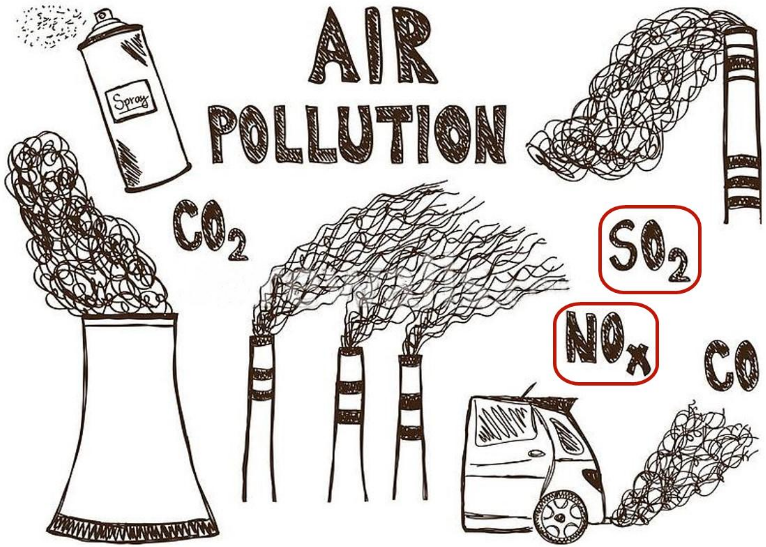
Regression Analysis of Air Pollution and Mortality

Yahui Peng
STA-6013 Regression Analysis

Objective

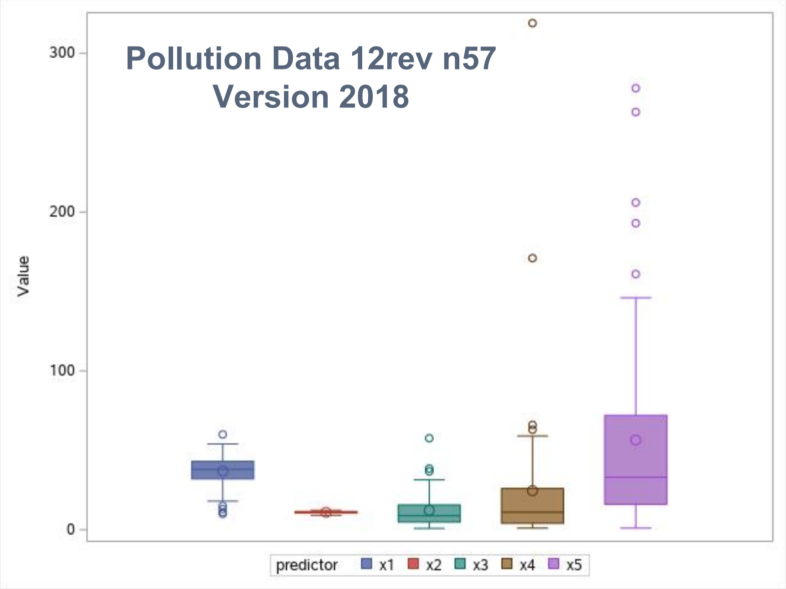


- Air pollution can contribute to climate change while the effects of climate change can contribute to and exacerbate the health risks posed by air pollution.

- Seek to obtain a prediction equation for mortality as a function of the predictor variables.

