Regression Analysis of Air Pollution and Mortality

STA-6013 Regression Analysis

Yahui Peng @ifg583

Date: 11/29/2021

1. Problem Description

The objective of this project is to seek to find the association between air pollution and mortality as well as to obtain a prediction equation for mortality as a function of the predictor variables. The dataset used is named 'Pollution Data 12rev n57 Version 2018' with all numeric data collected from 57 US cities. Below is the variable description and the header of the dataset:

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.

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

We follow the $PISEAS$ (Planning, Investigating, Specification, Estimation, Assess, and Selection) rule, a general system for performing a regression analysis except for 'P' since the data is designated here. MLR (multiple linear regression) is started with the most common OLS (ordinary least square) model. After investigation of the data, an increased power transformation on y is conducted based on Box-Cox test and estimate parameters are then obtained for specification of the fitted model. However, MSE (mean squared error) and R^2_{Adj} of the model are not desirable. Also, adequacy checking using diagnostic measures indicate the assumptions on error term do not hold well. To improve the fit, possible outliers and influential points are identified by influential analyses. After removing one outlier, the model is refitted, making the assumptions on error term hold while MSE and R^2_{Adj} do not improve much. Another transformation on y did not help much, likewise. Hence, we conclude that OLS model does not perform well on the data.

Since a model is supposed to be as informative as possible, deleting an outlier is not preferred or recommended. We try to accommodate it, such as downweighting it to a less impactful point, or almost to zero, by using WLS (weighted least square) model. Thus, the WLS model is fitted instead of the OLS one. To our delight, MSE, R^2_{Adj} and even CV (coefficient variance) improved significantly with all observations kept. Finally, variable selection measures were performed to obtain the optimal model, and the prediction equation is

$$\hat{y} = 803.64449 + 2.33451x_1 + 2.28908x_3 - 0.12425x_4 + 0.42937x_5$$

We conclude $Mortality$ is positively associated with mean annual precipitation, percentage of the population that is nonwhite, relative pollution potential of sulfur dioxide, and is negatively related to relative pollution potential of oxides of nitrogen.

Fortunately, multicollinearity was not observed throughout the whole regression analysis.

2. Investigation of the Data

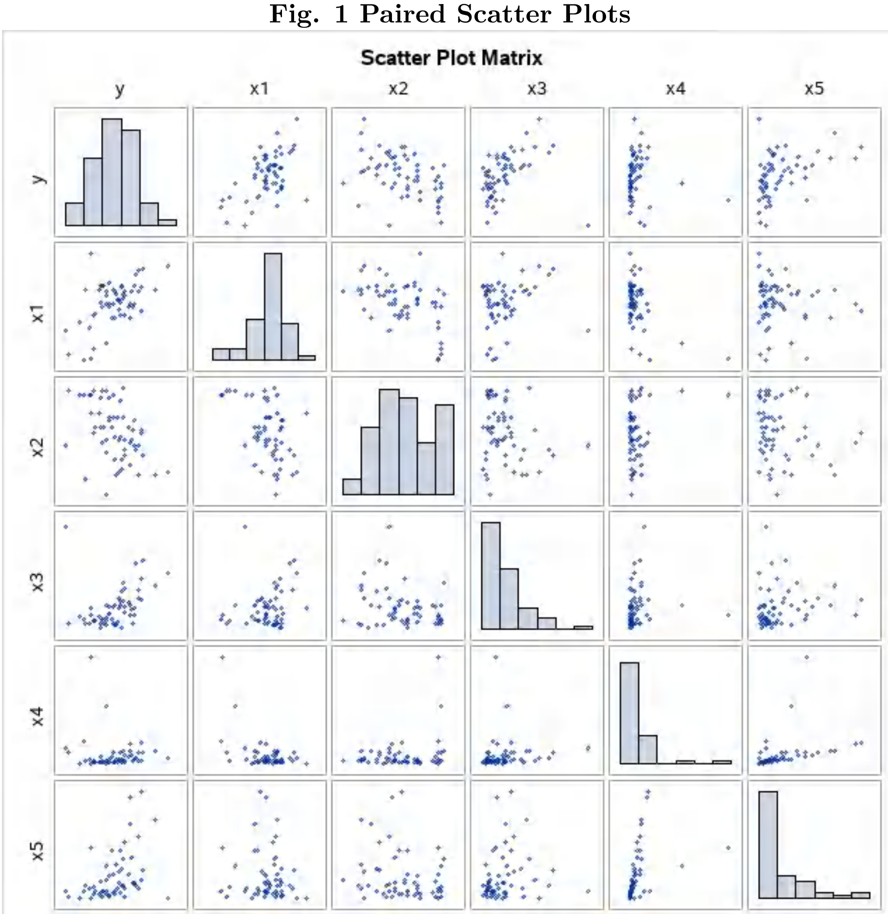
The summary statistics for response and all predictors are shown in Table 1. x_3, x_4 and x_5 show large standard deviance compared to the mean.

Table 1 Simple Summary Statistics

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
y	57	938.89702	65.79181	53517	782.37000	1113	y
x1	57	36.96491	10.21023	2107	10.00000	60.00000	x1
x2	57	10.93333	0.85802	623.20000	9.00000	12.30000	x2
x3	57	12.18246	10.72048	694.40000	0.80000	57.60000	x3
x4	57	24.59649	47.40210	1402	1.00000	319.00000	x4
x5	57	56.45614	63.89726	3218	1.00000	278.00000	x5

The pairwise scatter plots are shown in Fig.1, suggesting y is positively associated with x_1,x_3 and x_5, negatively associated with x_2, and not obviously related to x_4. However, considering the two leverage points in the plot of y vs x_4, there might be association between the two variables. Moreover, there appear to no or little linear pairwise association among the five predictors.

