

Project Report

The data set that I will analyze corresponds to information of air pollution and related values for 41 U.S. cities, which were collected from U.S. government publications. The data are means over the years 1969-1971. The idea is to find a correlation between a different set of variables affecting air quality, in order to find those variables that are producing the mayor effect.

Data file Name: Air Pollution

Data file Subjects: Environment

Story Names:

Reference: Sokal, R.R. and Rohlf, F.J. (1981) *Biometry*, 2nd edition, San Francisco: W.H. Freeman, 239. Also found in: Hand, D.J., *et al.* (1994) *A Handbook of Small Data Sets*, London: Chapman & Hall, 20-21.

Authorization: Contact Author

Description: These data give air pollution and related values for 41 U.S. cities and were collected from U.S. government publications. The data are means over the years 1969-1971.

Number of cases: 41

Variable Names:

1. City: City
2. SO2: Sulfur dioxide content of air in micrograms per cubic meter
3. Temp: Average annual temperature in degrees Fahrenheit
4. Man: Number of manufacturing enterprises employing 20 or more workers
5. Pop: Population size in thousands from the 1970 census
6. Wind: Average annual wind speed in miles per hour
7. Rain: Average annual precipitation in inches
8. RainDays: Average number of days with precipitation per year

The Data:

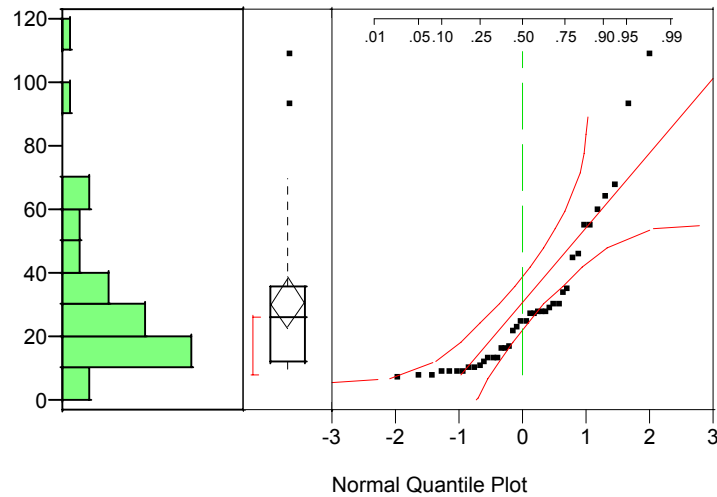
No	City	SO ₂	Temp	Man	Pop	Wind	Rain	RainDays
1	Phoenix	10	70.3	213	582	6	7.05	36
2	Little Rock	13	61	91	132	8.2	48.52	100
3	San Francisco	12	56.7	453	716	8.7	20.66	67
4	Denver	17	51.9	454	515	9	12.95	86
5	Hartford	56	49.1	412	158	9	43.37	127
6	Wilmington	36	54	80	80	9	40.25	114
7	Washington	29	57.3	434	757	9.3	38.89	111
8	Jacksonville	14	68.4	136	529	8.8	54.47	116
9	Miami	10	75.5	207	335	9	59.8	128
10	Atlanta	24	61.5	368	497	9.1	48.34	115
11	Chicago	110	50.6	3344	3369	10.4	34.44	122
12	Indianapolis	28	52.3	361	746	9.7	38.74	121
13	Des Moines	17	49	104	201	11.2	30.85	103
14	Wichita	8	56.6	125	277	12.7	30.58	82
15	Louisville	30	55.6	291	593	8.3	43.11	123
16	New Orleans	9	68.3	204	361	8.4	56.77	113
17	Baltimore	47	55	625	905	9.6	41.31	111
18	Detroit	35	49.9	1064	1513	10.1	30.96	129
19	Minn-St. Paul	29	43.5	699	744	10.6	25.94	137
20	Kansas City	14	54.5	381	507	10	37	99
21	St. Louis	56	55.9	775	622	9.5	35.89	105
22	Omaha	14	51.5	181	347	10.9	30.18	98
23	Albuquerque	11	56.8	46	244	8.9	7.77	58
24	Albany	46	47.6	44	116	8.8	33.36	135
25	Buffalo	11	47.1	391	463	12.4	36.11	166
26	Cincinnati	23	54	462	453	7.1	39.04	132
27	Cleveland	65	49.7	1007	751	10.9	34.99	155
28	Columbus	26	51.5	266	540	8.6	37.01	134
29	Philadelphia	69	54.6	1692	1950	9.6	39.93	115
30	Pittsburgh	61	50.4	347	520	9.4	36.22	147
31	Providence	94	50	343	179	10.6	42.75	125
32	Memphis	10	61.6	337	624	9.2	49.1	105
33	Nashville	18	59.4	275	448	7.9	46	119
34	Dallas	9	66.2	641	844	10.9	35.94	78
35	Houston	10	68.9	721	1233	10.8	48.19	103
36	Salt Lake City	28	51	137	176	8.7	15.17	89
37	Norfolk	31	59.3	96	308	10.6	44.68	116
38	Richmond	26	57.8	197	299	7.6	42.59	115
39	Seattle	29	51.1	379	531	9.4	38.79	164
40	Charleston	31	55.2	35	71	6.5	40.75	148

Exploring Data

As a starting point in the analysis, the existing data, forty observations, are explored for outliers, and normality. Every single variable will be analyzed separately.

1. Dependent Variable: SO₂

As it can be seen in the outlier box plot there are two possible outliers, observation 11 and observation 31, with 110 and 94 $\mu\text{g} / \text{m}^3$ respectively. The minimum value= $8 \mu\text{g} / \text{m}^3$, maximum= $110 \mu\text{g} / \text{m}^3$, $Q_1=12.25 \mu\text{g} / \text{m}^3$, $Q_2=26 \mu\text{g} / \text{m}^3$, $Q_3=35.75 \mu\text{g} / \text{m}^3$. These values suggest a not symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



Quantiles

100.0%	maximum	110.00
99.5%		110.00
97.5%		109.60
90.0%		64.60
75.0%	quartile	35.75
50.0%	median	26.00
25.0%	quartile	12.25
10.0%		10.00
2.5%		8.03
0.5%		8.00
0.0%	minimum	8.00

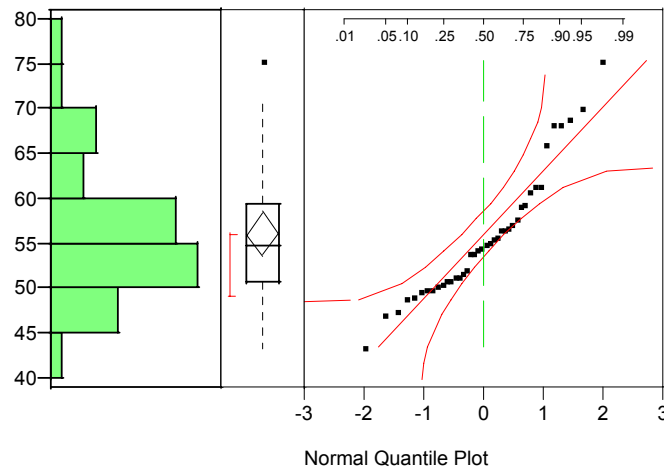
Moments

Mean	30.4
Std Dev	23.661935
Std Err Mean	3.7412805
upper 95% Mean	37.967454
lower 95% Mean	22.832546
N	40

It can also be seen that $media < mean$, hence there is a positive skew distribution. A good way to normalize the distribution would be to apply a natural logarithm.