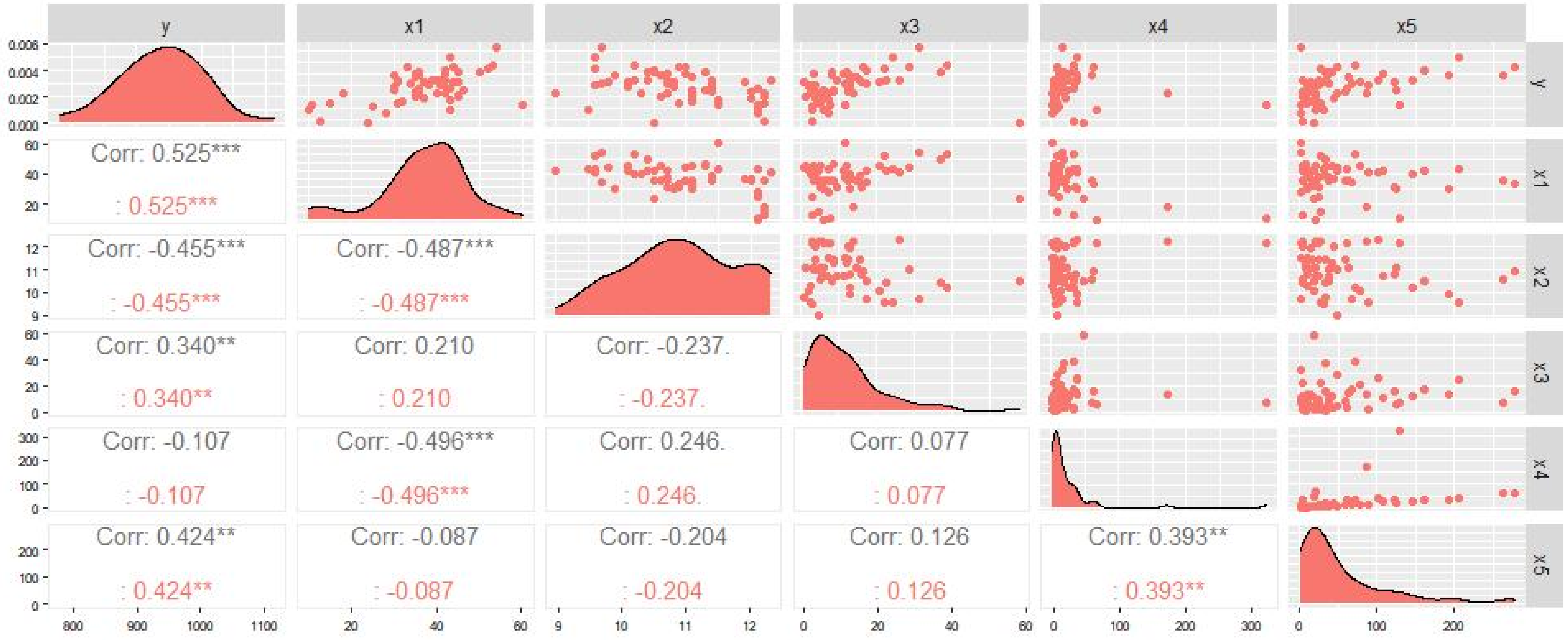
Data Descriptions

Obs	City	y	x1	x2	x3	x4	x5
1	San Jose, CA	790.73	13	12.2	3	32	3
2	Wichita, KS	823.76	28	12.1	7.5	2	1
3	San Diego, CA	839.71	10	12.1	5.9	66	20
4	Lancaster, PA	844.05	43	9.5	2.9	7	32
5	Minneapolis, MN	857.62	25	12.1	3	11	26
52	Chicago, IL	1024.89	33	10.9	16.3	63	278
53	Richmond, VA	1025.5	44	11	28.6	9	48
54	Birmingham, AL	1030.38	53	10.2	38.5	32	72
55	Baltimore, MD	1071.29	43	9.6	24.4	38	206
56	New Orleans, LA	1113.06	54	9.7	31.4	17	1
57	San Antonio, TX	782.37	24	10.5	57.6	48	19

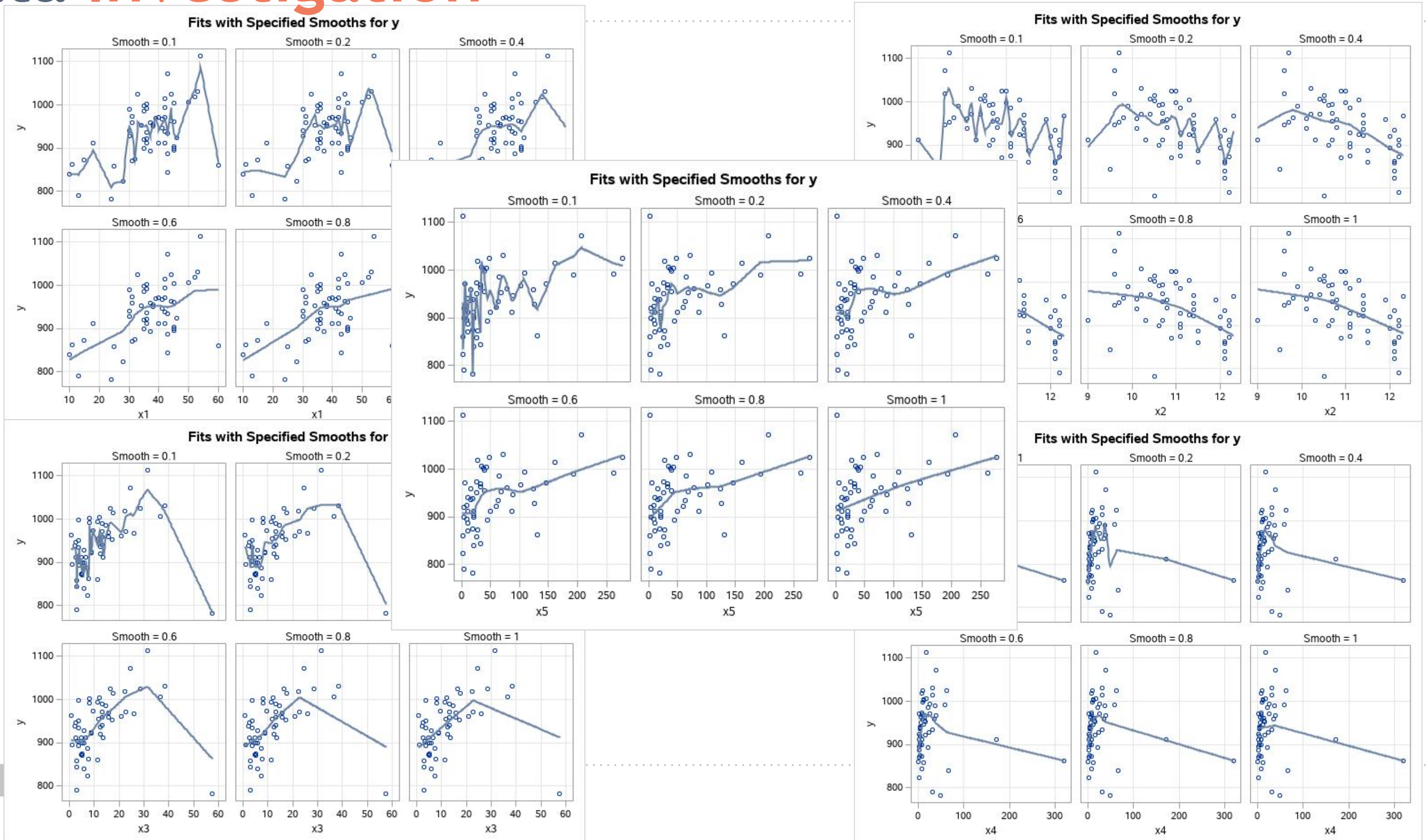


y = Total age-adjusted mortality from all causes in deaths per 100,000 population
x_1 = Mean annual precipitation (in inches)
x_2 = Median number of school years completed for persons of age 25 years or older
x_3 = Percentage of the population that is nonwhite
x_4 = Relative pollution potential of oxides of nitrogen
x_5 = Relative pollution potential of sulfur dioxide
Note: Relative pollution potential is the product of the tons emitted per day per square kilometer and a factor correcting for the dimensions of and exposure to the given area.

