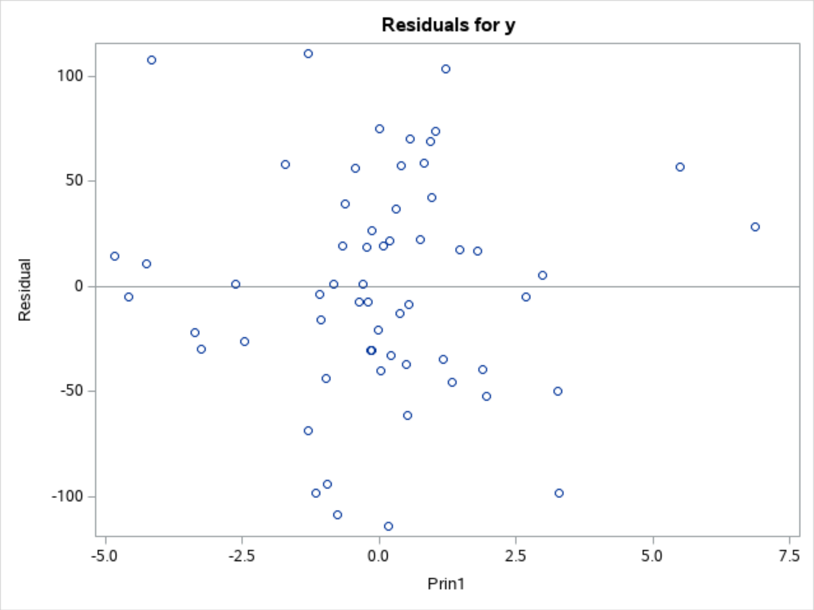
Huong Tran - Assignment 3

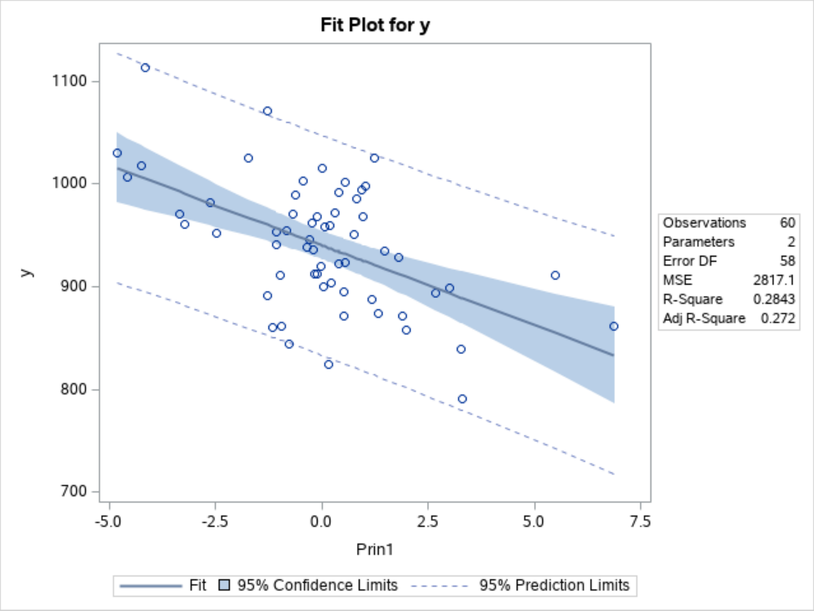
Exercise 4
Regression model by first PC

The REG Procedure
Model: MODEL1
Dependent Variable: y



Observations	60
Parameters	2
Error DF	58
MSE	2817.1
R-Square	0.2843
Adj R-Square	0.272





Observations	60
Parameters	2
Error DF	58
MSE	2817.1
R-Square	0.2843
Adj R-Square	0.272

Huong Tran - Assignment 3

Exercise 4
Regression model by first 2 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	66679	33339	11.76	<.0001
Error	57	161632	2835.64535		
Corrected Total	59	228311			

Root MSE	53.25078	R-Square	0.2921
Dependent Mean	940.35850	Adj R-Sq	0.2672
Coeff Var	5.66282		