2. Predictor Variable: Temp

As it can be seen in the outlier box plot there is one possible outlier, observation 79, with 75.5 degrees Fahrenheit. The minimum value=43.5 degrees Fahrenheit, maximum=75.5 degrees Fahrenheit, $Q_1=50.7$ degrees Fahrenheit, $Q_2=54.8$ degrees Fahrenheit, $Q_3=59.375$ degrees Fahrenheit. These values suggest a not symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .

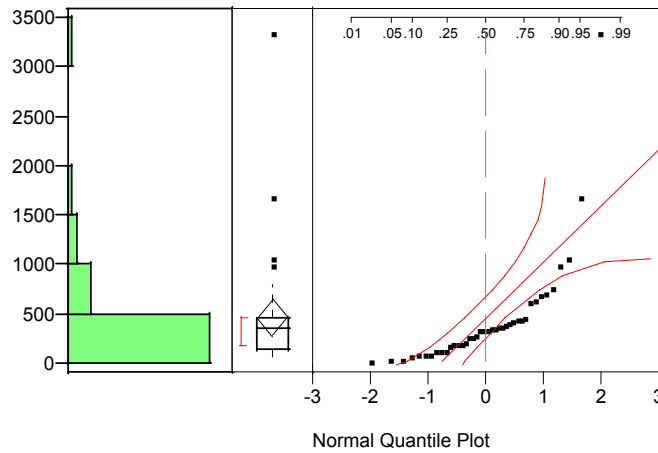


Quantiles			Moments	
100.0%	maximum	75.500	Mean	56.015
99.5%		75.500	Std Dev	7.1356652
97.5%		75.370	Std Err Mean	1.1282477
90.0%		68.390	upper 95% Mean	58.297096
75.0%	quartile	59.375	lower 95% Mean	53.732904
50.0%	median	54.800	N	40
25.0%	quartile	50.700		
10.0%		49.010		
2.5%		43.590		
0.5%		43.500		
0.0%	minimum	43.500		

It can also be seen that $media < mean$, hence there is a positive skew distribution. A good way to normalize the distribution would be to apply a natural logarithm.

3. Predictor Variable: Man

As it can be seen in the outlier box plot there are four possible outliers, observations 11, 18, 27 and 29 with 3344, 1064 manufacturing enterprises respectively, 1007, 1692 manufacturing enterprises. The minimum value=35 manufacturing enterprises, maximum=3344 manufacturing enterprises, $Q_1=148$ manufacturing enterprises, $Q_2=345$ manufacturing enterprises, $Q_3=460$ manufacturing enterprises. These values suggest a not symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



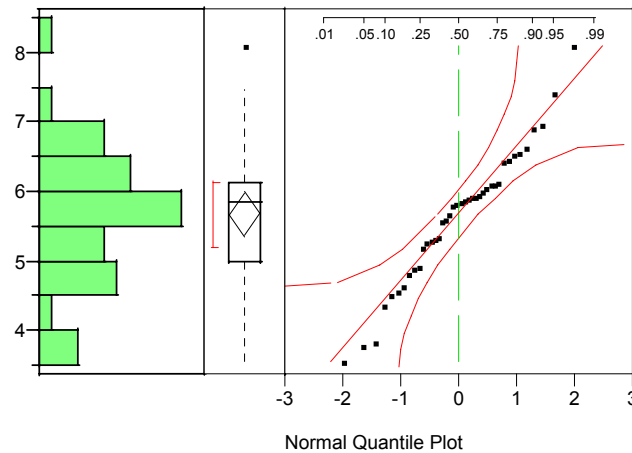
Quantiles

100.0%	maximum	3344.0
99.5%		3344.0
97.5%		3302.7
90.0%		983.8
75.0%	quartile	460.0
50.0%	median	345.0
25.0%	quartile	148.0
10.0%		81.1
2.5%		35.2
0.5%		35.0
0.0%	minimum	35.0

Moments

Mean	460.45
Std Dev	570.39392
Std Err Mean	90.187198
upper 95% Mean	642.87083
lower 95% Mean	278.02917
N	40

It can also be seen that $media < mean$, hence there is a positive skew distribution. A good way to normalize the distribution would be to apply a natural logarithm. If this transformation is made the normal plot and the histogram would look like the plot that follows.



Quantiles

100.0%	maximum	8.1149
99.5%		8.1149
97.5%		8.0979
90.0%		6.8885
75.0%	quartile	6.1312
50.0%	median	5.8435
25.0%	quartile	4.9896
10.0%		4.3949
2.5%		3.5611
0.5%		3.5553
0.0%	minimum	3.5553

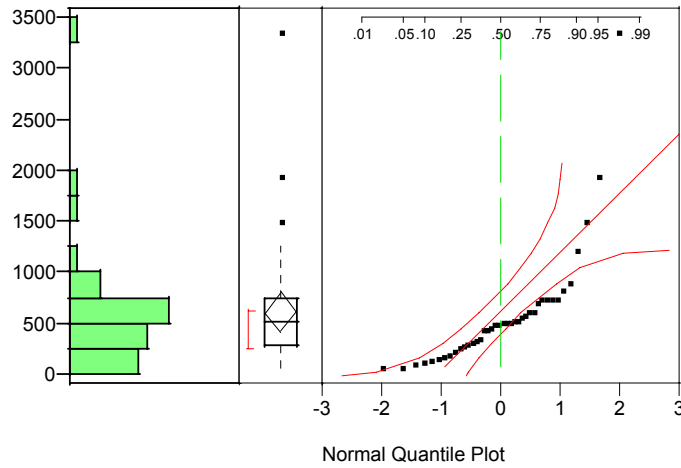
Moments

Mean	5.6757996
Std Dev	0.9699698
Std Err Mean	0.1533657
upper 95% Mean	5.986011
lower 95% Mean	5.3655882
N	40

The normal plot shows a distribution looking much more normal. Mean and media are very similar and the differences between quartiles are also smaller. Though there still is one outlier.

4. Predictor Variable: Pop

As it can be seen in the outlier box plot there are three possible outliers, observations 11,18 and 29 with 3369,1513 and 1950 thousands of people. The minimum value=71 thousands of people, maximum=3369 thousands of people, $Q_1=282.5$ thousands of people, $Q_2=511$ thousands of people, $Q_3=737$ thousands of people. These values suggest a symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



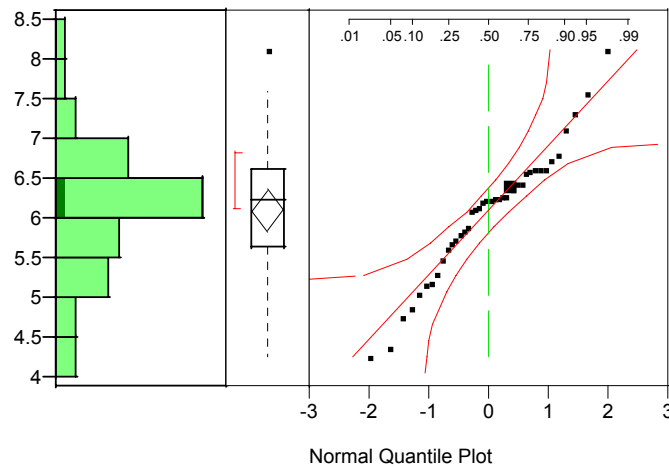
Quantiles

100.0%	maximum	3369.0
99.5%		3369.0
97.5%		3333.5
90.0%		1200.2
75.0%	quartile	737.0
50.0%	median	511.0
25.0%	quartile	282.5
10.0%		134.6
2.5%		71.2
0.5%		71.0
0.0%	minimum	71.0

Moments

Mean	605.9
Std Dev	586.22726
Std Err Mean	92.690668
upper 95% Mean	793.38457
lower 95% Mean	418.41543
N	40

It can also be seen that $media < mean$, hence there is a positive skew distribution. A good way to normalize the distribution would be to apply a natural logarithm. If this transformation is made the normal plot and the histogram would look like the plot that follows.