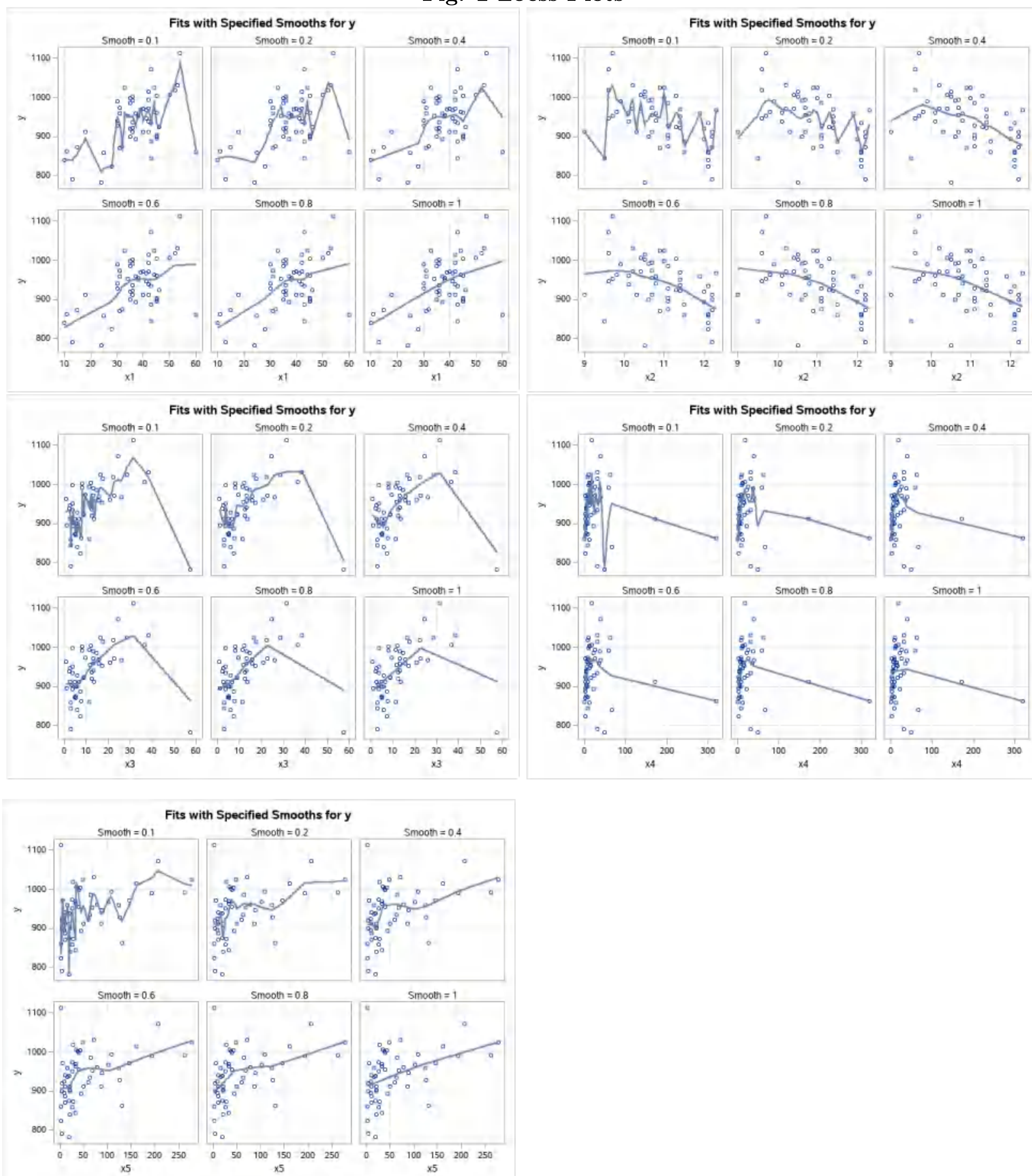
Fig. 1 Paired Scatter Plots



Loess plots are obtained for all predictors using various smoothing parameters of 0.1, 0.2, 0.4, 0.6, 0.8 and 1. As seen in Fig.2, all predictors achieve the best Loess fit with *smooth* equals 0.6 or 0.8. However, the strong nonlinear trend with x_3 and x_4 suggests transformation of the original data, commonly on the power of response. To obtain the appropriate λ for y in the transformed form y^λ, Box-Cox analysis is recommended.

Observation 7, 17, 57 should be carefully scrutinized as possible outliers since they have a large x_4 or x_3 value and are off the trend of the majority data.

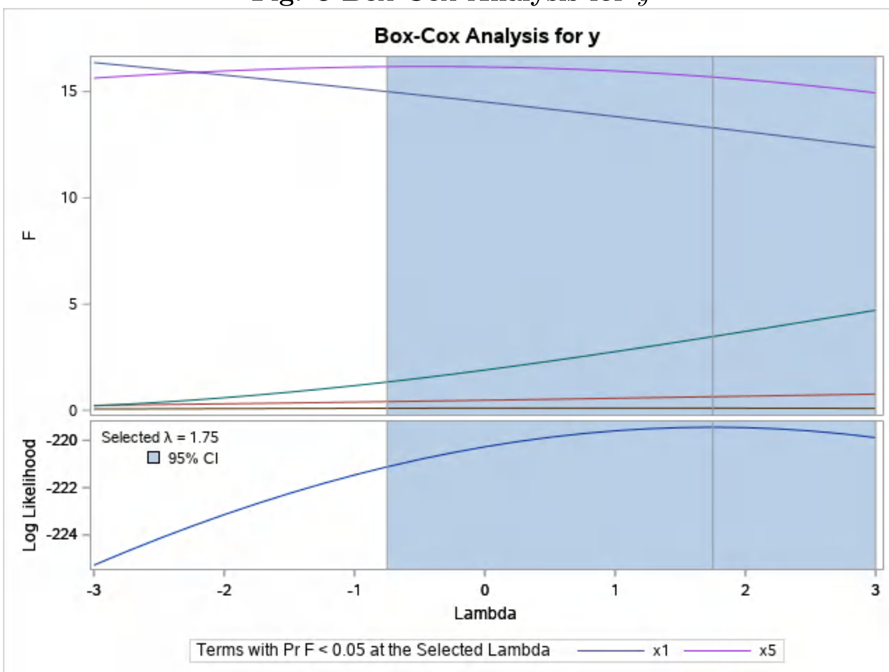
Fig. 2 Loess Plots



3. Specification of the Model

Base on preliminary judgments from the Investigation stage in Section 2 above, Box-Cox analysis suggests $y^{1.75}$ is a good transformation of the response. To simplify the prediction equation, we take value of 2, the nearest integer around 1.75, to be the proper order transformation on the response variable.

Fig. 3 Box-Cox Analysis for y



Therefore, the initial proposed MLR model is

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

For ith observation, y_i^2 is the response variable, all $x_{ij}s$ are predictors for $j = 1, 2, 3, 4, 5$, and ϵ_i is the error term. $\beta_{ij}s$ are parameters where β_{0j} is the response value when all predictors are set to zero, and other $\beta_{ij}s$ represent the change in y_i per unit change in x_{ij}.

The model assumptions are:
1. Linear relationship between response and predictors in coefficient β_{ij}, or $\beta_{ij} \neq 0$ for at least one j
2. Error term ϵ_i follows $N(0, \sigma^2)$. In other words, the probability distribution of ϵ_i shows normality and constant variance.
3. No or light multicollinearity among all predictors x_{ij}.

4. Estimation of the Appropriate Model

According to Table 2, the prediction equation is obtained with the parameter estimates as

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

The ANOVA analysis result given in Table 2 shows the p-value for F-test is smaller than 0.0001, indicating rejection of the H_0 that all $\beta_{ij}s = 0$. The assumption (1) in Section 3 holds. A good CV of 8.71% shows the ratio of standard deviation to mean is low, suggesting a low variability in the data. However, a R^2_{Adj} of 49.04% means less than 50% of the variability in the data can be explained by the model. Also, the MSE is 13922^2, which is extremely large.

Table 2 ANOVA Analysis for the Initial Model

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		

Read from the parameter estimates Table 3, the p-values obtained for x_2 and x_4 in t-test are greater than 0.05, indicating insignificant of these two variables at 0.05 level. The absolute value of $\hat{\beta}_2$ is extremely large while that of $\hat{\beta}_4$ is too small in magnitude compared with other predictor coefficients, suggesting the model does not fit well.

All VIFs (variance inflation factors) are smaller than 2, meaning only moderate correlations are observed among the predictors. Therefore, the assumption (3) in Section 3 holds. Multicollinearity is not a concern for the data.

Table 3 Parameter Estimates for the Initial Model

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	184.15681	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	-118.64690	85.37865
x5	x5	1	133.20095	33.63467	3.96	0.0002	0.43643	1.33460	65.67653	200.72537

Also, the collinearity diagnostics table with intercept adjusted in Table 4 further confirms it as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{1.90269}{0.37954} = 1.32 < 10$$

Table 4 Collinearity Diagnostics (intercept adjusted)

Collinearity Diagnostics (intercept adjusted)							
			Proportion of Variation				
Number	Eigenvalue	Condition Index	x1	x2	x3	x4	x5
1	1.90269	1.00000	0.12208	0.09084	0.02132	0.08798	0.00850
2	1.44511	1.14745	0.00071289	0.06619	0.14254	0.07239	0.23271
3	0.83232	1.51195	0.00146	0.08551	0.71547	0.00625	0.15161
4	0.44034	2.07869	0.33870	0.75730	0.00188	0.04940	0.36311
5	0.37954	2.23899	0.53705	0.00015162	0.11880	0.78356	0.20406

Overall, the assumption (2) in Section 3 does not hold with the initial model. The poor R^2_{Adj} and MSE indicate poor adequacy of the model.

5. Assessment of the Chosen Prediction Equation

Shapiro-Wilk test result and QQ-plot of R-student residual vs predicted value in Table 5 suggests rejection of H_0 in normality test and we conclude the normality assumption on the error term does not hold.