Data Investigation



Data Investigation

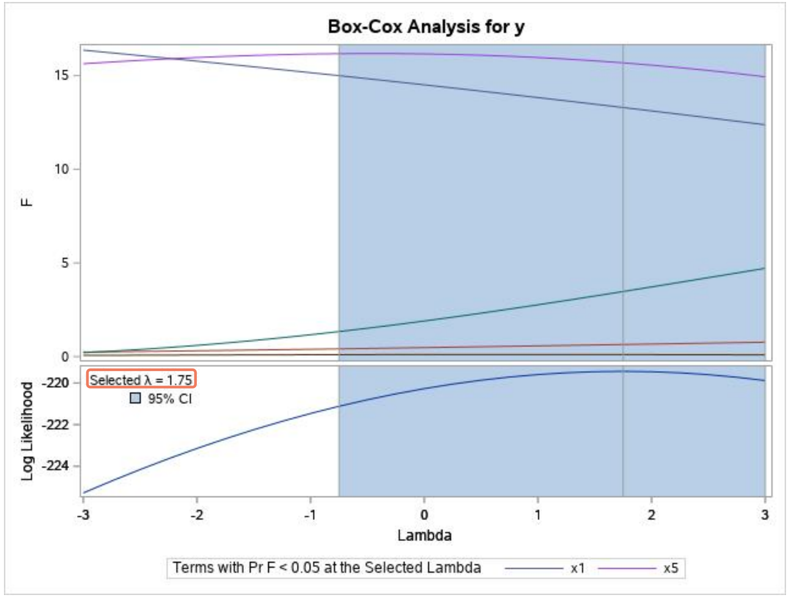


MLR Model Specification

$$y_i^2 = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, \quad i = 1, \ldots, n$$

Ordinary Least Squares (OLS) Regression Model Assumptions:

- $\beta_j \neq 0$ for at least one $j = 1, \ldots, 5$
- $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$
- No or little multicollinearity



Model Estimation

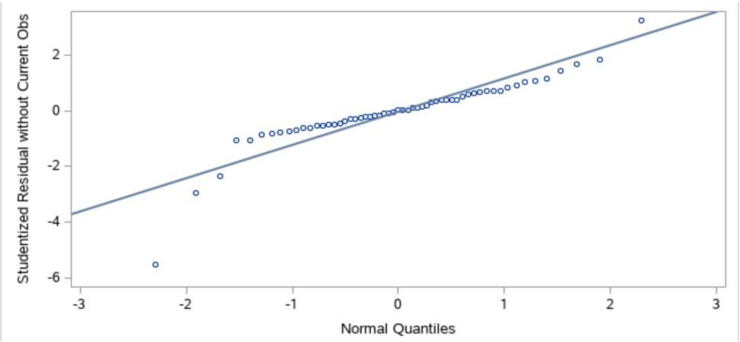
$$\hat{y}_2 = 139629 + 866.83842x_1 - 2122.80478x_2 + 343.27633x_3 - 16.63512x_4 + 133.20095x_5$$

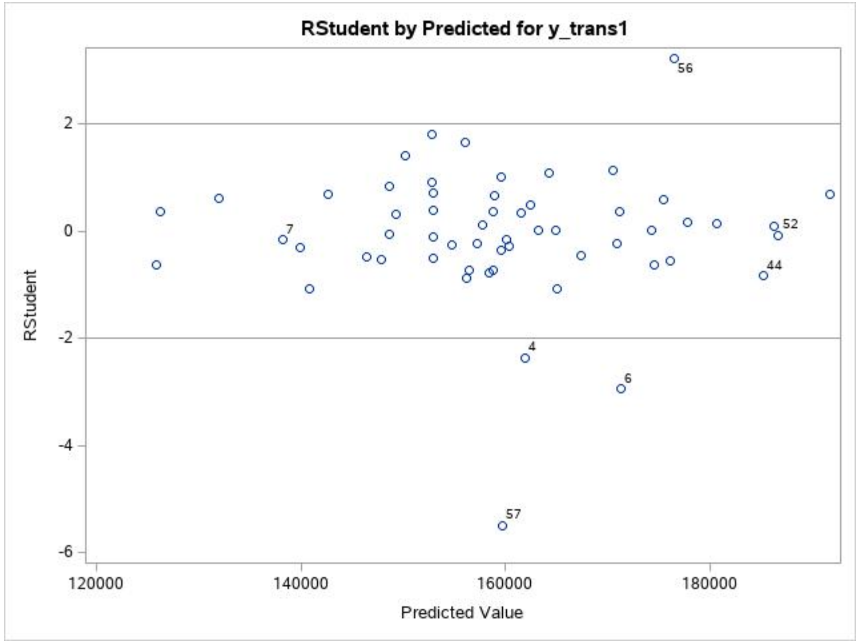
Parameter Estimates										
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	Intercept	1	139629	33991	4.11	0.0001	0	0	71389	207870
x1	x1	1	866.83842	237.71280	3.65	0.0006	0.45383	1.70211	389.61006	1344.06678
x2	x2	1	-2122.80478	2642.57367	-0.80	0.4255	-0.09340	1.48544	-7427.99279	3182.38322
x3	x3	1	343.27633	184.15581	1.86	0.0681	0.18870	1.12619	-26.43189	712.98455
x4	x4	1	-16.63512	50.81421	-0.33	0.7447	-0.04043	1.67640	-118.64890	85.37865
x5	x5	1	133.20095	33.63467	3.96	0.0002	0.43643	1.33460	65.67653	200.72537

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	11413879112	2282775822	11.78	<.0001
Error	51	9884331014	193810412		
Corrected Total	56	21298210126			

Root MSE	13922	**R-Square**	0.5359
Dependent Mean	159755	**Adj R-Sq**	0.4904
Coeff Var	8.71432		