Quantiles

100.0%	maximum	8.1224
99.5%		8.1224
97.5%		8.1087
90.0%		7.0863
75.0%	quartile	6.6025
50.0%	median	6.2363
25.0%	quartile	5.6431
10.0%		4.9008
2.5%		4.2657
0.5%		4.2627
0.0%	minimum	4.2627

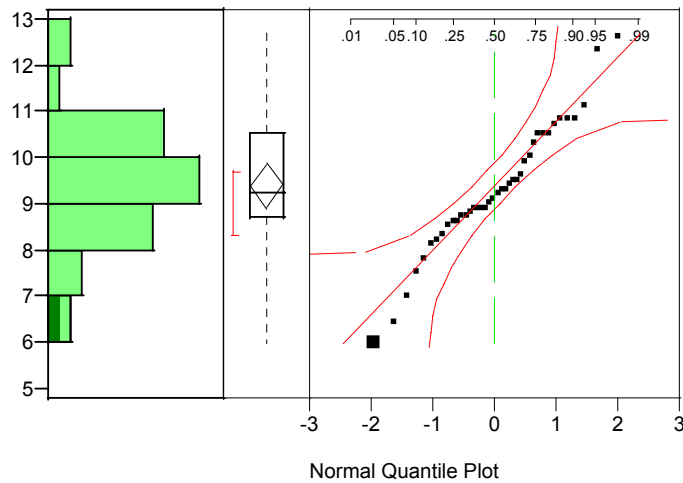
Moments

Mean	6.0898707
Std Dev	0.8110599
Std Err Mean	0.1282398
upper 95% Mean	6.3492603
lower 95% Mean	5.8304812
N	40

The normal plot shows a distribution looking much more normal. Mean and media are very similar and the differences between quartiles are also smaller. Though there still is one outlier.

5. Predictor Variable: Wind

As it can be seen in the outlier box plot there are no outliers. The minimum value=6 miles per hour, maximum=12.7 miles per hour, $Q_1=8.7$ miles per hour, $Q_2=9.25$ miles per hour, $Q_3=10.55$ miles per hour. These values suggest a symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



Quantiles

100.0%	maximum	12.700
99.5%		12.700
97.5%		12.693
90.0%		10.900
75.0%	quartile	10.550
50.0%	median	9.250
25.0%	quartile	8.700
10.0%		7.630
2.5%		6.013
0.5%		6.000
0.0%	minimum	6.000

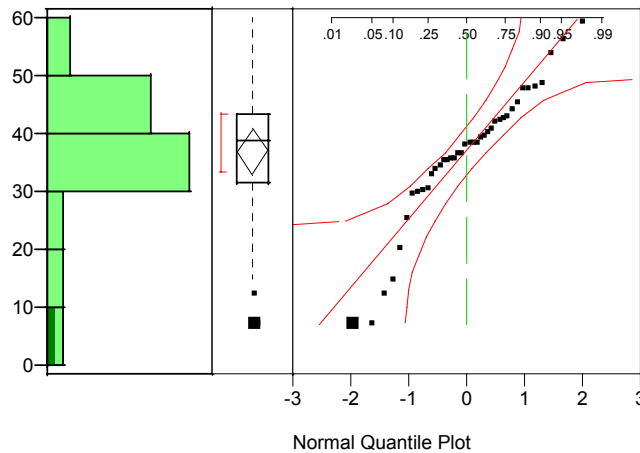
Moments

Mean	9.385
Std Dev	1.3955148
Std Err Mean	0.2206503
upper 95% Mean	9.8313073
lower 95% Mean	8.9386927
N	40

It can also be seen that $media \approx mean$, even though the media still smaller than the mean, hence there is a positive skew distribution.

6. Predictor Variable: Rain

As it can be seen in the outlier box plot there are three possible outliers, observations 1, 4 and 23 with 7.05, 12.95 and 7.77 inches respectively. The minimum value=7.05 inches, maximum=59.8 inches, $Q_1=43.305$ inches, $Q_2=38.765$ inches, $Q_3=31.56$ inches. These values suggest a symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



Quantiles

100.0%	maximum	59.800
99.5%		59.800
97.5%		59.724
90.0%		49.042
75.0%	quartile	43.305
50.0%	median	38.765
25.0%	quartile	31.560
10.0%		15.719
2.5%		7.068
0.5%		7.050
0.0%	minimum	7.050

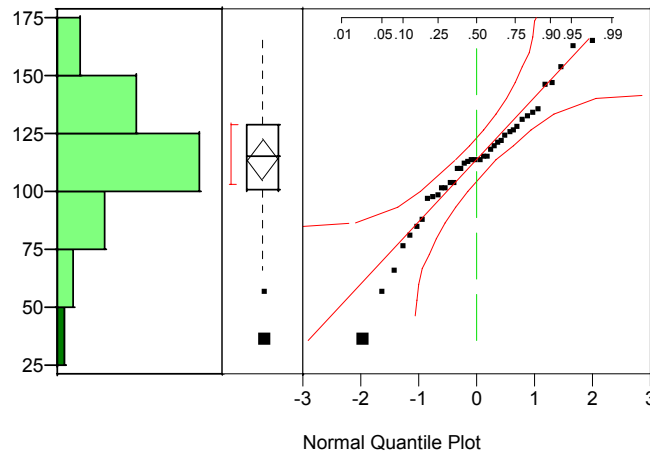
Moments

Mean	36.9615
Std Dev	11.855993
Std Err Mean	1.8745971
upper 95% Mean	40.753231
lower 95% Mean	33.169769
N	40

It can also be seen that $media \approx mean$, even though the mean still smaller than the media, hence there is a negative skew distribution.

7. Predictor Variable: Rain Days

As it can be seen in the outlier box plot there are two possible outliers, observations 1 and 23 with 36 and 58 days per year respectively. The minimum value=36 days per year, maximum=166 days per year, $Q_1=100.75$ days per year, $Q_2=115$ days per year, $Q_3=128.75$ days per year. These values suggest a symmetric distribution, since Q_1 and Q_3 are not equidistant from Q_2 .



Quantiles

100.0%	maximum	166.00
99.5%		166.00
97.5%		165.95
90.0%		147.90
75.0%	quartile	128.75
50.0%	median	115.00
25.0%	quartile	100.75
10.0%		78.40
2.5%		36.55
0.5%		36.00
0.0%	minimum	36.00

Moments

Mean	113.675
Std Dev	26.803547
Std Err Mean	4.2380129
upper 95% Mean	122.24719
lower 95% Mean	105.10281
N	40

It can also be seen that $media \approx mean$, even though the mean still smaller than the media, hence there is a negative skew distribution, as one would think, since the amount of rain days per year is usually related to the amount of precipitation in inches.

Summary

From the previous analysis one can state the following:

- The observation 11, corresponding to Chicago, is a possible outlier in three of the seven variables analyzed (SO_2 , Man and Pop).
- The observation 18, corresponding to Detroit, is a possible outlier in two of the seven variables analyzed (Man and Pop).
- The observation 23, corresponding to Albuquerque, is a possible outlier in two of the seven variables analyzed (Rain and Rain days).
- The observation 29, corresponding to Philadelphia, is a possible outlier in two of the seven variables analyzed (Man and Pop). It could be noticed that almost all the outliers from variables Man and Pop are the same. There could collinearity between them, this would explain this phenomenon.