Table 5 Shapiro-Wilk Test on Normality

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

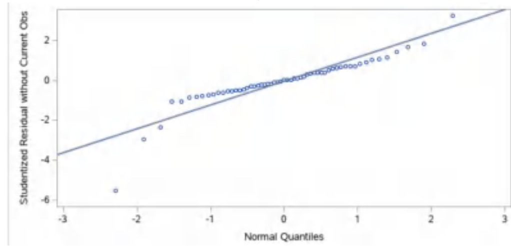
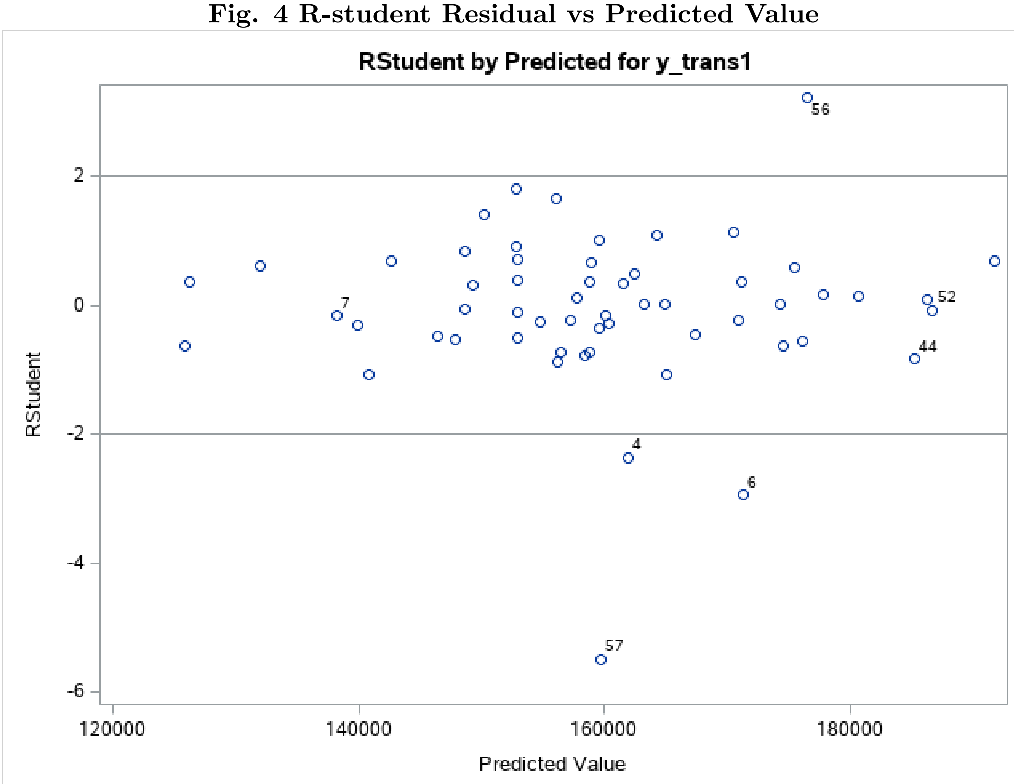


Fig. 4 shows the of R-student residual vs predicted value. The variance on error term has an outward-funnel pattern, likely caused by observation 57, a possible outlier with R-student residual value of less than -4. Therefore, we conclude the constant variance assumption on the error term does not hold, and the assumption (2) in Section 3 does not hold.

Fig. 4 R-student Residual vs Predicted Value



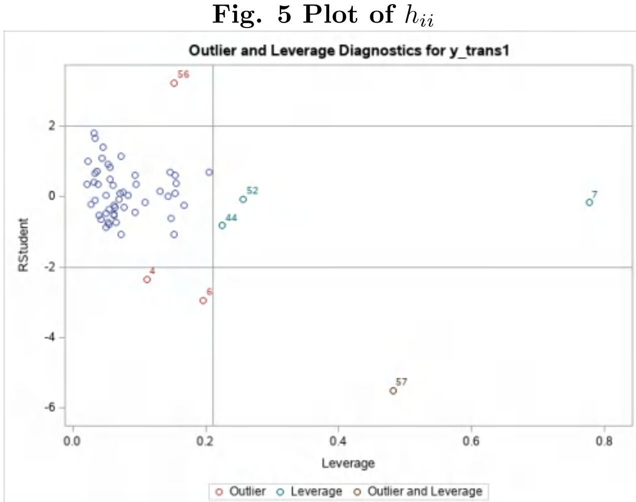
Influential analysis is summarized in Table 6. According to *RStudent* values in Fig. 4, observation 57 is an outlier. Next, H_{ii}, $Cook'sD$, $DFFITS$, and $DFBETAS$ are evaluated to find out possible influential points.