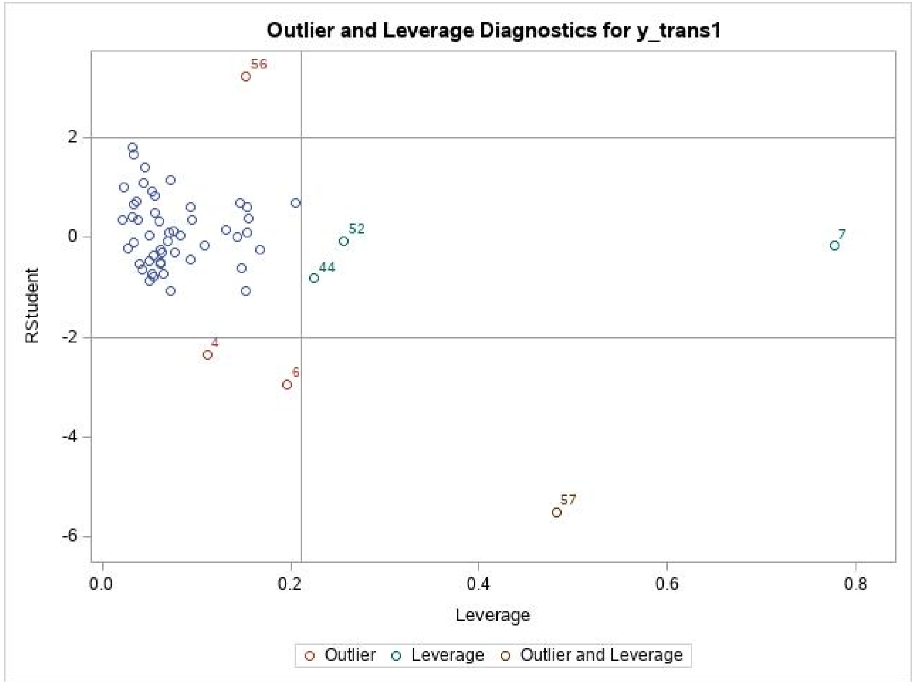
Model Adequacy Checking

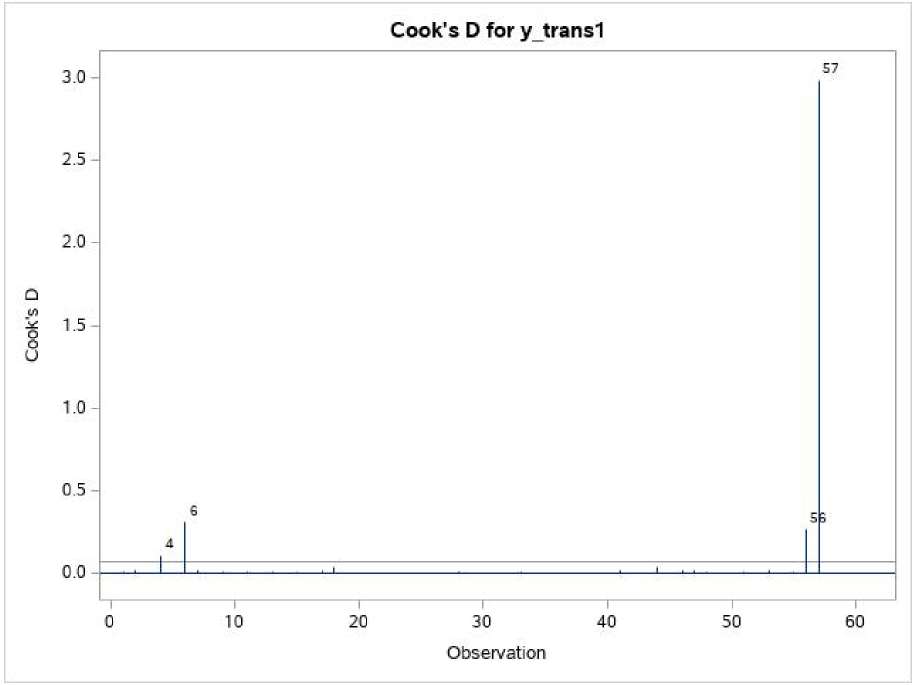




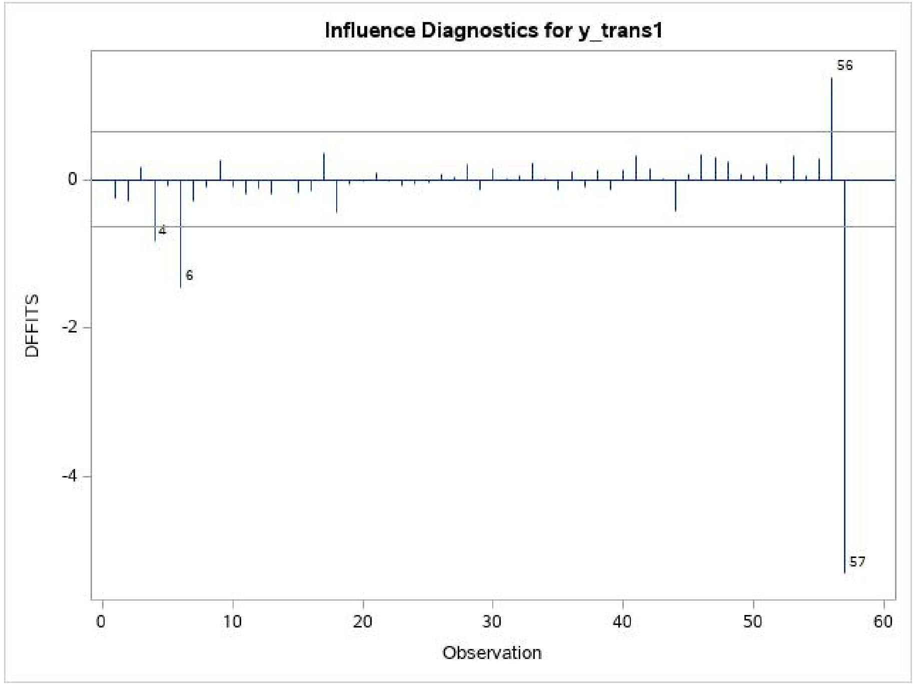
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.84934	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.154373	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.321018	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.054926	Pr > A-Sq	<0.0050