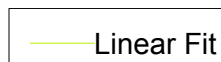
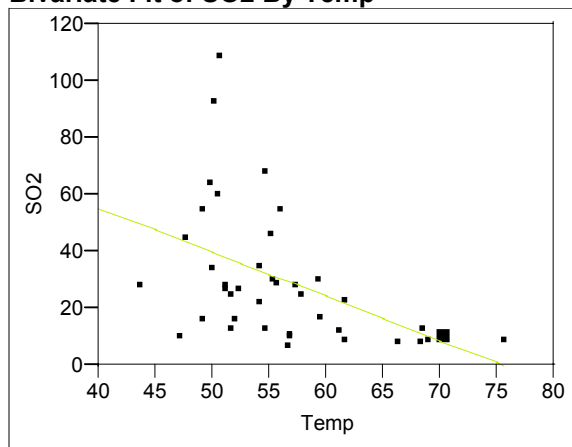
Bivariate Analysis

It is very useful to study bivariate relationships between the response variable and each of the predictor variables by making scatter plots.

1. SO₂ vs X predictors

The r^2 for the variables Temp and Man are not as good as one would like them to be, but they are certainly much more significant than other variables such as wind, rain and raindays. It does not seem they have a clear pattern to be sure the observations follow a certain trend or even that there is an evident transformation.

Bivariate Fit of SO₂ By Temp



Linear Fit

$$SO_2 = 117.48739 - 1.5547155 \text{ Temp}$$

Summary of Fit

RSquare	0.219822
RSquare Adj	0.199291
Root Mean Square Error	21.17326
Mean of Response	30.4
Observations (or Sum Wgts)	40

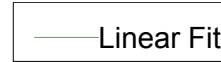
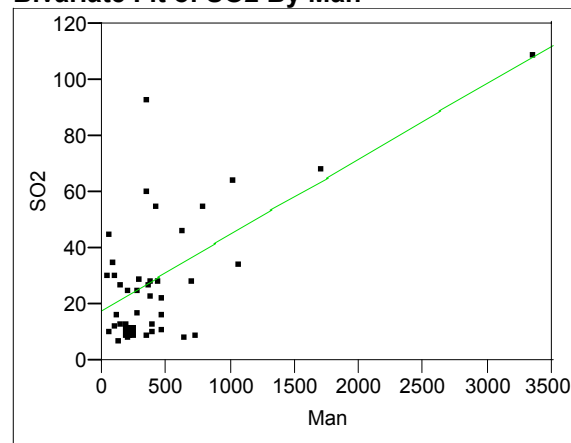
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4799.935	4799.94	10.7068
Error	38	17035.665	448.31	Prob > F
C. Total	39	21835.600		0.0023

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	117.48739	26.82466	4.38	<.0001
Temp	-1.554715	0.475139	-3.27	0.0023

Bivariate Fit of SO₂ By Man



Linear Fit

$$SO_2 = 17.96636 + 0.0270032 \text{ Man}$$

Summary of Fit

RSquare	0.423722
RSquare Adj	0.408557
Root Mean Square Error	18.19729
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

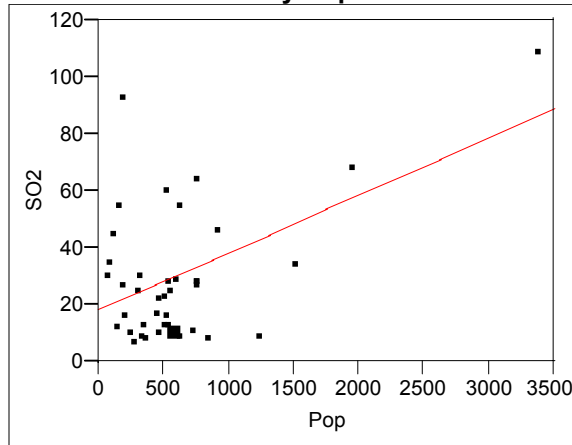
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	9252.221	9252.22	27.9404
Error	38	12583.379	331.14	Prob > F
C. Total	39	21835.600		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	17.96636	3.716394	4.83	<.0001
Man	0.0270032	0.005109	5.29	<.0001

The r^2 for the variable Pop is much better than the variable wind, rain and raindays, but still it has a low correlation with SO_2 . The variable wind in the other hand is not a good predictor, since there is practically no correlation with the response variable SO_2 . It does not seem they have a clear patron to be sure the observations follow a certain trend or even that there is a evident transformation. It looks like there is an influential observation for the variable Pop, but this will be analyzed in the next pages.

Bivariate Fit of SO2 By Pop



— Linear Fit

Linear Fit

$SO_2 = 18.192313 + 0.020148 \text{ Pop}$

Summary of Fit

RSquare	0.24917
RSquare Adj	0.229412
Root Mean Square Error	20.77119
Mean of Response	30.4
Observations (or Sum Wgts)	40

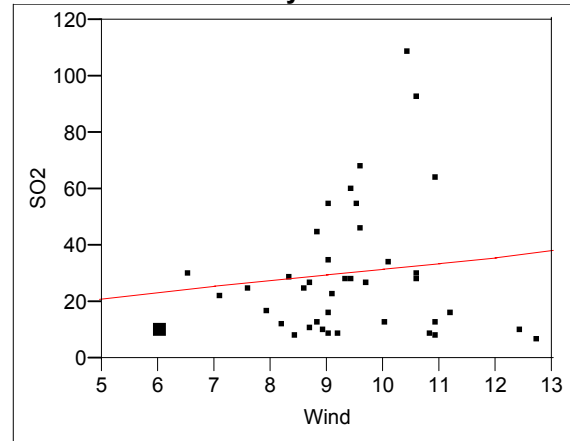
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5440.784	5440.78	12.6107
Error	38	16394.816	431.44	Prob > F
C. Total	39	21835.600		0.0010

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	18.192313	4.754328	3.83	0.0005
Pop	0.020148	0.005674	3.55	0.0010

Bivariate Fit of SO2 By Wind



— Linear Fit

Linear Fit

$SO_2 = 10.513206 + 2.1189978 \text{ Wind}$

Summary of Fit

RSquare	0.015618
RSquare Adj	-0.01029
Root Mean Square Error	23.78332
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

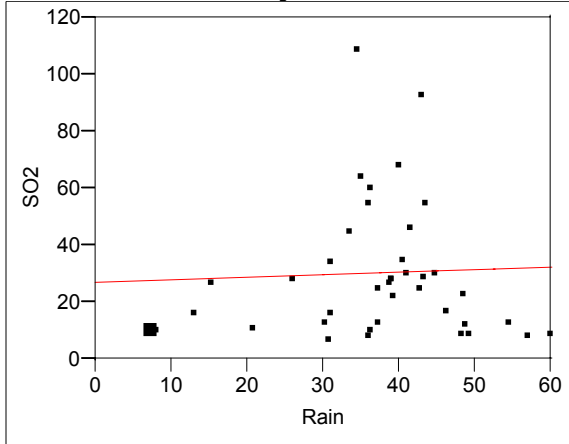
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	341.032	341.032	0.6029
Error	38	21494.568	565.647	Prob > F
C. Total	39	21835.600		0.4423

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	10.513206	25.88639	0.41	0.6869
Wind	2.1189978	2.729014	0.78	0.4423

The r^2 for both variables Rain and Raindays is extremely low. This indicates that there is practically no correlation with SO_2 . By this means one would expect that these two variables and the variable Wind should not be included in the final model, since their predictor power is null.