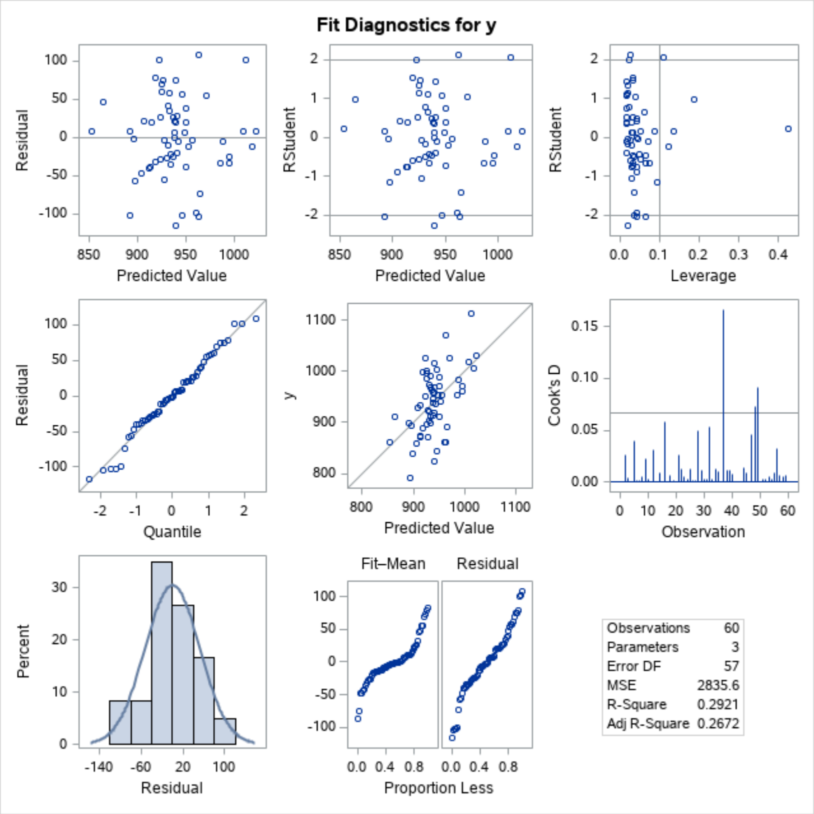
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	940.35850	6.87465	136.79	<.0001
Prin1	1	-15.58781	3.25783	-4.78	<.0001
Prin2	1	3.29131	4.17688	0.79	0.4340

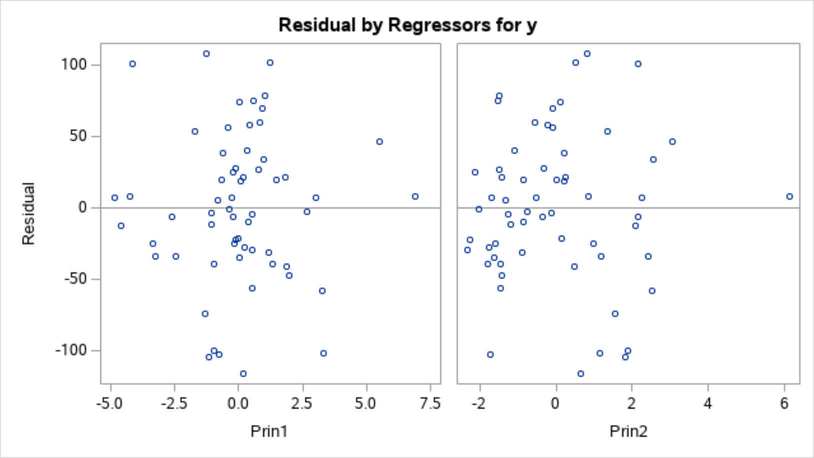
Huong Tran - Assignment 3

Exercise 4
Regression model by first 2 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y



Observations	60
Parameters	3
Error DF	57
MSE	2835.6
R-Square	0.2921
Adj R-Square	0.2672



Huong Tran - Assignment 3

Exercise 4
Regression model by first 3 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	114337	38112	18.73	<.0001
Error	56	113974	2035.25098		
Corrected Total	59	228311			