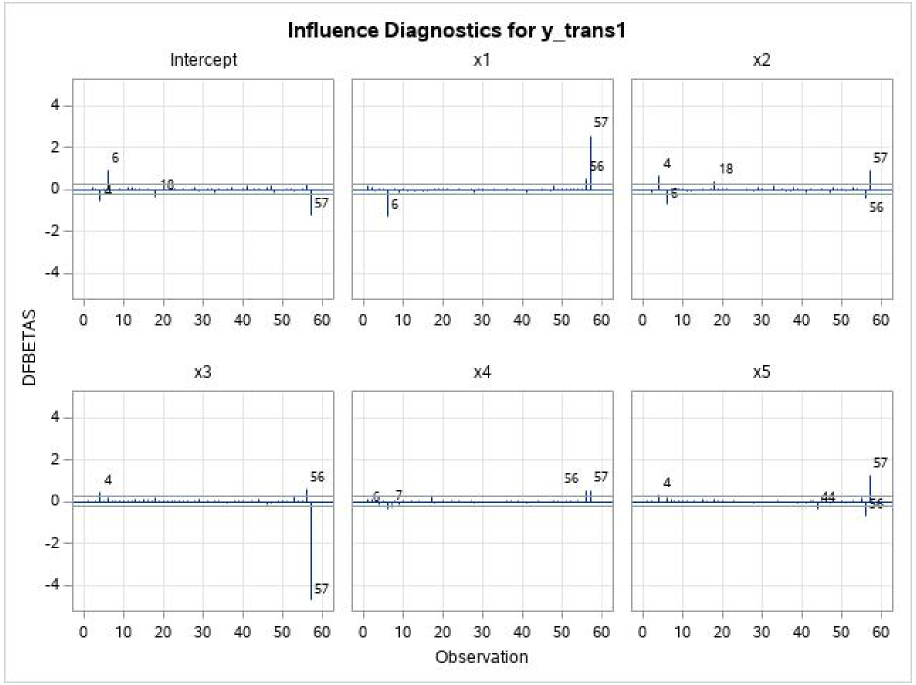
Model Diagnostics





Model Diagnostics



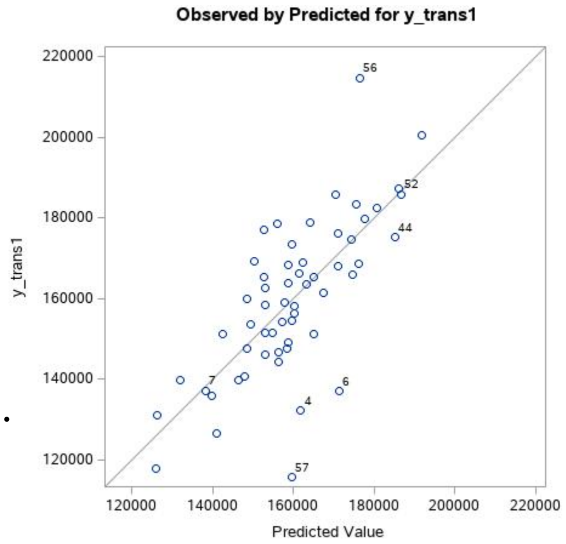


Influence Analysis

Point 57 is clearly influential

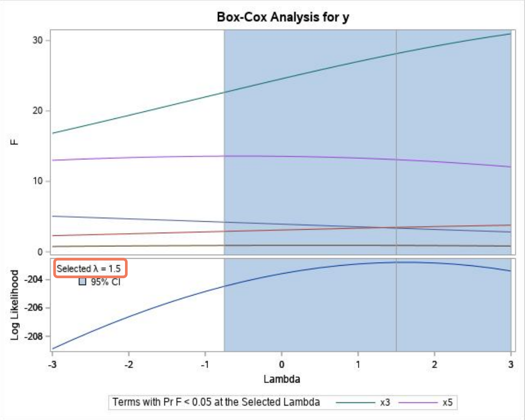
Point 56 has effect on $\widehat{\beta}s$, $\widehat{Y}_i$ and $COVRATIO_i$

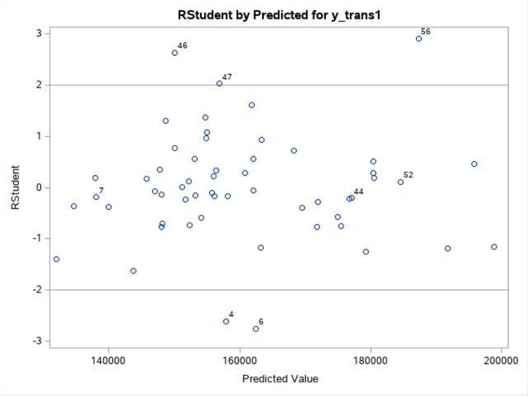
Point 4, 6 have moderate effect on $\widehat{\beta}s$ and $\widehat{Y}_i$