Table 6 Influential Analysis Summary

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	x1	x2	x3	x4	x5
1	117616	125897	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3907	12851	-0.622	0.011	-0.6178	0.1476	1.2827	-0.2574	0.0454	0.1614	0.0171	0.0046	0.0719	0.0477
2	126602	140866	3704	133332	148468	111307	[illegible]	-14236	13403	-1.047	0.019	-1.0623	0.0791	1.0608	-0.2966	0.0791	0.1047	0.1466	-0.0218	0.1276	0.0373
3	120886	136203	8477	115207	137199	96489	156007	-5383	12796	0.374	0.004	0.3708	0.1548	1.3107	0.1568	0.0410	0.1206	0.0098	0.0057	0.0166	0.0256
4	132171	161878	4636	152572	171184	132431	191306	-29706	13137	-2.340	0.106	-2.3622	0.1006	0.8713	-0.8347	0.6990	0.0068	0.3488	0.4264	0.1766	0.2787
5	135916	139824	3644	132267	147582	110890	168919	-4910	13380	-0.300	0.001	-0.2970	0.0760	1.2862	-0.0856	0.0154	0.0369	-0.0346	0.0115	0.0364	0.0042
6	136976	171297	6165	158914	183680	140735	201858	-34316	12492	-2.749	0.307	-2.9486	0.1961	0.5351	-1.4604	0.9386	1.3057	-0.7444	0.1717	0.3662	0.1604
7	137084	136166	13276	110603	162798	100911	175420	9081	8577.8	-0.164	0.016	-0.1628	0.7766	5.0296	-0.3087	-0.0185	-0.0362	0.0247	0.0496	0.2724	0.0700
8	139742	146095	3082	140188	152803	117786	175025	-8853	13574	-0.490	0.002	-0.4884	0.0493	1.1515	-0.1100	-0.0503	0.0602	-0.0399	0.0296	0.0293	0.0407
9	139653	131944	3455	120880	142896	101926	161681	7530	12808	0.818	0.012	0.8145	0.1538	1.2717	0.2617	0.0163	0.1651	0.0498	0.0160	0.1401	0.0169
10	140589	147862	2745	140552	152333	115367	178337	7284	13649	-0.534	0.003	-0.5299	0.0387	1.1326	-0.1064	0.0346	0.0471	0.0186	0.0316	0.0122	0.0260
11	144407	156345	3124	149672	162937	127801	184888	11944	13586	-0.860	0.007	-0.8784	0.0604	1.0816	-0.2023	0.1946	-0.1044	-0.1034	0.0450	0.0382	0.0566
12	146181	152976	3464	146033	159800	124175	181777	-6811	13484	-0.505	0.003	-0.5017	0.0619	1.1640	-0.1286	0.0856	-0.0481	-0.0938	0.0417	0.0080	-0.0304
13	146857	156433	3800	138291	153874	127598	185280	-9752	13439	-0.727	0.006	-0.7234	0.0683	1.1319	-0.1912	0.0274	-0.1072	0.0294	0.1148	0.0074	0.0822
14	147873	148576	3046	142514	166603	119613	177478	-903.2123	13430	-0.067	0.000	-0.0665	0.0693	1.2094	-0.0182	0.0117	-0.0031	-0.0137	0.0020	0.0045	-0.0006
15	147750	158421	3328	151960	164901	129931	187111	-10671	13543	-0.788	0.006	-0.7850	0.0656	1.1089	-0.1971	-0.0302	-0.0790	0.0429	0.0566	0.0901	0.0596
16	149080	158836	3223	162969	165308	130151	187528	-9755	13542	-0.720	0.005	-0.7168	0.0606	1.1191	-0.1706	0.0460	-0.1624	-0.0376	0.0604	0.0353	0.0016
17	151266	142646	6295	130070	156296	111974	173321	8616	12417	0.694	0.021	0.6905	0.2045	1.3373	0.3501	0.0069	-0.0094	0.0449	0.0133	0.2527	-0.0606
18	151301	164972	5541	154060	175896	134965	194900	-13672	12814	-1.067	0.034	-1.0684	0.1527	1.1605	-0.4535	-0.3821	0.0569	0.4005	0.1835	0.0783	0.1182
19	151411	154728	3454	147793	161663	125932	183524	-3317	13486	-0.246	0.001	-0.2437	0.0616	1.1915	-0.0624	0.0356	0.0511	0.0352	-0.0346	-0.0971	0.0275
20	151455	152966	2559	147840	158096	124563	181384	-1514	13688	-0.111	0.000	-0.1096	0.0337	1.1837	-0.0206	-0.0105	0.0053	0.0067	0.0089	0.0016	0.0060
21	152605	149286	3426	142409	156164	120504	178069	4319	13488	0.320	0.001	0.3172	0.0606	1.1644	0.0805	0.0437	0.0024	0.0527	0.0193	-0.0232	-0.0113
22	154231	157286	2330	152608	161584	128931	185600	-3054	13727	-0.221	0.000	-0.2180	0.0278	1.1817	-0.0370	0.0146	-0.0024	-0.0185	0.0048	0.0112	-0.0107
23	154629	159663	3288	153011	166064	130849	188257	-4803	13525	-0.364	0.001	-0.3606	0.0548	1.1730	-0.0868	0.0409	-0.0572	-0.0342	0.0177	0.0717	0.0222
24	156389	160330	3533	153258	167401	131500	189159	-3961	13468	-0.294	0.001	-0.2914	0.0640	1.1910	-0.0762	0.0019	0.0271	-0.0378	0.0176	0.0340	-0.0504
25	156067	160146	4561	146606	168982	130816	188574	-4138	13143	-0.163	0.001	-0.1611	0.1068	1.2697	-0.0563	0.0415	-0.0378	-0.0414	0.0164	0.0087	-0.0126
26	156460	152567	3485	147067	157966	124606	181347	-5663	13586	-0.402	0.001	-0.3584	0.0315	1.1413	-0.0723	0.0456	0.0069	-0.0264	0.0144	0.0230	-0.0162
27	159132	157683	3727	150201	165164	128780	186615	1451	13414	0.108	0.000	0.1071	0.0717	1.2114	0.0296	0.0071	0.0022	0.0140	0.0174	0.0055	-0.0110
28	159605	148604	3262	142015	155192	119880	177319	11225	13528	0.837	0.007	0.8348	0.0556	1.0974	0.2024	0.1119	0.1411	-0.0789	0.0416	-0.0671	-0.0783
29	161446	167366	4242	158873	175906	138133	196607	-5921	13260	-0.450	0.003	-0.4467	0.0929	1.2121	-0.1429	-0.0865	0.0398	0.1007	0.0792	0.0384	-0.0519
30	162780	152920	2675	147549	158290	124480	181361	9861	13662	0.720	0.003	0.7168	0.0389	1.0886	0.1404	0.0087	0.0303	0.0490	0.0620	-0.0190	-0.0524
31	163626	163220	4010	155154	177292	134133	192310	406.6413	13329	0.030	0.000	0.0301	0.0603	1.2284	0.0091	0.0077	-0.0052	-0.0072	0.0015	0.0011	-0.0010
32	163891	158960	2031	154773	162928	130805	187094	5043	13773	0.366	0.000	0.3606	0.0213	1.1326	0.0555	0.0190	0.0059	-0.0184	0.0026	-0.0284	0.0139
33	165217	152786	3208	146348	159225	124105	181467	12431	13547	0.916	0.009	0.9162	0.0531	1.0762	0.2169	0.1351	0.0095	0.1545	0.0499	-0.0446	-0.0197
34	165232	164833	3122	158665	171202	136292	165576	398.8233	13567	0.029	0.000	0.0291	0.0503	1.1857	0.0067	0.0011	-0.0032	-0.0004	-0.0036	-0.0023	0.0041
35	165872	174562	3569	166763	180301	148018	203106	-8690	13519	-0.638	0.003	-0.6343	0.0430	1.1316	-0.1346	0.0314	-0.0256	0.0431	0.0489	0.0002	-0.0727
36	166277	161537	4379	152546	170128	132298	190776	4140	13248	0.356	0.002	0.3547	0.0945	1.2380	-0.1148	0.0675	-0.0749	-0.0741	-0.0717	0.0258	-0.0204
37	167929	170504	5693	159415	182234	140299	201103	2576	12794	0.204	0.002	0.2020	0.1672	1.3436	0.1040	0.0464	0.0457	0.0659	0.0466	0.0111	-0.0424
38	168146	158970	2502	153726	164014	130452	187028	9220	13684	0.561	0.003	0.6775	0.0339	1.1035	0.1269	0.0615	0.0213	0.1698	0.0448	0.0336	0.0524
39	168746	178126	3460	158715	165560	141327	204429	7387	13484	0.548	0.003	0.5441	0.0616	1.1587	0.1397	0.0252	0.0019	0.0297	0.0142	0.0234	-0.0466
40	169636	162566	3212	155711	169098	133459	191537	6620	13522	0.490	0.002	0.4922	0.0586	1.1610	0.1151	0.0060	-0.0072	-0.0146	0.0019	0.0008	0.0616
41	169146	150062	3665	143309	158218	121594	178637	19084	13602	1.403	0.016	1.4156	0.0454	0.9315	0.3048	0.1460	-0.2025	-0.1458	-0.0115	-0.0397	-0.0630
42	171476	159677	2065	155447	163912	131387	187957	13853	13766	1.006	0.004	1.0069	0.0224	1.0219	0.1526	-0.0146	-0.0341	0.0329	0.0478	0.0470	-0.0526
43	174602	174208	5776	163824	184788	146236	204060	393.4320	13886	0.030	0.000	0.0298	0.1413	1.3145	0.0093	0.0040	0.0042	-0.0026	-0.0002	0.0027	0.0081
44	175126	165165	8593	171829	195401	154340	216039	-10039	12301	-0.816	0.032	-0.8183	0.2343	1.3410	-0.3391	0.0584	-0.0295	-0.0509	0.0099	0.0811	-0.0973
45	176100	171155	3867	165760	176550	140990	199619	5012	13660	0.367	0.001	0.3636	0.0372	1.1512	0.0716	-0.0137	-0.0202	0.0087	-0.0144	0.0300	0.0410
46	177165	153846	3469	147660	157803	124483	181232	24321	13701	1.778	0.017	1.8148	0.0314	0.7930	0.3298	0.0756	-0.0537	-0.0391	-0.1759	0.0558	-0.0271
47	178430	156066	2521	150995	161110	127652	184450	22364	13681	1.633	0.015	1.6614	0.0320	0.8434	0.3059	0.1466	-0.0887	-0.1746	-0.0900	0.0170	-0.0896
48	178916	164214	2803	158521	170162	135653	193078	14704	13626	1.080	0.009	1.0603	0.0444	1.0266	0.2312	0.1443	0.1672	0.1377	0.0066	0.0126	-0.0101
49	179853	177722	5039	167605	187838	147596	207848	2131	12916	0.164	0.001	0.1626	0.1310	1.2619	0.0651	0.0069	0.0161	0.0036	0.0472	0.0061	-0.0136
50	180502	180667	3634	172661	189344	151584	209697	-1851	13386	-0.136	0.000	-0.1377	0.0764	1.2153	-0.0392	0.0065	0.0001	-0.0040	0.0044	-0.0089	-0.0010
51	183310	175410	4260	166841	183979	146777	204043	7834	13231	0.596	0.006	0.5953	0.0940	1.1919	0.1916	0.0660	0.0682	0.0034	0.0356	0.0076	-0.0950
52	188848	188874	9048	172591	200830	155849	217698	-3028	12001	-0.266	0.003	-0.0846	0.2582	1.6127	-0.0497	0.0111	0.0019	-0.0116	-0.0029	0.0118	0.0465
53	188838	170461	3759	162902	178017	141554	199429	15338	13406	1.140	0.017	1.1449	0.0727	1.0386	0.3216	0.1176	-0.0866	0.2108	0.0463	-0.0491	-0.0856
54	197390	188183	5468	175920	197187	158158	218219	9198	12804	0.095	0.000	0.0925	0.1512	1.3300	0.0395	0.0074	0.0106	0.0002	0.0269	0.0080	-0.0016
55	200676	191706	5323	181000	202415	161739	221637	6895	12958	0.662	0.034	0.6962	0.1466	1.2473	0.2856	0.0466	0.0527	0.0866	0.0650	0.0346	0.1965
56	214490	176477	5433	165569	187386	146475	208478	38014	12816	2.966	0.283	3.2280	0.1523	0.4267	1.3592	0.2600	0.4870	-0.4351	0.5949	0.4675	-0.8569
57	115737	159645	9061	140541	178057	125623	193575	-43912	10016	-4.383	2.982	-5.4877	0.4822	0.1375	-5.3351	-1.2827	2.5754	0.8974	-4.7227	5.5427	1.2715