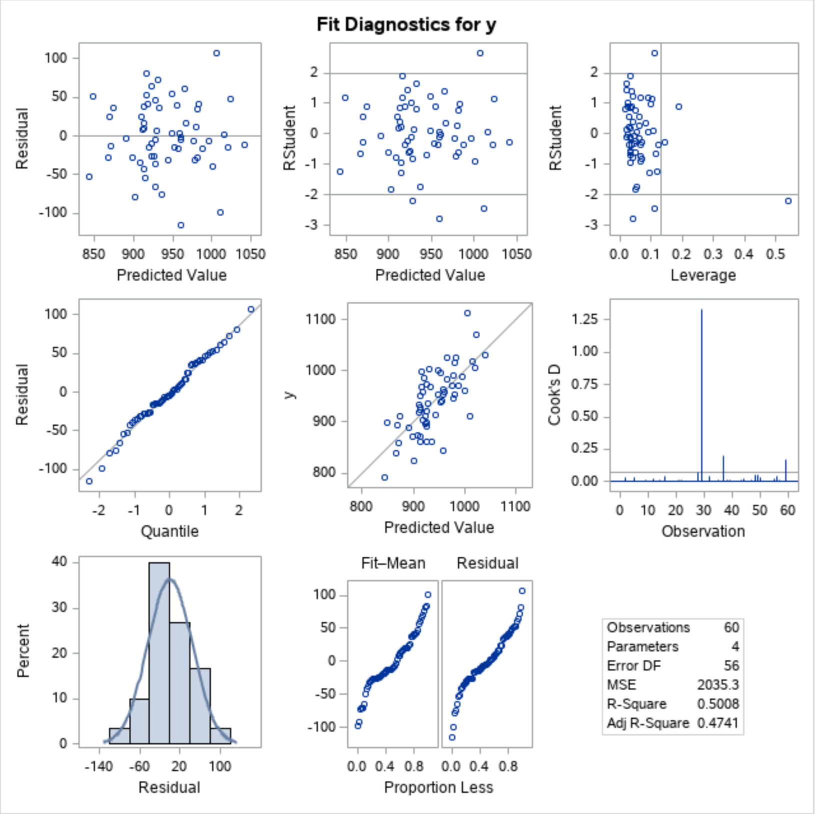
Root MSE	45.11376	R-Square	0.5008
Dependent Mean	940.35850	Adj R-Sq	0.4741
Coeff Var	4.79751		

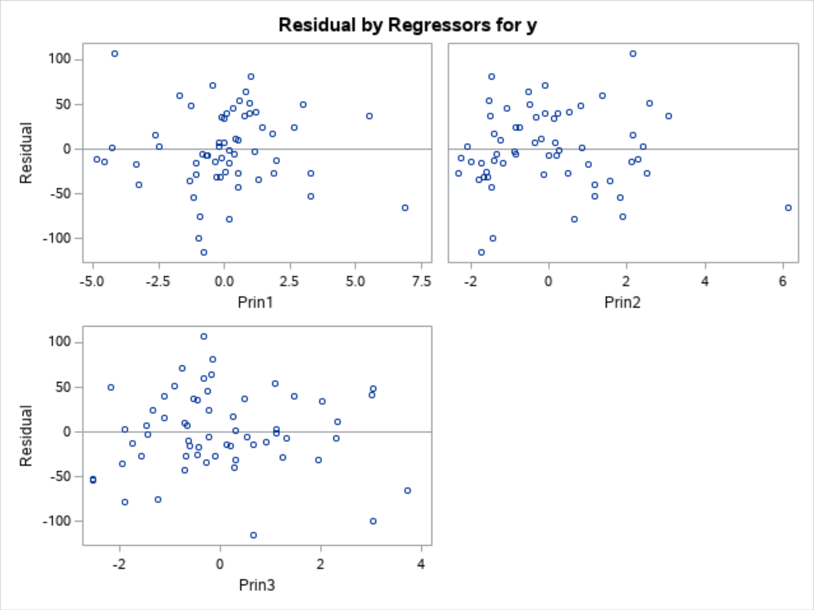
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	940.35850	5.82416	161.46	<.0001
Prin1	1	-15.58781	2.76001	-5.65	<.0001
Prin2	1	3.29131	3.53863	0.93	0.3563
Prin3	1	19.82857	4.09764	4.84	<.0001

Huong Tran - Assignment 3

Exercise 4
Regression model by first 3 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y





Huong Tran - Assignment 3

Exercise 4
Regression model by first 4 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	114917	28729	13.93	<.0001
Error	55	113394	2061.70868		
Corrected Total	59	228311			