- $\boldsymbol{Rstudent\ Residual}$: cutoff is 3. Point 57 is an outlier.
- $\boldsymbol{h_{ii}}$: cutoff is 2p/n=0.2105. Point 7, 44, 52 and 57 are leverage points.
- $\boldsymbol{h_{ii}}$ & $\boldsymbol{Rstudent\ Residual}$: point 57 is likely influential.
- ***Cook'sD*** : cutoff is 1. Point 57 is influential.
- ***DFFITS***$_i$: cutoff is $2\sqrt{p/n} = 0.5923$. Point 4, 6, 56, and 57 are most likely influential.
- ***DFBETAS***$_i$: cutoff is $2\sqrt{n} = 0.2649$. Effect: 57(on all $\widehat{\beta}s$, especially $\widehat{\beta}_{1,\ 3}$), 56 (on $\widehat{\beta}_{1\text{-}5}$), 4 (on $\widehat{\beta}_{0,2,3,5}$, especially $\widehat{\beta}_2$), 6 (on $\widehat{\beta}_{0,1,2,4}$, especially $\widehat{\beta}_{0,1}$). Small effect: 7 (on $\widehat{\beta}_4$), 18 (on $\widehat{\beta}_2$), 44 (on $\widehat{\beta}_5$).
- $\boldsymbol{COVRATIO_i}$: cutoff is $1 \pm \frac{3p}{n}$, or 1.316 and 0.684. Point 6, 8, 15, 52, 56, and 57 are influential.

Refit after Discarding Point 57





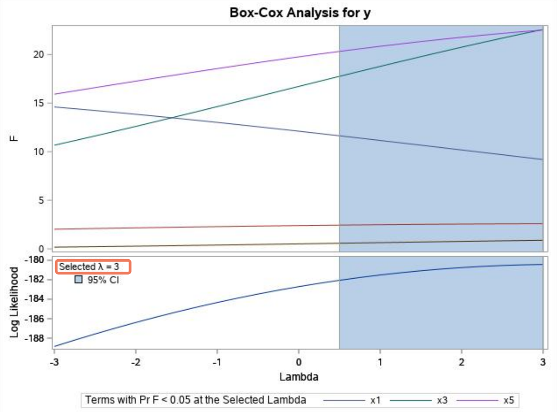
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.963682	Pr < W	0.0898
Kolmogorov-Smirnov	D	0.099041	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.129012	Pr > W-Sq	0.0453
Anderson-Darling	A-Sq	0.760501	Pr > A-Sq	0.0460

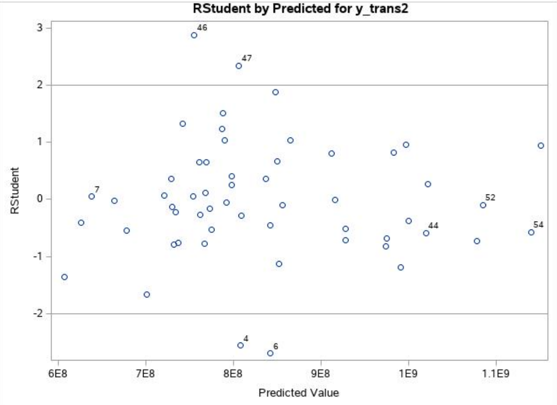
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	314016453	62803291	21.27	<.0001
Error	50	147645256	2952905		
Corrected Total	55	461661709			

Root MSE	1718.40191	**R-Square**	0.6802
Dependent Mean	28945	**Adj R-Sq**	0.6482
Coeff Var	5.93675		

Parameter Estimates										
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	Intercept	1	30933	4304.94869	7.19	<.0001	0	0	22287	39580
x1	x1	1	59.45619	32.40188	1.83	0.0725	0.20831	2.01480	-5.62491	124.53729
x2	x2	1	-616.32518	330.50150	-1.86	0.0681	-0.18375	1.51795	-1280.15697	47.50662
x3	x3	1	158.99476	29.96653	5.31	<.0001	0.48729	1.31872	98.80522	219.18430
x4	x4	1	-6.01188	6.30271	-0.95	0.3447	-0.09903	1.68523	-18.67124	6.64748
x5	x5	1	15.42109	4.26126	3.62	0.0007	0.34211	1.39720	6.86210	23.98009

Refit after Discarding Point 56 and 57





Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.972292	Pr < W	0.2329
Kolmogorov-Smirnov	D	0.108408	Pr > D	0.1042
Cramer-von Mises	W-Sq	0.088981	Pr > W-Sq	0.1571
Anderson-Darling	A-Sq	0.563653	Pr > A-Sq	0.1423

Analysis of Variance

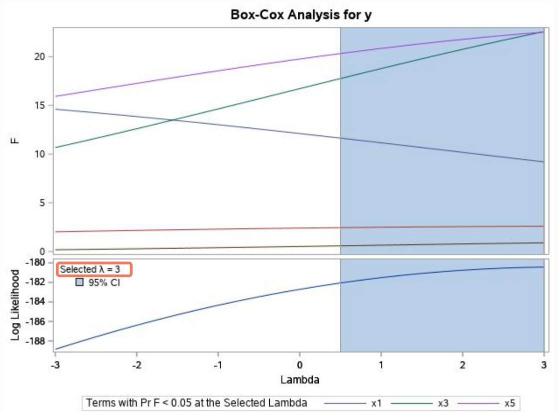
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8.920817E17	1.784163E17	21.78	<.0001
Error	49	4.014492E17	8.19284E15		
Corrected Total	54	1.293531E18			

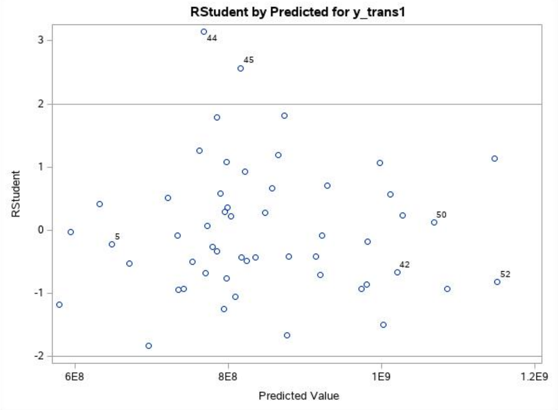
Root MSE	90514308	**R-Square**	0.6896
Dependent Mean	836377214	**Adj R-Sq**	0.6580
Coeff Var	10.82219		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	890606490	228415281	3.90	0.0003	0	0	431588797	1349624183
x1	x1	1	2839689	1709344	1.66	0.1030	0.18312	1.91847	-595367	6274744
x2	x2	1	-27300141	17651015	-1.55	0.1284	-0.15080	1.50087	-62771184	8170902
x3	x3	1	8221737	1635466	5.03	<.0001	0.45306	1.28238	4935145	11508329
x4	x4	1	-487358	334945	-1.46	0.1520	-0.15163	1.71468	-1160454	185739
x5	x5	1	1060602	231582	4.58	<.0001	0.44136	1.46634	595220	1525983