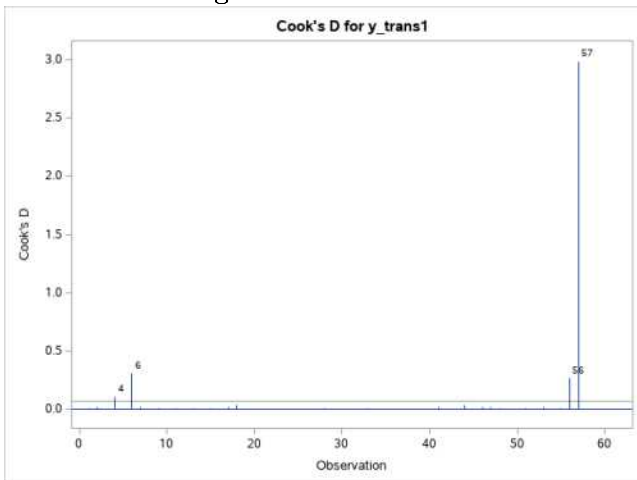
Leverage points are determined by $h_{ii} > 2p/n = (2)(6)/57 = 0.2105$, suggesting point 7, 44, 52 and 57 to be leverage points. Combined with previous finding the point 57 is an outlier, then it is likely to be influential.

Fig. 5 Plot of h_{ii}



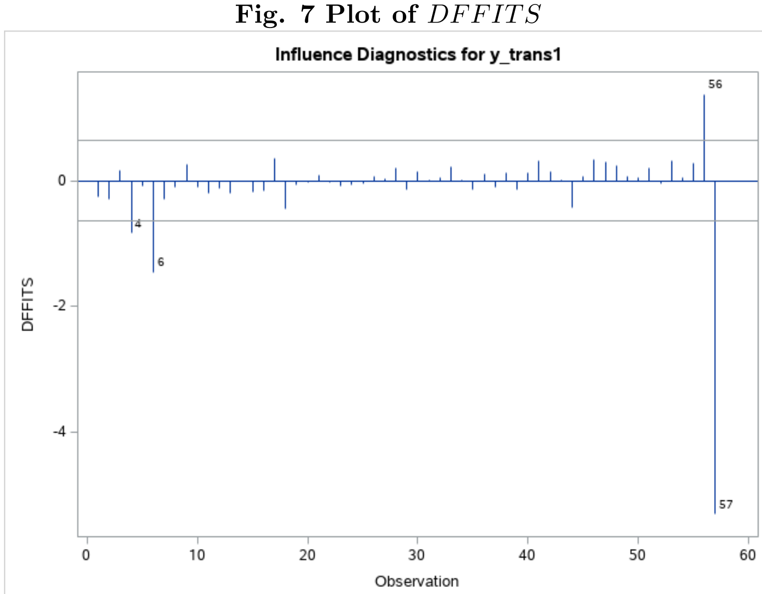
$Cook'sD$ value greatly exceeds $Di > 1$ for observation 57, suggesting it is influential.

Fig. 6 Plot of $CookD$



The cutoff for $DFFITS_i$ is $|DFFITS_i| > 2\sqrt{p/n} = 2\sqrt{5/57} = 0.5923$. Observations 4, 6, 56, and 57 have values of $|DFFITS_i|$ that exceed this value, therefore are most likely influential.

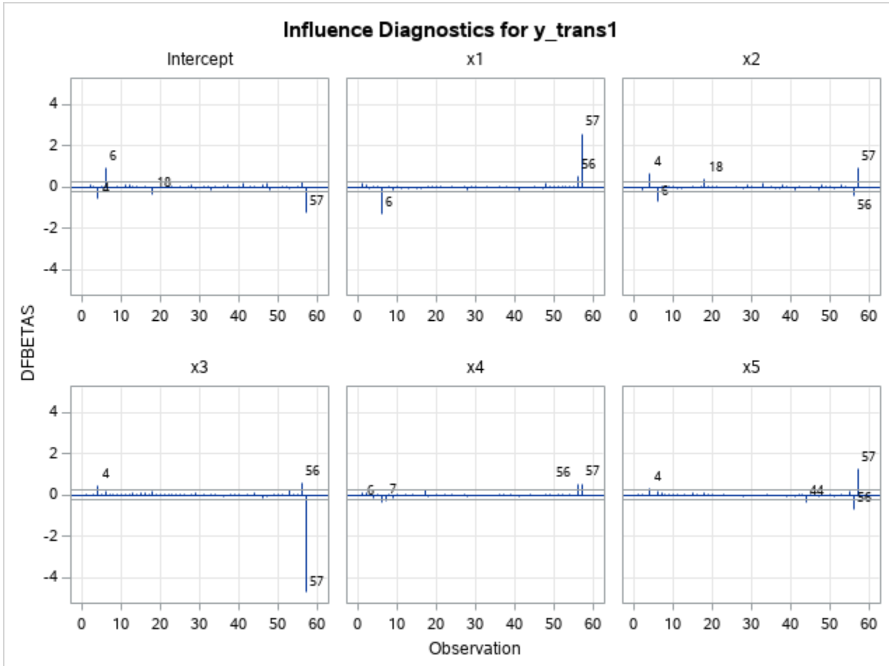
Fig. 7 Plot of $DFFITS$



The cutoff for $DFBETAS_i$ is $|DFBETAS_i| > 2/\sqrt{n} = 2/\sqrt{57} = 0.2649$, we immediately noticed that observation 57 has effect on all five parameters and intercept, and its effect on $\hat{\beta}_1$ and $\hat{\beta}_3$ is large. Point 56 has effect on all five parameters other than intercept. Point 4 has effect on intercept, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_5$, especially for $\hat{\beta}_2$. Point 6 has effect on intercept, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_4$, especially for intercept and $\hat{\beta}_1$. Point 18 has small effect on intercept and $\hat{\beta}_2$. Point 7 has small effect on $\hat{\beta}_4$. Point 44 has small effect on $\hat{\beta}_5$.

The cutoff for $COVRATIO_i$ is $1 \pm 3p/n = 1 \pm (3)(6)/57$, or 1.316 and 0.684. Note that the values of observation 6, 8, 15, 52, 56, and 57 clearly exceed these limits, indicating these points are influential. However, point 4 barely exceed the cutoff, so the influence of it, from a practically point of view, is fairly small.

Fig. 8 Plot of $DFBETA$



Adopting a diagnostic view, point 57 is clearly influential since it has effect on H_{ii}, $RStudent$, $Cook'sD$, $\hat{\beta}$, $\hat{Y}_i$ and $COVRATIO_i$, point 56 have effect on $\hat{\beta}$, $\hat{Y}_i$ and $COVRATIO_i$, and point 4, 6 have moderate effect on $\hat{\beta}$ and $\hat{Y}_i$. Hence, we conclude that point 57 are influential.

After removing the influential point 57, the model specified in Section 3 is refitted and analyzed as follows.