Bivariate Fit of SO2 By Rain



— Linear Fit

Linear Fit

$SO_2 = 27.101619 + 0.0892383 \text{ Rain}$

Summary of Fit

RSquare	0.001999
RSquare Adj	-0.02426
Root Mean Square Error	23.94728
Mean of Response	30.4
Observations (or Sum Wgts)	40

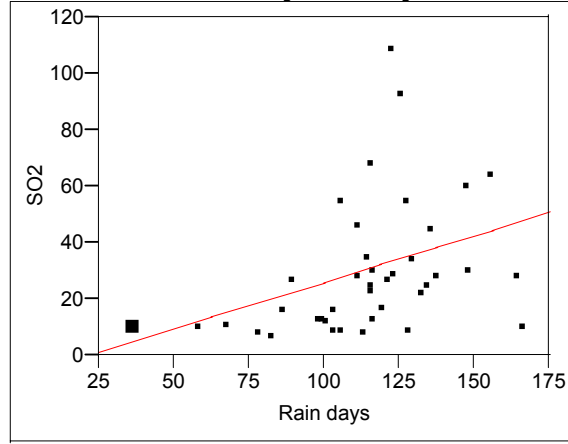
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	43.656	43.656	0.0761
Error	38	21791.944	573.472	Prob > F
C. Total	39	21835.600		0.7841

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	27.101619	12.53992	2.16	0.0370
Rain	0.0892383	0.323434	0.28	0.7841

Bivariate Fit of SO2 By Rain days



— Linear Fit

Linear Fit

$SO_2 = -7.44545 + 0.3329268 \text{ Rain days}$

Summary of Fit

RSquare	0.142227
RSquare Adj	0.119654
Root Mean Square Error	22.20123
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3105.607	3105.61	6.3008
Error	38	18729.993	492.89	Prob > F
C. Total	39	21835.600		0.0164

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-7.44545	15.48034	-0.48	0.6333
Rain days	0.3329268	0.132633	2.51	0.0164

2. Multicollinearity and Confidence Intervals

Multicollinearity can cause serious numerical and statistical difficulties in fitting the regression model unless extra predictors are deleted. Hence, in order to avoid difficulties multicollinearity is measured by three methods.

1.1. Correlation Matrix

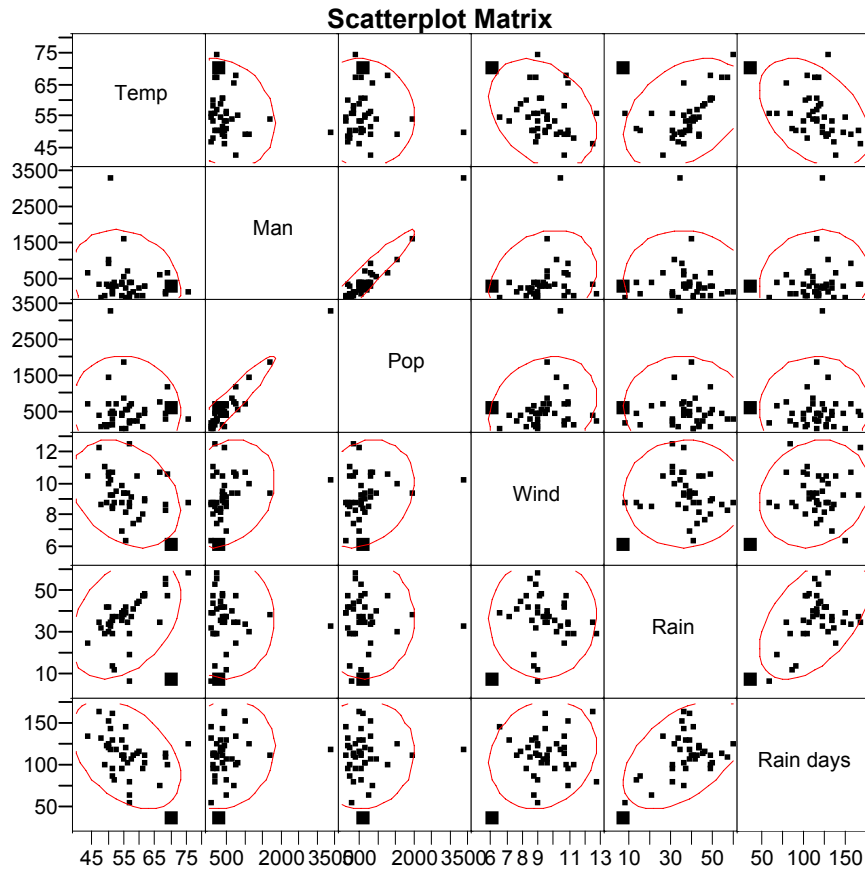
As it can be seen in the correlation matrix provided by JMP, there are two variables highly correlated: Man (the number of manufacturing enterprises) and Pop (population). In a certain way one would expect that the amount of companies are highly correlated to the places or cities where the populations are higher, hence this sort of explain the collinearity between the two variables.

Multivariate Correlations

	Temp	Man	Pop	Wind	Rain	Rain days
Temp	1.0000	-0.1882	-0.0575	-0.3094	0.3743	-0.4294
Man	-0.1882	1.0000	0.9552	0.2386	-0.0294	0.1304
Pop	-0.0575	0.9552	1.0000	0.2124	-0.0231	0.0405
Wind	-0.3094	0.2386	0.2124	1.0000	0.0153	0.1553
Rain	0.3743	-0.0294	-0.0231	0.0153	1.0000	0.5054
Rain days	-0.4294	0.1304	0.0405	0.1553	0.5054	1.0000

1.2. Scatterplot Matrix

To help you visualize the correlations, a scatterplot for each pair of response variables displays in a matrix arrangement. The correlation of the variables is seen by the collapsing of the ellipse along the diagonal axis. If the ellipse is fairly round and is not diagonally oriented, the variables are uncorrelated. On the contrary the scatterplot matrix shows that the variables Man and Pop are highly correlated, proving what was found in the correlation matrix.



1.3. Variance Inflation Factor (VIF) and Confidence Interval (CI)

A direct measure of multicollinearity is the VIF. In order for JMP to get the VIF value a multilinear regression was entered. The VIF values are highlighted for the variables Man and Pop, showing that these values are greater than 10. Therefore only one of these two variables should be deleted from the final model.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	112.90148	46.87245	2.41	0.0218	17.538768	208.26418	.
Temp	-1.353108	0.618749	-2.19	0.0359	-2.611961	-0.094254	3.618242
Man	0.0630477	0.015664	4.02	0.0003	0.0311784	0.0949169	14.817418
Pop	-0.037688	0.015038	-2.51	0.0173	-0.068283	-0.007093	14.424811
Wind	-2.709137	1.834513	-1.48	0.1492	-6.441482	1.0232075	1.2164963
Rain	0.5080641	0.359286	1.41	0.1667	-0.222909	1.2390376	3.3678871
Rain days	-0.055034	0.160474	-0.34	0.7338	-0.381521	0.2714541	3.4339698

From this JMP output table it is notorious that the CI of variables Wind, Rain and Rain days contain zero, which means, as commented before in the bivariate analysis, there is evidence these variables should be excluded from the model. This is corroborated by the variable selection method Stepwise, where the variables selected where Temp and Man as predictors (Pop was taken out because of colinearity with Man).