Refit after Discarding Point 4, 6, 56 and 57





Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.958339	Pr < W	0.0621
Kolmogorov-Smirnov	D	0.090183	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.088039	Pr > W-Sq	0.1625
Anderson-Darling	A-Sq	0.575207	Pr > A-Sq	0.1337

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8.967591E17	1.793518E17	28.18	<.0001
Error	47	2.991384E17	6.364647E15		
Corrected Total	52	1.195897E18			

Root MSE	79778738	R-Square	0.7499
Dependent Mean	844531545	Adj R-Sq	0.7233
Coeff Var	9.44651		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	824467013	218027713	3.78	0.0004	0	0	385851829	1263082197
x1	1	5068250	1671590	3.03	0.0039	0.32082	2.10374	1705444	8431055
x2	1	-26715005	16604544	-1.61	0.1143	-0.14864	1.60372	-60119038	6689028
x3	1	7002771	1473748	4.75	<.0001	0.39789	1.31753	4037972	9967570
x4	1	-282011	299925	-0.94	0.3519	-0.09094	1.75744	-885381	321360
x5	1	975121	205470	4.75	<.0001	0.41817	1.45881	561768	1388474

WLS Regression Overview

- Apart from the main function of Weighted Least Squares (WLS) regression in correcting for non-constant variance of residual error, it is sometimes also used to adjust fit to give less weight to distant points and outliers, or to give less weight to observations thought to be less reliable.

- A model is supposed to be as informative as possible, discarding an outlier or influential point is not preferred or recommended.

- We try to accommodate it, down-weighting it to a less impactful point, or almost to zero, by using WLS regression.

WLS Model Estimation

Weight: wt2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	349.11200	69.82240	29.26	<.0001
Error	51	121.70630	2.38640		
Corrected Total	56	470.81830			

Root MSE	1.54480	R-Square	0.7415
Dependent Mean	934.69189	Adj R-Sq	0.7162
Coeff Var	0.16527		

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	823.10440	50.03136	16.45	<.0001	0	0	722.66226	923.54654
x1	x1	1	2.41085	0.48239	5.00	<.0001	0.61271	2.96529	1.44242	3.37929
x2	x2	1	-1.85437	4.28649	-0.43	0.6671	-0.04280	1.93109	-10.45986	6.75112
x3	x3	1	2.15583	0.77624	2.78	0.0076	0.28730	2.11127	0.59747	3.71419
x4	x4	1	-0.11054	0.06868	-1.61	0.1137	-0.19447	2.88061	-0.24842	0.02735
x5	x5	1	0.42484	0.06398	6.64	<.0001	0.65790	1.93689	0.29639	0.55329

Output Statistics												
Obs	Weight	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	9.90E-04	791	836.0270	15.2658	805.3795	866.6744	732.8233	939.2307	-45.2970	46.654	-0.971	0.017
2	9.91E-04	824	884.5429	10.6461	863.1700	905.9157	783.7501	985.3357	-60.7829	47.895	-1.269	0.013
3	8.04E-04	840	838.6958	14.9088	808.7651	868.6264	725.3140	952.0775	1.0142	52.393	0.019	0.000
4	1.59E-03	844	928.2275	11.0597	906.0242	950.4308	847.3407	1009	-84.1775	37.131	-2.267	0.076
5	2.54E-03	858	877.2352	10.1033	856.9519	897.5185	812.4088	942.0616	-19.6152	28.958	-0.677	0.009
6	2.10E-03	861	971.5367	11.3160	948.8188	994.2546	900.1505	1043	-110.0967	31.753	-3.467	0.254
7	1.67E-02	862	863.9691	11.8027	840.2741	887.6641	830.2280	897.7101	-2.1391	1.964	-1.089	7.134
8	9.75E-04	871	892.5397	8.1221	876.2339	908.8455	791.9117	993.1677	-21.1997	48.790	-0.435	0.001
9	1.22E-03	872	857.7874	15.2249	827.2222	888.3526	763.9037	951.6711	13.9826	41.513	0.337	0.003
10	1.28E-03	874	897.6523	7.2586	883.0800	912.2245	809.8050	985.4995	-23.3723	42.537	-0.549	0.001
11	1.82E-03	887	924.8846	5.3632	914.1174	935.6517	851.4655	998.3036	-37.4146	35.776	-1.046	0.004
12	1.14E-02	894	914.1861	3.5788	907.0013	921.3708	884.2459	944.1262	-20.1961	14.028	-1.440	0.022
13	5.05E-03	896	916.2322	5.7698	904.6488	927.8156	871.0923	961.3721	-20.5322	20.952	-0.980	0.012
14	2.85E-03	899	904.8722	6.5898	891.6426	918.1017	845.3052	964.4391	-5.6122	28.169	-0.199	0.000
15	1.44E-03	900	924.6196	7.3945	909.7745	939.4646	841.4360	1008	-25.0896	40.093	-0.626	0.002
16	4.01E-03	904	926.3937	4.9523	916.4516	936.3358	876.4175	976.3699	-22.2337	23.888	-0.931	0.006
17	1.44E-03	912	891.0456	9.9382	871.0937	910.9974	806.9576	975.1336	20.6544	39.457	0.523	0.003
18	1.18E-03	912	937.9516	12.2768	913.3048	962.5984	844.4587	1031	-26.1316	43.212	-0.605	0.005
19	1.56E-03	912	910.2802	8.2098	893.7984	926.7620	830.0146	990.5458	1.9198	38.258	0.050	0.000
20	1.06E-03	912	912.2204	6.8095	898.5498	925.8910	815.9837	1008	0.1296	46.959	0.003	0.000
21	7.56E-04	920	913.6525	10.5306	892.5114	934.7935	798.9002	1028	6.0775	55.185	0.110	0.000
22	2.33E-03	922	931.1340	4.9179	921.2609	941.0070	866.0782	996.1898	-9.2640	31.650	-0.293	0.000
23	1.46E-03	923	935.0876	6.5815	921.8747	948.3006	852.7668	1017	-11.8576	39.935	-0.297	0.000
24	2.18E-02	929	937.9126	6.3322	925.2002	950.6250	913.3490	962.4763	-8.7626	8.337	-1.051	0.106
25	2.67E+01	935	934.6813	0.2990	934.0810	935.2817	933.8322	935.5305	0.0187	0.00762	2.447	1534.921
26	7.29E-04	936	921.8428	8.9541	903.8667	939.8190	805.6160	1038	14.3872	56.492	0.255	0.000
27	1.84E-03	939	917.5788	8.9269	899.6572	935.5004	843.1605	991.9972	20.9212	34.853	0.600	0.004
28	4.38E-04	941	907.8752	10.7012	886.3917	929.3588	758.1163	1058	33.3048	73.045	0.456	0.001
29	6.18E-03	946	946.5628	10.1118	926.2624	966.8632	902.1897	990.9359	-0.3828	16.853	-0.023	0.000
30	3.10E-03	951	911.8374	4.6304	902.5416	921.1332	855.3680	968.3068	38.8326	27.355	1.420	0.010
31	4.39E-04	954	951.3972	10.3911	930.5363	972.2582	801.8856	1101	2.1628	73.009	0.030	0.000
32	8.56E-04	954	935.2300	7.2810	920.6128	949.8471	828.2181	1042	19.2100	52.300	0.367	0.000