The ANOVA analysis result given in Table 7 shows the p-value for F-test is smaller than 0.0001, indicating rejection of the H_0 that all $\beta_{ij}s = 0$. The assumption (1) in Section 3 holds. A slightly decreased CV of 6.91% shows the ratio of standard deviation to mean is low, suggesting a low variability in the data. However, a slightly higher R^2_{Adj} of 64.94% means only less than 70% of the variability in the data can be explained by the model. Also, the MSE of 11100^2 remains extremely large.

Table 7 ANOVA Analysis after Deleting the Outlier

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	13165586677	2633117335	21.37	<.0001
Error	50	6160393749	123207875		
Corrected Total	55	19325980426			

Root MSE	11100	R-Square	0.6812
Dependent Mean	160541	Adj R-Sq	0.6494
Coeff Var	6.91405		

Read from the parameter estimates Table 8, the p-values obtained for x_1, x_2 and x_4 in t-test are greater than 0.05, indicating insignificant of these three variables at 0.05 level, which will cause the model to be less informative due to great loss of predictors. The absolute value of $\hat{\beta}_2$ is extremely large while that of $\hat{\beta}_4$ and $\hat{\beta}_5$ are too small in magnitude compared with other predictor coefficients, suggesting the model does not fit well.

All VIFs are smaller than 3, meaning only moderate correlations are observed among the predictors. Therefore, the assumption (3) in Section 3 holds. Multicollinearity is not a concern for the data.

Table 8 Parameter Estimates after Deleting the Outlier

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	173851	27808	6.25	<.0001	0	0	117998	229705
x1	x1	1	378.71431	209.29784	1.81	0.0764	0.20507	2.01480	-41.67278	799.10140
x2	x2	1	-4013.61098	2134.85271	-1.88	0.0659	-0.18495	1.51795	-8301.58885	274.36688
x3	x3	1	1036.70608	193.56682	5.36	<.0001	0.49108	1.31872	647.91569	1425.49647
x4	x4	1	-38.62336	40.71193	-0.95	0.3473	-0.09833	1.68523	-120.39568	43.14896
x5	x5	1	99.10320	27.52533	3.60	0.0007	0.33981	1.39720	43.81695	154.38945

Shapiro-Wilk test result and QQ-plot of R-student residual vs predicted value in Table 9 indicates we fail to reject H_0 in normality test and we conclude the normality assumption on the error term hold.

Table 9 Shapiro-Wilk Test on Normality

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

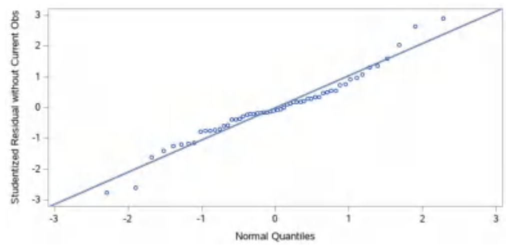
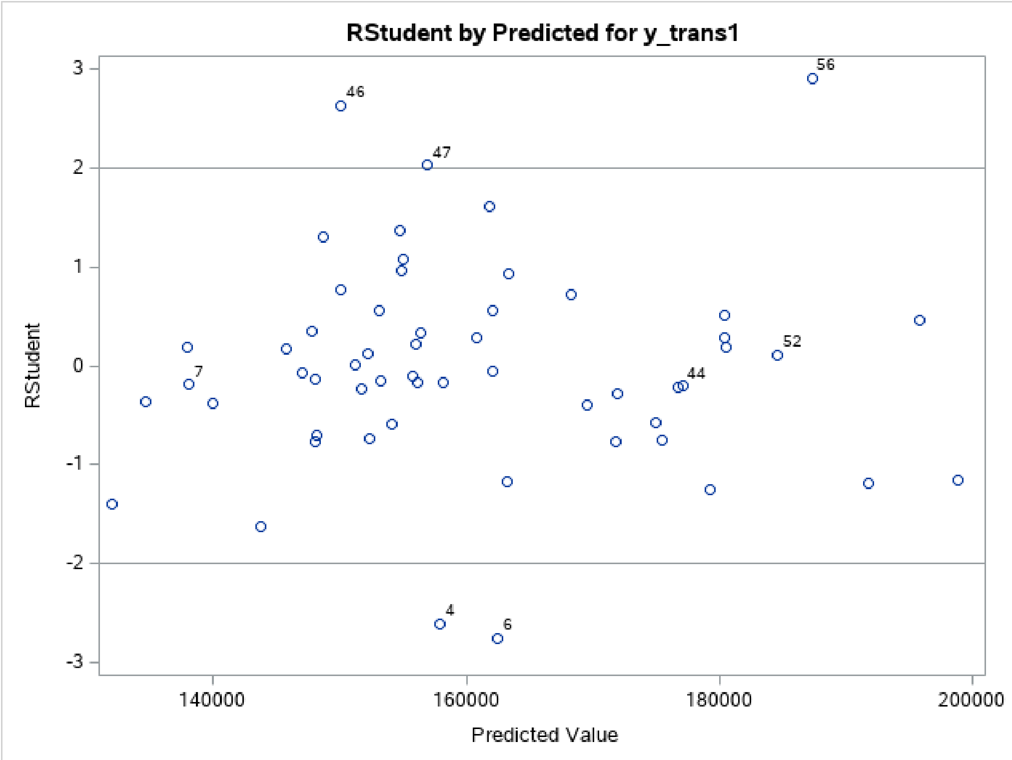


Fig. 9 shows the of R-student residual vs predicted value. The variance on error term has an decent even pattern along size the zero line, suggesting the constant variance assumption on the error term holds. Therefore, the assumption (2) in Section 3 holds.

Fig. 9 R-student Residual vs Predicted Value



Overall, the assumptions (1), (2) and (3) in Section 3 all hold with the refitted model after deleting the outlier. And there is no sign of strong multicollinearity among the predictors. However, the decent R^2_{Adj} of 64.94% and the poor MSE of 11100^2 indicate lack of adequacy of the refitted model.
Box-Cox analysis is performed again, and the result in Fig. 10 suggests $y^{1.5}$ is a proper transformation of the response after deleting the outlier. However, the R^2_{Adj} and MSE do not improve much as shown in Table 10.

Fig. 10 Box-Cox Analysis after Deleting the Outlier

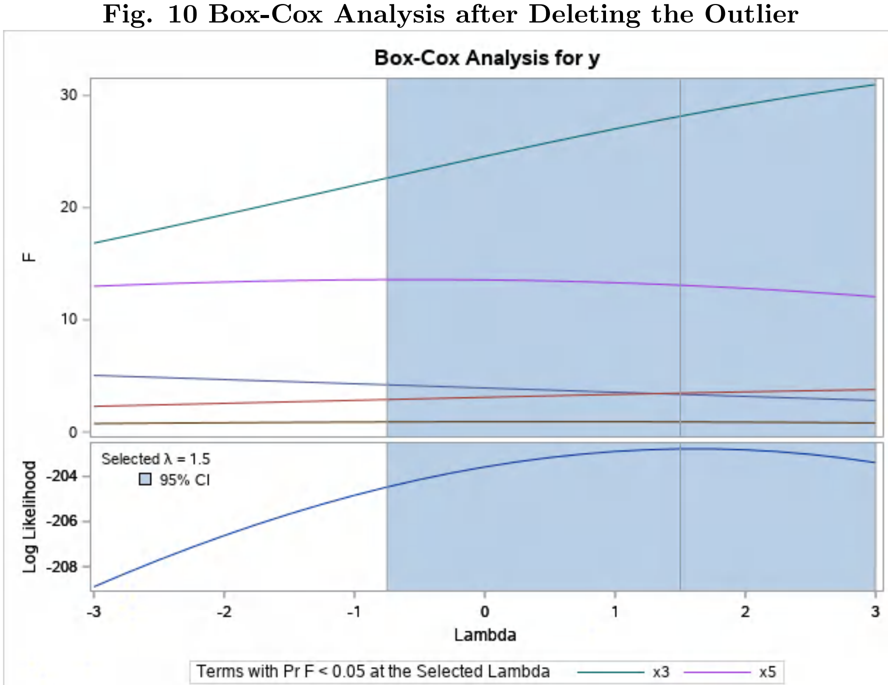


Table 10 ANOVA Analysis after Second Transformation on Response

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		

Hence, we conclude MLR with an OLS model does not provide a desirable fit for the data, especially for the complete data. To make the model as informative as possible, we want to keep all the observations and consider WLS as a method to accommodate the outlier via downweighting it to a less impactful point, or almost to zero.