Root MSE	45.40604	**R-Square**	0.5033
Dependent Mean	940.35850	**Adj R-Sq**	0.4672
Coeff Var	4.82859		

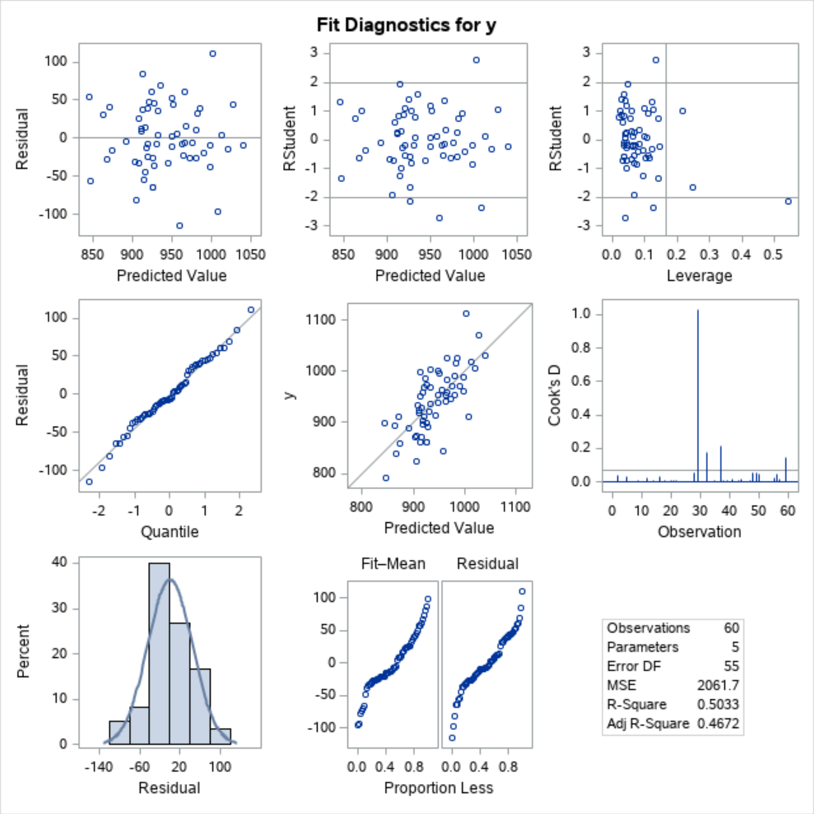
Parameter Estimates

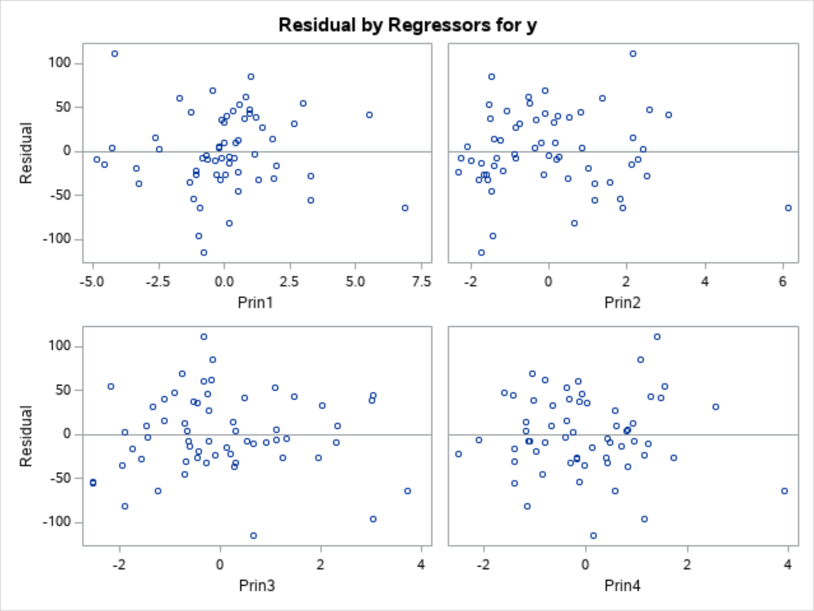
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	940.35850	5.86189	160.42	<.0001
Prin1	1	-15.58781	2.77789	-5.61	<.0001
Prin2	1	3.29131	3.56155	0.92	0.3595
Prin3	1	19.82857	4.12419	4.81	<.0001
Prin4	1	-2.70028	5.09073	-0.53	0.5979