Feature Selection

Weight: wt2

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x5		x5	1	0.2545	0.2545	94.0797	18.78	<.0001
2	x1		x1	2	0.3988	0.6533	17.4088	62.10	<.0001
3	x3		x3	3	0.0662	0.7195	6.3446	12.51	0.0009
4	x4		x4	4	0.0211	0.7406	4.1872	4.22	0.0449

Number of Observations Read	57
Number of Observations Used	57

Weight: wt2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	348.66539	87.16635	37.11	<.0001
Error	52	122.15292	2.34909		
Corrected Total	56	470.81830			

Root MSE	1.53268	R-Square	0.7406
Dependent Mean	934.69189	Adj R-Sq	0.7206
Coeff Var	0.16398		

C(p) Selection Method

Number of Observations Read	57
Number of Observations Used	57

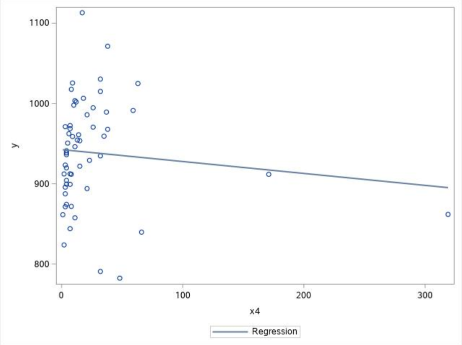
Weight: wt2

Number in Model	C(p)	R-Square	Adjusted R-Square	MSE	Variables in Model
4	4.1872	0.7406	0.7206	2.34909	x1 x3 x4 x5
5	6.0000	0.7415	0.7162	2.38640	x1 x2 x3 x4 x5
3	6.3446	0.7195	0.7036	2.49197	x1 x3 x5
4	6.5901	0.7284	0.7075	2.45937	x1 x2 x3 x5
4	11.7133	0.7024	0.6795	2.69449	x1 x2 x4 x5
3	11.9304	0.6912	0.6737	2.74347	x1 x2 x5
3	12.5085	0.6882	0.6706	2.76950	x1 x4 x5
2	17.4088	0.6533	0.6404	3.02316	x1 x5
4	28.9774	0.6149	0.5853	3.48677	x2 x3 x4 x5
3	29.2262	0.6035	0.5811	3.52224	x3 x4 x5

Conclusion

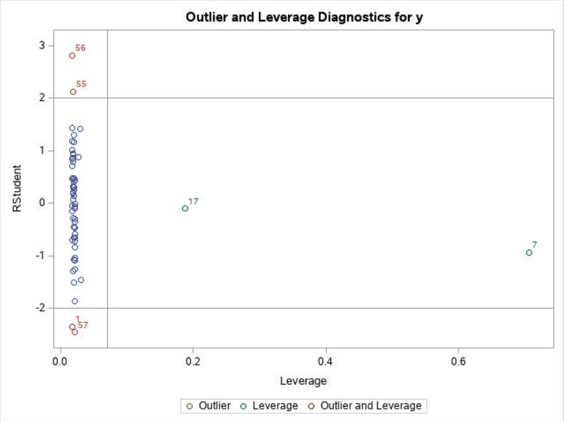
Final Prediction Equation for WLS Regression

$$\widehat{Mort} = 803.6445 + 2.3345Precep + 2.2891Nonwhite - 0.1243NO_X + 0.4294SO_2$$



Positive relationship with mortality
- annual precipitation (x_1)
- non-white percentage (x_3)
- relative pollution potential of SO_2 (x_5)

Negative relationship with mortality
- relative pollution potential of NOx (x_4)



Comment
- Are the outliers in y vs x_4 model valid?
- If not, refit after discarding them.

Thank You
For Your Attention

Any
Questions?