To our delight, read from Tabel 11 and Table 12, the ANOVA analysis shows the p-value for F-test is smaller than 0.0001, indicating rejection of the H_0 that all $\beta_{ij}s = 0$. A smaller CV of 0.17%, a higher R^2_{Adj} of 71.62%, and a profoundly decreased MSE of 1.5448^2 are obtained. All VIFs are smaller than 3, meaning only moderate correlations are observed among the predictors. Therefore, multicollinearity is not a concern for the data. We conclude the WLS model provides a better fit to the data than the OLS model.

Table 11 ANOVA Analysis of the WLS Model

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		

Table 12 Parameter Estimates of the WLS Model

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	0.06866	-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

The weights obtained for the WLS model are listed in the below Table 13.

Table 13 Weights of the WLS Model

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	15.2858	805.3799	866.6744	732.8233	939.2307	-45.2970	46.664	-0.971	0.017
2	9.81E-04	824	884.5425	10.6461	863.1700	905.9157	783.7501	985.3367	-60.7025	47.096	-1.289	0.015
3	9.04E-04	840	838.9858	14.9086	808.7851	868.8084	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-02	856	877.2580	10.1032	866.9619	887.5185	812.4088	942.0016	-19.8152	28.588	-0.677	0.089
6	2.10E-03	861	971.0087	11.3160	948.0188	994.2648	900.1505	1043	-110.0987	31.763	-3.467	0.254
7	1.57E-02	862	863.9891	11.8029	840.2741	887.6641	830.2280	897.7161	-2.1381	1.864	-1.009	7.104
8	3.79E-04	871	892.3087	8.1221	876.2359	908.8456	791.5117	993.1077	-21.1987	48.790	-0.435	0.001
9	1.22E-03	872	857.7874	15.2245	827.2022	888.3626	763.9037	951.5711	13.9026	41.513	0.337	0.003
10	1.28E-03	924	897.6523	7.2506	883.0890	912.2245	809.8050	985.4995	23.3721	42.537	0.549	0.001
11	1.92E-02	887	924.8840	5.3630	914.1174	935.6517	891.4655	958.3286	-37.4140	35.776	-1.046	0.004
12	1.14E-02	884	914.1861	3.5788	907.0013	921.3709	884.2469	944.1252	20.1961	14.026	1.440	0.022
13	5.05E-03	890	916.2322	5.7659	904.0488	927.5159	871.0923	961.3721	20.5322	26.662	0.980	0.012
14	2.05E-03	899	904.8722	6.5658	891.0426	918.1017	845.3062	964.4381	5.6122	20.189	0.199	0.000
15	1.44E-03	900	924.6196	7.3946	909.7746	939.4640	841.4560	1006	36.0896	40.095	0.628	0.012
16	4.91E-03	904	926.0807	4.9500	916.4516	936.5250	876.4175	976.3099	22.0237	21.080	0.901	0.080
17	1.44E-03	912	891.0456	9.9303	871.0957	910.9974	806.9578	975.1236	20.0544	39.467	0.520	0.005
18	1.16E-03	912	937.9616	12.2788	913.3046	962.6884	844.4567	1031	36.1316	42.212	0.605	0.005
19	1.56E-03	912	910.2902	6.2059	893.7964	926.7620	830.0146	990.5450	1.9140	38.296	0.050	0.000
20	1.06E-03	912	910.2204	6.8096	896.5496	925.8910	815.9837	1000	0.1296	46.659	0.000	0.000
21	7.56E-04	920	913.6625	10.6306	892.6114	934.7505	790.8009	1020	6.0775	56.185	0.110	0.000
22	2.33E-03	922	931.1340	4.9179	921.2609	941.0070	806.0782	996.1090	-9.2040	31.050	-0.290	0.000
23	1.46E-03	920	935.0876	6.6815	921.0747	948.3008	853.7668	1017	11.8576	39.935	0.297	0.000
24	2.18E-02	909	937.9126	6.3302	925.2002	950.6250	913.3490	962.4768	-6.7080	6.237	-1.061	0.106
25	2.67E+01	905	924.6813	0.2996	904.0810	935.2217	938.8325	935.5305	0.0107	0.00782	2.447	1504.921
26	7.29E-04	906	921.8420	8.9541	906.0867	939.6190	805.6160	1035	14.3872	58.492	0.255	0.000
27	1.94E-03	908	917.5790	8.9289	898.6573	935.5004	843.1805	991.3973	20.9242	34.083	0.600	0.014
28	4.36E-04	941	907.6752	10.7012	886.3917	929.3680	750.1163	1065	33.3048	73.046	0.456	0.001
29	6.16E-03	946	946.5620	10.1118	926.2624	966.8032	902.1697	990.9059	0.3020	18.665	0.020	0.000
30	3.10E-02	951	911.8074	4.6304	902.5416	921.1302	856.3600	968.3400	38.6030	27.365	1.420	0.040
31	4.09E-04	964	951.3972	10.3911	930.5983	972.2682	801.6666	1101	2.1628	73.009	0.030	0.000
32	6.56E-04	964	935.2000	7.2810	920.0126	949.6471	826.2181	1042	19.2100	52.300	0.367	0.000
33	3.59E-04	963	923.8601	9.7017	904.3813	943.3360	816.2779	1031	24.5619	51.801	0.675	0.003
34	1.43E-03	969	960.6669	7.4510	946.7265	975.6453	877.2449	1044	-1.4669	40.305	-0.036	0.000
35	5.62E-04	961	989.7259	12.3012	965.0303	1014	856.9687	1123	-28.7159	64.005	-0.449	0.001
36	3.32E-03	962	926.0902	10.6504	904.8923	947.2870	866.7044	986.3956	36.2598	26.065	1.391	0.053
37	9.71E-04	960	994.1097	16.7379	960.5068	1020	889.0537	1099	36.3097	48.670	0.564	0.007
38	7.29E-04	969	942.8647	9.8660	923.0579	962.6715	826.2997	1058	25.7953	56.361	0.458	0.001
39	4.45E-03	970	987.8021	6.6281	974.7004	1001	939.4026	1036	-17.3021	22.226	-0.780	0.009
40	2.66E-04	971	955.6047	15.7740	924.1371	987.4724	762.6052	1149	15.3153	93.624	0.164	0.000
41	6.59E-04	972	908.3087	6.3903	891.6024	925.1509	708.8405	1028	64.1332	58.273	1.101	0.004
42	9.12E-04	986	943.4560	6.0647	927.2291	959.6905	838.4820	1047	42.4502	50.510	0.841	0.003
43	5.46E-03	989	983.2169	7.8331	967.4913	998.9425	938.4738	1028	6.0531	19.336	0.313	0.003
44	2.34E-03	991	1013	10.8763	991.0766	1035	945.2462	1081	-21.6216	29.891	-0.721	0.011
45	4.29E-03	995	971.2679	5.2158	961.4173	982.3684	923.3699	1020	22.7521	22.014	0.909	0.000
46	2.75E-03	998	910.0849	5.6949	898.6619	921.5079	849.6749	970.5140	87.7951	28.959	3.027	0.059
47	1.06E-03	1002	922.2795	6.7257	908.7767	935.7513	825.6152	1019	79.6215	47.201	1.687	0.010
48	1.57E-03	1004	953.3512	6.9483	909.4019	967.3005	873.8382	1033	50.1400	38.466	1.307	0.009
49	1.76E-04	1006	1016	25.3307	964.9616	1067	778.1094	1254	-9.4456	112.9	-0.084	0.000
50	3.76E-03	1015	1007	8.1049	991.0462	1024	954.4412	1061	7.5419	23.747	0.318	0.002
51	3.36E-04	1018	989.1126	16.6715	956.6431	1023	816.4466	1163	28.4974	82.712	0.345	0.001
52	1.16E-02	1025	1029	10.4390	1008	1050	995.3236	1064	-3.8410	9.650	-0.398	0.031
53	3.08E-04	1028	989.3080	18.6013	952.4946	1027	809.2747	1170	36.6920	88.003	0.415	0.001
54	2.11E-04	1030	1042	25.9990	989.0203	1094	821.9026	1263	-11.6057	103.2	-0.113	0.000
55	1.84E-02	1071	1045	12.6696	1019	1070	964.2821	1125	26.4024	35.931	0.735	0.011
56	1.76E-04	1113	1002	23.7037	953.9546	1049	764.6821	1239	111.5181	113.2	0.985	0.007
57	5.70E-05	782	986.4081	41.1731	905.7776	1071	569.3054	1408	-206.9081	200.5	-1.028	0.007

6. Selection of Variable Subset

Stepwise selection and all possible regressions are used for variable selection of the WLS model at 0.05 level, the outputs are shown in Table 14 and Table 15, respectively. All of the three statistics, R_p^2 , C_p (close to $p+1=6$), and MS_{Res} optimize when x_1, x_3, x_4 and x_5 are included in the model, indicating we should remove x_2 in the final model.