Response:
SO2

Stepwise Regression Control

Prob to Enter 0.050
Prob to Leave 0.050

Direction:

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
		Intercept	85.9560901	1	0	0.000	1.0000
X	X	Temp	-1.1907382	1	2715.894	10.184	0.0029
	X	Man	0.02420048	1	7168.18	26.878	0.0000
		Wind	0	1	393.7917	1.496	0.2292
		Rain	0	1	980.617	3.972	0.0539
		Rain days	0	1	579.0505	2.244	0.1428

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Man	Entered	0.0000	9252.221	0.4237	15.836	2
2	Temp	Entered	0.0029	2715.894	0.5481	6.648	3
3	Rain	Entered	0.0539	980.617	0.5930	4.6085	4
4	Wind	Entered	0.1193	603.5575	0.6207	4.1222	5
5	Wind	Removed	0.1193	603.5575	0.5930	4.6085	4
6	Rain	Removed	0.0539	980.617	0.5481	6.648	3

Candidate Models

Several models can be identified with the information developed in previous pages. In this way all of the possible predictor variables are fitted in the JMP subroutines for getting a final model for predicting the variable SO₂.

Models and Goodness of fit

Regression 1

In the output table one can realize some important matters, that were already disused. First of all, and even though the $r^2=0.68$, the variables Wind, Rain and Rain days (highlighted in red), have zero on their CI, therefore they will be deleted from future model tries. This can also be seen from the output table Effect Test, where the p-values are greater than 0.05. One of the variables with high correlation, Man or Pop (highlighted in blue), has to be deleted from the model. In order to do this new model try outs will be made for making sure that the less significant of these variables will be taken out.

Equation:

$$SO_2 = 112.90 - 1.35Temp + 0.06Man - 0.04Pop - 2.71Wind + 0.51Rain - 0.06Raindays$$

Summary of Fit

RSquare	0.682449
RSquare Adj	0.624713
Root Mean Square Error	14.49546
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	14901.690	2483.61	11.8201
Error	33	6933.910	210.12	Prob > F
C. Total	39	21835.600		<.0001

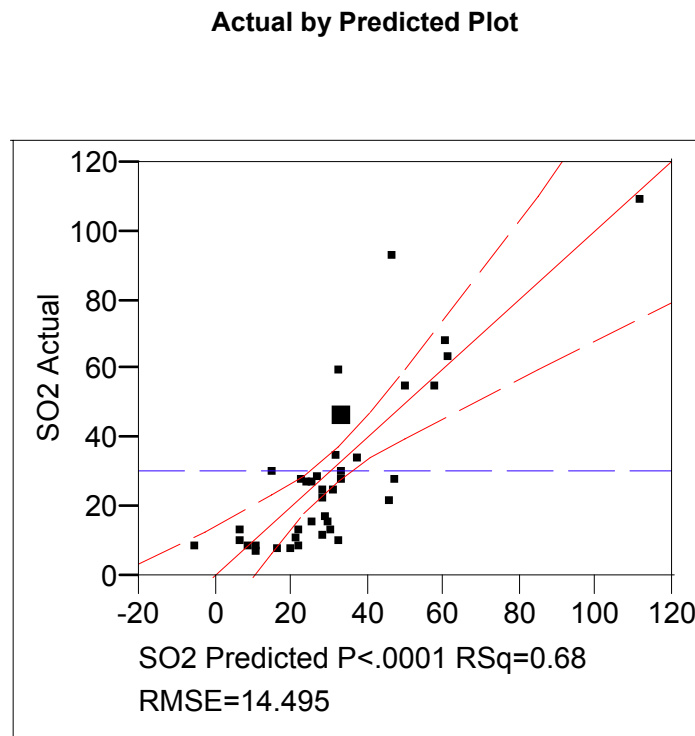
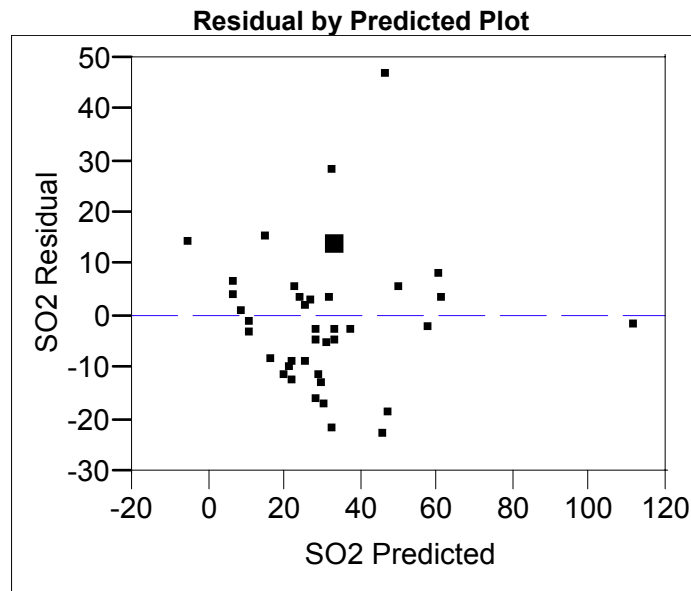
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	112.90148	46.87245	2.41	0.0218	17.538768	208.26418	.
Temp	-1.353108	0.618749	-2.19	0.0359	-2.611961	-0.094254	3.618242
Man	0.0630477	0.015664	4.02	0.0003	0.0311784	0.0949169	14.817418
Pop	-0.037688	0.015038	-2.51	0.0173	-0.068283	-0.007093	14.424811
Wind	-2.709137	1.834513	-1.48	0.1492	-6.441482	1.0232075	1.2164963
Rain	0.5080641	0.359286	1.41	0.1667	-0.222909	1.2390376	3.3678871
Rain days	-0.055034	0.160474	-0.34	0.7338	-0.381521	0.2714541	3.4339698

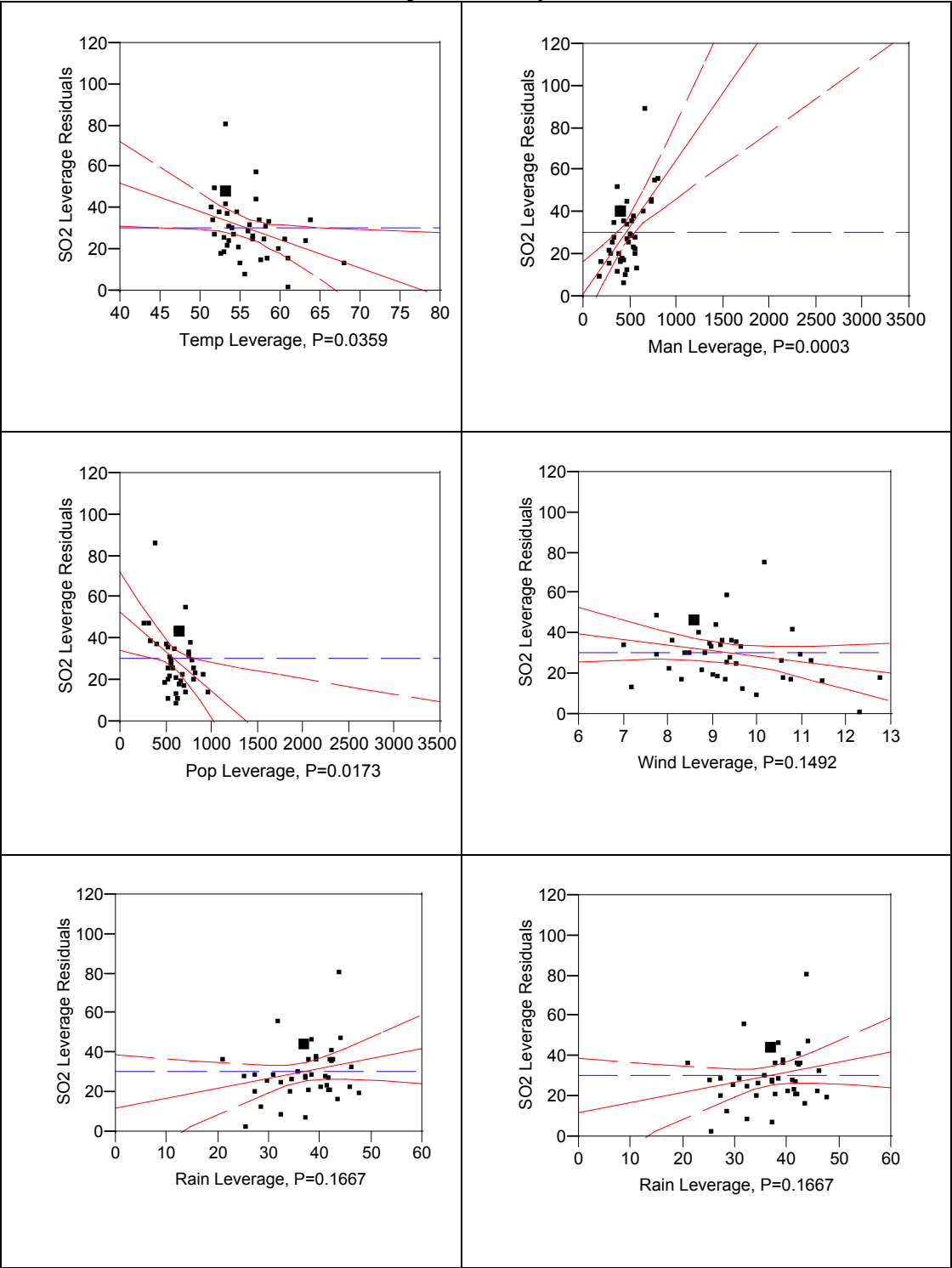
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	1004.8485	4.7823	0.0359
Man	1	1	3403.9250	16.2000	0.0003
Pop	1	1	1319.7435	6.2809	0.0173
Wind	1	1	458.2313	2.1808	0.1492
Rain	1	1	420.1651	1.9997	0.1667
Rain days	1	1	24.7120	0.1176	0.7338