Huong Tran - Assignment 3

Exercise 4
Regression model by first 4 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y





Huong Tran - Assignment 3

Exercise 4
Regression model by first 5 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	114954	22991	10.95	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Error	54	113357	2099.19804		
Corrected Total	59	228311			

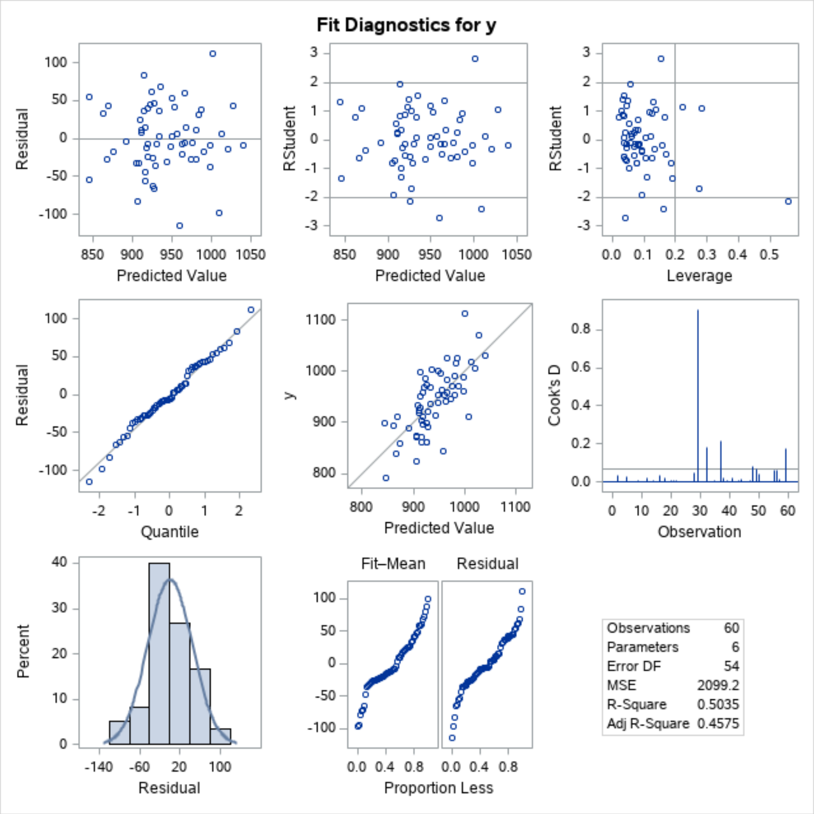
Root MSE	45.81701	R-Square	0.5035
Dependent Mean	940.35850	Adj R-Sq	0.4575
Coeff Var	4.87229		

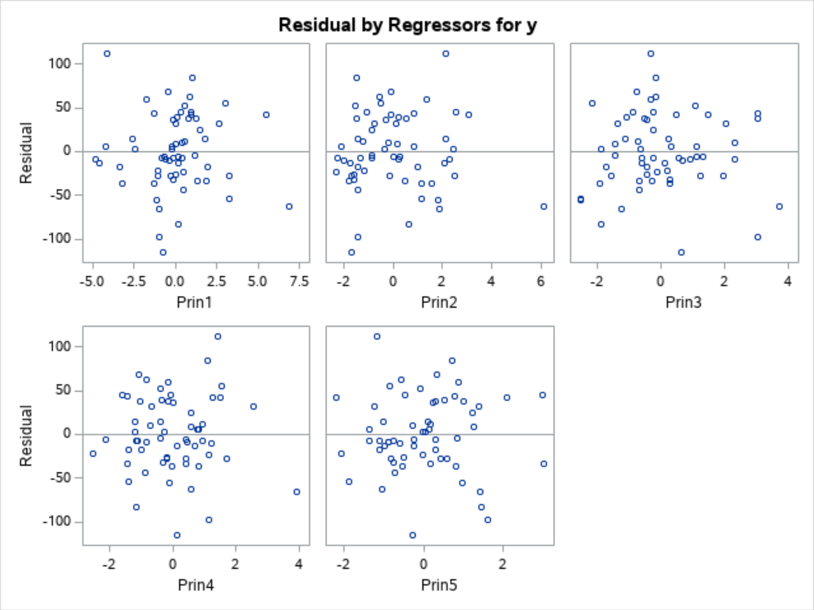
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	940.35850	5.91495	158.98	<.0001
Prin1	1	-15.58781	2.80304	-5.56	<.0001
Prin2	1	3.29131	3.59379	0.92	0.3638
Prin3	1	19.82857	4.16151	4.76	<.0001
Prin4	1	-2.70028	5.13680	-0.53	0.6013
Prin5	1	0.71875	5.39322	0.13	0.8945