Table 14 Backward 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

Table 15 All Possible Regressions

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

7. Conclusion

After using the OLS and WLS models, WLS is chosen as the best estimation method. And the final prediction equation is

$$\hat{y} = 803.64449 + 2.33451x_1 + 2.28908x_3 - 0.12425x_4 + 0.42937x_5$$

We conclude $Mortality$ is positively associated with mean annual precipitation (x_1), percentage of the population that is nonwhite (x_3), relative pollution potential of sulfur dioxide (x_5), and is negatively related to relative pollution potential of oxides of nitrogen (x_4). However, $Mortality$ is not associated with Median number of school years completed for persons of age 25 years or older (x_2). An advantage of the WLS model is relatively even magnitudes of the predictor coefficients. A concern is the negative association between y and x_4, but a reasonable explanation maybe the existed large composition of nitrogen oxides in the air we breathe. Further investigation should be focused on whether the outlier data is reliable, because it is influential to modeling.

8. Appendix

```
/* 1. import database */
filename reffile '/home/u49400069/Regression Analysis/Final_Project/Pollution_Mortality.xls';

proc import datafile=reffile
        dbms=xls
        out=work.pollution;
        getnames=yes;
run;
proc print data=work.pollution;run;
proc contents data=work.pollution; run;



/* 2. investigation of the data */
/* Boxplots */
data long;
   set pollution;
   array tm(*) x1 - x5;

   do i=1 to dim(tm);
      hour=compress(vname(tm(i)), 'kd');
      Value=tm(i);
      output;
   end;
   keep hour value;
run;

proc sgplot data=long;
```

```
   vbox value / group=hour;
run;

proc sgplot data=have;
   vbox x1 x2 x3 x4 x5;
run;

* correlation table & scatter plots ;
proc corr data=pollution plots=matrix(histogram nvar=all);
run;

* loess plots ;
proc sgplot data=pollution;
	reg x=x1 y=y / clm cli;
run;

proc loess data=pollution;
	model y=x1/smooth=0.1 0.2 0.4 0.6 0.8 1.0;
run;

proc sgplot data=pollution;
	reg x=x2 y=y / clm cli;
run;

proc loess data=pollution;
	model y=x2/smooth=0.1 0.2 0.4 0.6 0.8 1.0;
run;

proc sgplot data=pollution;
```

```
        reg x=x3 y=y / clm cli;
run;

proc loess data=pollution;
        model y=x3/smooth=0.1 0.2 0.4 0.6 0.8 1.0;
run;

proc sgplot data=pollution;
        reg x=x4 y=y / clm cli;
run;

proc loess data=pollution;
        model y=x4/smooth=0.1 0.2 0.4 0.6 0.8 1.0;
run;

proc sgplot data=pollution;
        reg x=x5 y=y / clm cli;
run;

proc loess data=pollution;
        model y=x5/smooth=0.1 0.2 0.4 0.6 0.8 1.0;
run;


/* 3. specification of the model */
* box-cox analysis and transformation on y ;
proc transreg data=pollution;
        model boxcox(y)=identity(x1 x2 x3 x4 x5);
run;
```

```
data trans1;
	set pollution;
	y_trans1=y**2;
run;


/* 4. estimation of the appropriate model */
proc reg data=trans1 plots(label)=(cooksd RSTUDENTBYPREDICTED dfbetas dffits diagnostics
		observedbypredicted);
	model y_trans1=x1 x2 x3 x4 x5/
	alpha=.05 r p clb cli clm stb vif partial influence collinoint collin;
	output out=one r=resid student=sresid p=pred rstudent=rs r=y_res;
	run;
proc univariate data=one plot normal;
	var rs; * qqplot r-student residual vs predicted ;
run;


/* 5. assessment of the chosen prediction equation */
* delete an outlier ;
data pollution_new;
	set trans1 end=last;
	if not last then
		output;
run;

* refit the model ;
proc reg data=pollution_new plots(label)=(cooksd RSTUDENTBYPREDICTED dfbetas dffits diagnostics
```

```
        observedbypredicted);
    model y_trans1=x1 x2 x3 x4 x5/alpha=.05 r p clb cli clm stb vif partial
        influence collinoint collin;
    output out=one r=resid student=sresid p=pred rstudent=rs r=y_res;
    run;

proc univariate normal plot data=one;
    var rs; * qqplot r-student vs predicted ;
run;

* box-cox analysis and transformation on y ;
proc transreg data=pollution_new;
    model boxcox(y)=identity(x1 x2 x3 x4 x5);
run;

data trans2;
    set pollution_new;
    y_trans2=y**1.5;
run;

* refit the model ;
proc reg data=trans2 plots(label)=(cooksd dfbetas dffits RSTUDENTBYPREDICTED diagnostics
observedbypredicted);
    model y_trans2=x1 x2 x3 x4 x5/alpha=.05 r p clb cli clm stb vif partial
        influence collinoint collin;
    output out=one r=resid student=sresid p=pred rstudent=rs r=y_res;
    run;

proc univariate normal plot data=one;
```

```
            var rs; * qqplot r-student vs predicted ;
run;

/* weighted ls est */
* step0: initial step ;
proc reg data=pollution;
            model y=x1 x2 x3 x4 x5 / clb;
            output out=result1 p=yhat r=resid;
            run;

* step1: estimate standard dev. function ;
data result1;
            set result1;
            absres=abs(resid);
run;

proc reg data=result1;
            model absres=x1 x2 x3 x4 x5;
            output out=step1 p=preds1 r=ress;
            run;

data step1;
            set step1;
            wt1=1/(preds1)**2;
run;

proc reg data=step1;
            model y=x1 x2 x3 x4 x5 /p clb;
            weight wt1;
```

```
            output out=result2 p=wyhat r=wres;
            run;

* step2: estimate the standard dev. function ;
data result2;
            set result2;
            abswres=abs(wres);
run;

proc reg data=result2;
            model abswres=x1 x2 x3 x4 x5;
            output out=step2 p=preds2;
            run;

data step2;
            set step2;
            wt2=1/(preds2)**2;
run;

proc reg data=step2 plots(label)=(RSTUDENTBYPREDICTED);
            model y=x1 x2 x3 x4 x5 /p r clb cli clm stb vif partial collinoint collin;
            weight wt2;
            output out=one r=resid student=sresid p=pred rstudent=rs r=y_res;
            run;



/* 6. variable selection */
* forward selection ;
proc reg data=step2;
```

```
    model y=x1 x2 x3 x4 x5 / selection=forward slentry=0.25;
    weight wt2;
    run;

* backward selection ;
proc reg data=step2;
    model y=x1 x2 x3 x4 x5 / selection=backward slstay=0.1;
    weight wt2;
    run;

* stepwise selection ;
proc reg data=step2;
    model y=x1 x2 x3 x4 x5 / selection=stepwise slentry=0.15 slstay=0.15;
    weight wt2;
    run;

* all possible selection ;
proc reg data=step2;
    model y=x1 x2 x3 x4 x5 / selection=cp rsquare mse adjrsq p clm cli best=10;
    weight wt2;
    run;
```