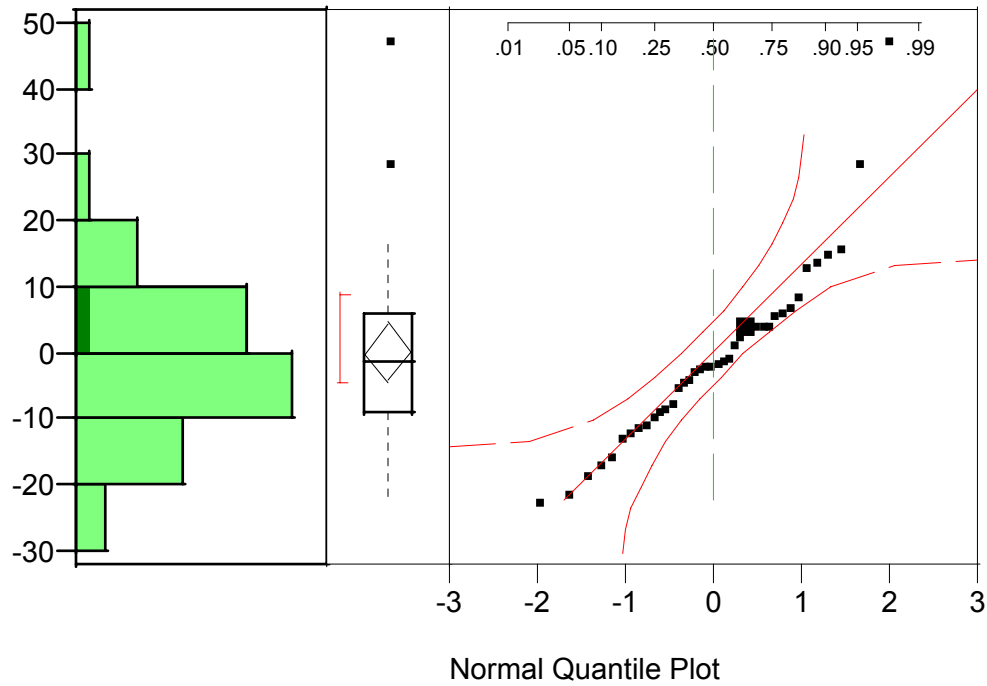
The residual plots will be analyzed in those regressions that seem to be a better model.



Residual by Individual predictor values



Normal Plot of the residual



Regression 2

In this regression candidate the variables taken into the model are Tem and Man. The correlation between the predictors and the response variable is 54.8 %.

Equation: $SO_2 = 85.96 - 1.19Temp + 0.02Man$

Summary of Fit

RSquare	0.548101
RSquare Adj	0.523674
Root Mean Square Error	16.33061
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	11968.115	5984.06	22.4384
Error	37	9867.485	266.69	Prob > F
C. Total	39	21835.600		<.0001

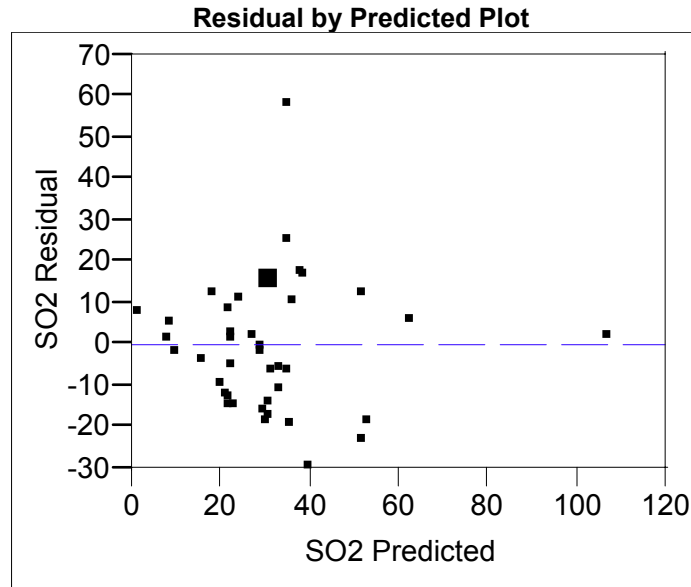
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	85.95609	21.56485	3.99	0.0003	42.261561	129.65062	.
Temp	-1.190738	0.373132	-3.19	0.0029	-1.946775	-0.434701	1.0367006
Man	0.0242005	0.004668	5.18	<.0001	0.0147424	0.0336585	1.0367006

Effect Tests

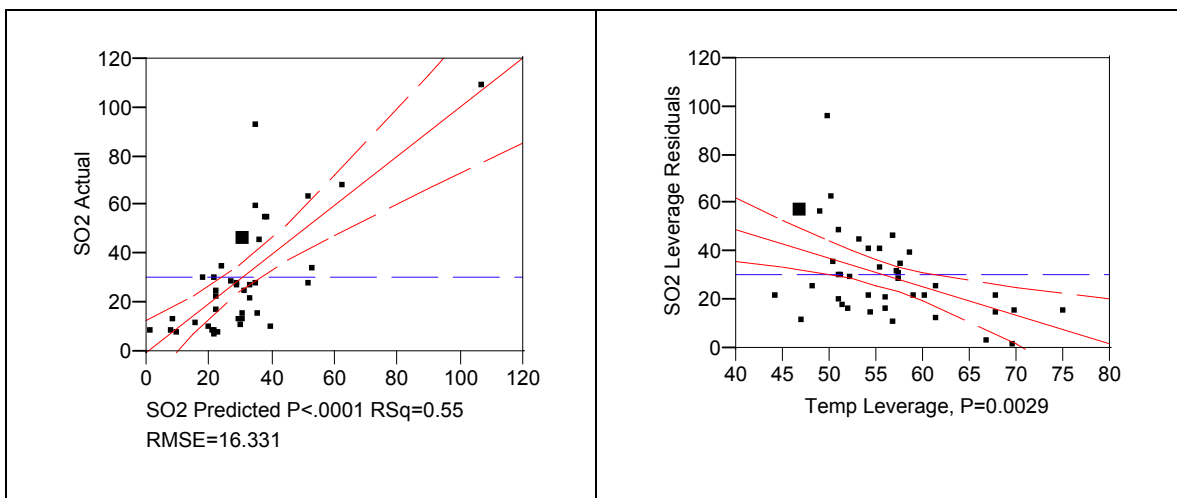
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	2715.8936	10.1838	0.0029
Man	1	1	7168.1797	26.8784	<.0001