Huong Tran - Assignment 3

Exercise 4
Regression model by first 5 PC's

The REG Procedure
Model: MODEL1
Dependent Variable: y





Huong Tran - Assignment 3

Exercise 4
Root Square Mean of each model corresponding upto the first 5 PC

Obs	Model	RootMSE
1	1	53.0765
2	2	53.2508
3	3	45.1138
4	4	45.4060
5	5	45.8170

Huong Tran - Assignment 3

Exercise 4
Regression moddel with CP criteria

The REG Procedure
Model: MODEL1
Dependent Variable: y

C(p) Selection Method

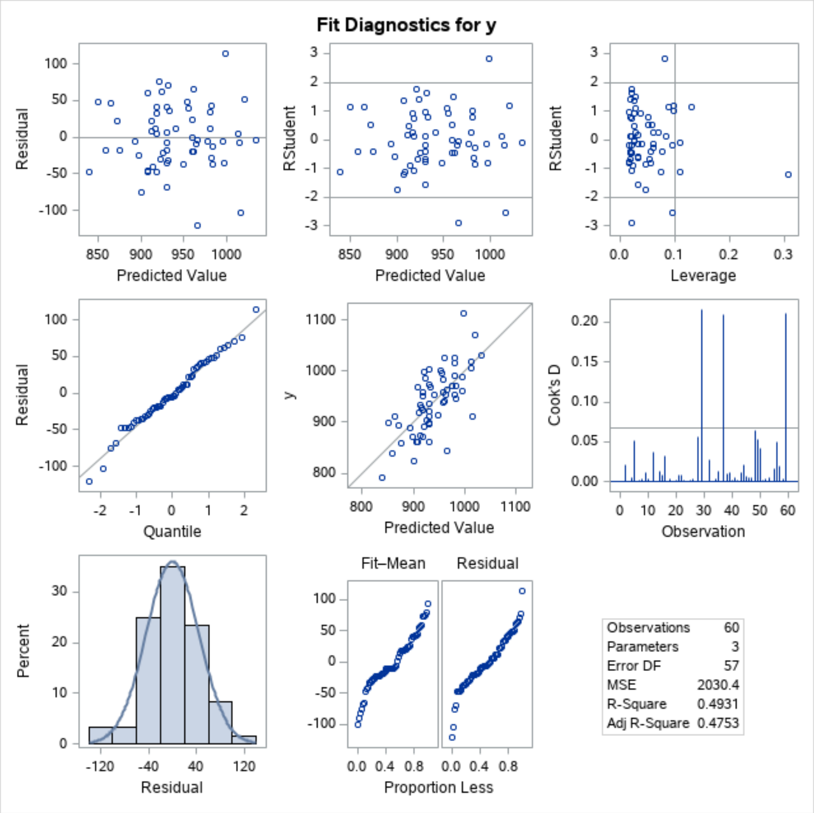
Number of Observations Read	60
Number of Observations Used	60

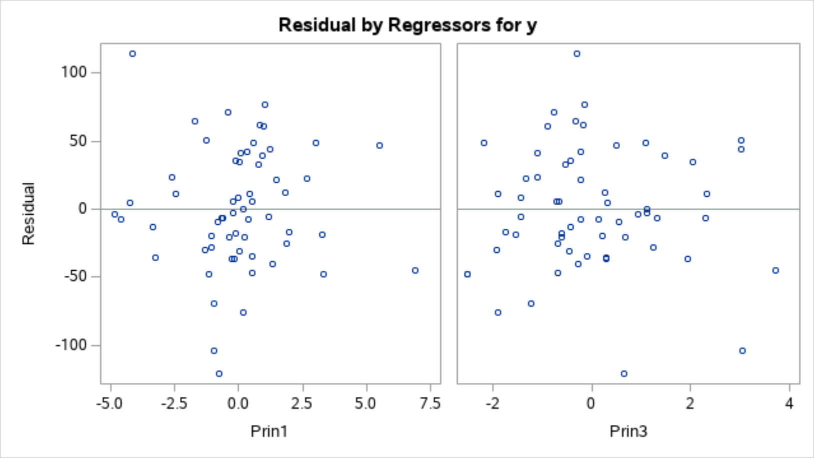
Number in Model	C(p)	R-Square	Variables in Model
2	1.1328	0.4931	Prin1 Prin3
3	2.2941	0.5008	Prin1 Prin2 Prin3
3	2.8565	0.4956	Prin1 Prin3 Prin4
3	3.1151	0.4932	Prin1 Prin3 Prin5
4	4.0178	0.5033	Prin1 Prin2 Prin3 Prin4
4	4.2763	0.5010	Prin1 Prin2 Prin3 Prin5
4	4.8387	0.4958	Prin1 Prin3 Prin4 Prin5
5	6.0000	0.5035	Prin1 Prin2 Prin3 Prin4 Prin5
1	21.8357	0.2843	Prin1
2	22.9969	0.2921	Prin1 Prin2
2	23.5593	0.2869	Prin1 Prin4
2	23.8179	0.2845	Prin1 Prin5
3	24.7206	0.2946	Prin1 Prin2 Prin4
3	24.9792	0.2922	Prin1 Prin2 Prin5
3	25.5416	0.2870	Prin1 Prin4 Prin5
4	26.7028	0.2948	Prin1 Prin2 Prin4 Prin5
1	30.0581	0.2087	Prin3
2	31.2193	0.2165	Prin2 Prin3
2	31.7817	0.2113	Prin3 Prin4
2	32.0403	0.2089	Prin3 Prin5
3	32.9430	0.2190	Prin2 Prin3 Prin4
3	33.2015	0.2166	Prin2 Prin3 Prin5
3	33.7640	0.2114	Prin3 Prin4 Prin5

Number in Model	C(p)	R-Square	Variables in Model
4	34.9252	0.2192	Prin2 Prin3 Prin4 Prin5
1	51.9221	0.0077	Prin2
1	52.4845	0.0025	Prin4
1	52.7431	0.0002	Prin5
2	53.6458	0.0103	Prin2 Prin4
2	53.9044	0.0079	Prin2 Prin5
2	54.4668	0.0027	Prin4 Prin5
3	55.6280	0.0104	Prin2 Prin4 Prin5

Exercise 4
Regression moddel with CP criteria

The REG Procedure
Model: MODEL1
Dependent Variable: y





```
footnote2 j = r height= 8pt italic "Huong Tran - Assignment 3";
*** import data from txt file***;
data air_pollution;
infile '/home/u59404828/sasuser.v94/STA5221/HW3/airpollution.txt' delimiter="," firstobs=2;
input  x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 y;
run;

proc print data = air_pollution (obs=10);
title "Exercise 4";
title2 "First 10 observations of data";

proc reg data = air_pollution;
model y = x1-x15 / vif;
title2 "Regression model of air pollution";
run;

proc princomp data = air_pollution out=airdata;
var x1-x15;
title2 "PCA - Air pollution";
run;


proc reg data = airdata;
model y = Prin1;
title2 "Regression model by first PC";
run;

proc reg data = airdata;
model y = prin1 prin2;
title2 "Regression model by first 2 PC's";
run;

proc reg data = airdata;
model y = prin1 prin2 prin3;
title2 "Regression model by first 3 PC's";
run;

proc reg data = airdata;
model y = prin1 prin2 prin3 prin4;
title2 "Regression model by first 4 PC's";
run;
proc reg data = airdata;
model y = prin1 prin2 prin3 prin4 prin5;
title2 "Regression model by first 5 PC's";
run;

data RootMSE;
input Model RootMSE;
lines;
1  53.07647
2  53.25078
3  45.11376
4  45.40604
5  45.81701
;
proc print data = RootMSE;
title2 "Root Square Mean of each model corresponding upto the first 5 PC";
run;

proc reg data = airdata;
model y = prin1 prin2 prin3 prin4 prin5 / selection=cp ;
title2 "Regression moddel with CP criteria";
run;
```