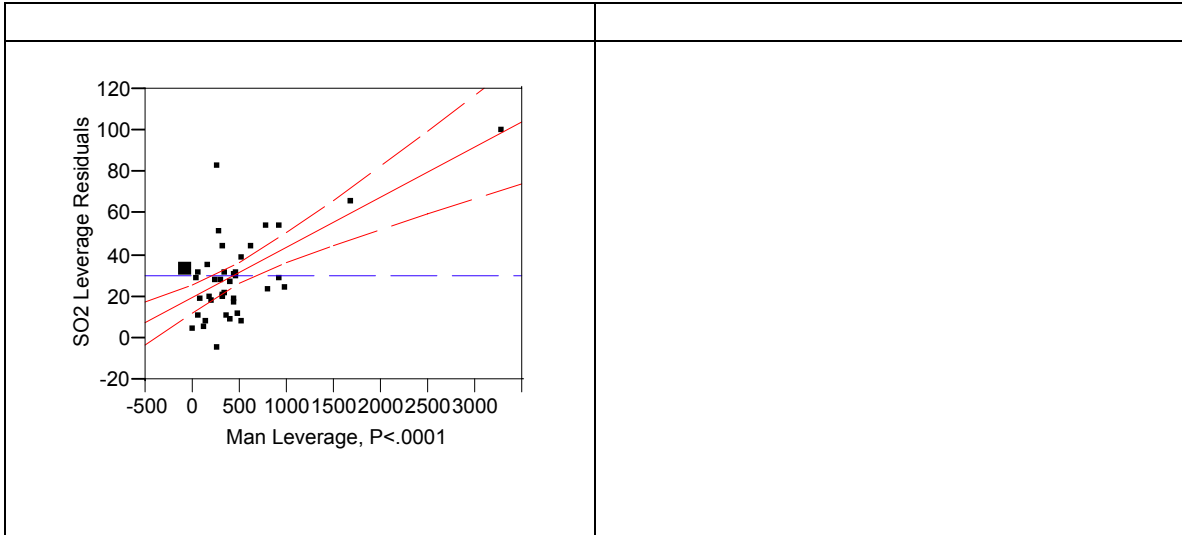
The plot below, Residuals against fitted values, does not shows a clear pattern, but it could be that the variability in the residuals could slightly show a pattern, sort of a cone, which means that $\text{Var}(Y)_{\alpha\mu^2}$ or $\text{SD}(Y)_{\alpha\mu}$. A transformation could be made in order to clarify this doubt.



The plots of the residuals against individual predictor variables, Temp and Man, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

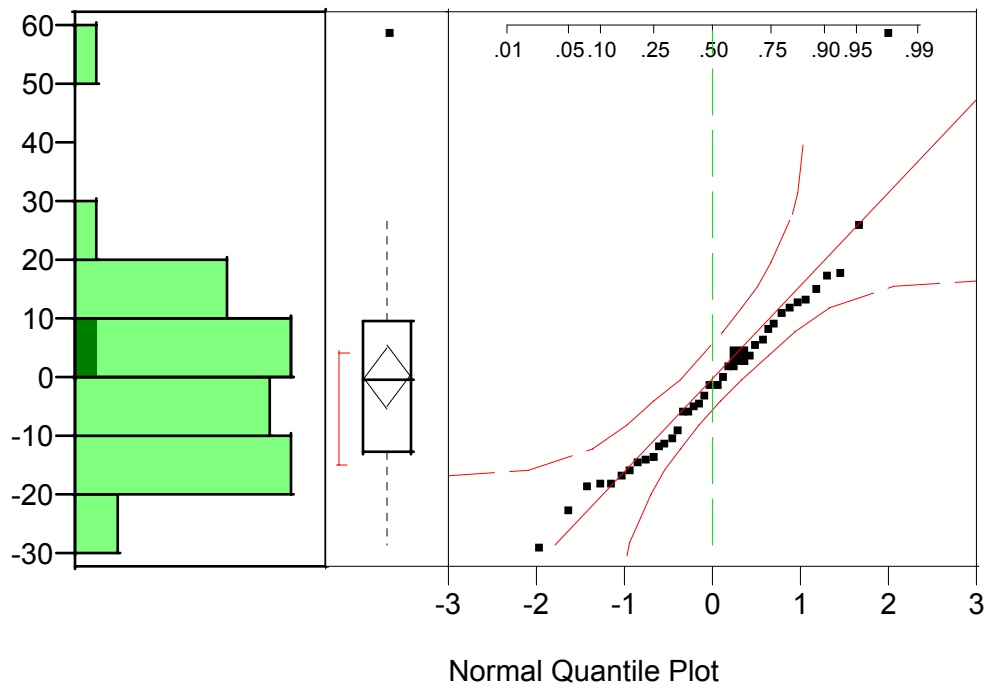
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Regression 3

In this regression candidate the variables taken into the model are Tem and Pop. The idea is to identify if Man or Pop is the most significant variable for the model. The correlation between the predictors and the response variable is 44.4 %. Because of the r^2 value, the variable Man appears to be statistically more significant than Pop for this kind of multilinear regression.

Equation: $SO_2 = 100.84 - 1.46Temp + 0.02Pop$

Summary of Fit

RSquare	0.443556
RSquare Adj	0.413478
Root Mean Square Error	18.12142
Mean of Response	30.4
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9685.317	4842.66	14.7468
Error	37	12150.283	328.39	Prob > F
C. Total	39	21835.600		<.0001

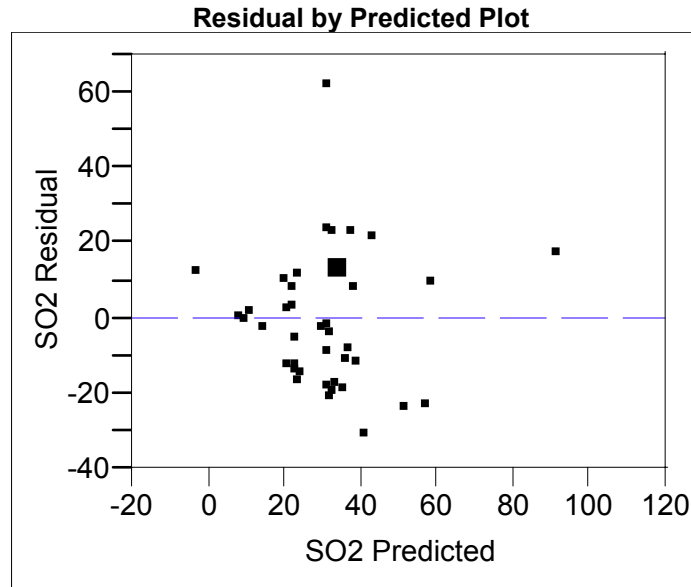
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	100.84268	23.3603	4.32	0.0001
Temp	-1.464423	0.407328	-3.60	0.0009
Pop	0.0191236	0.004958	3.86	0.0004

Effect Tests

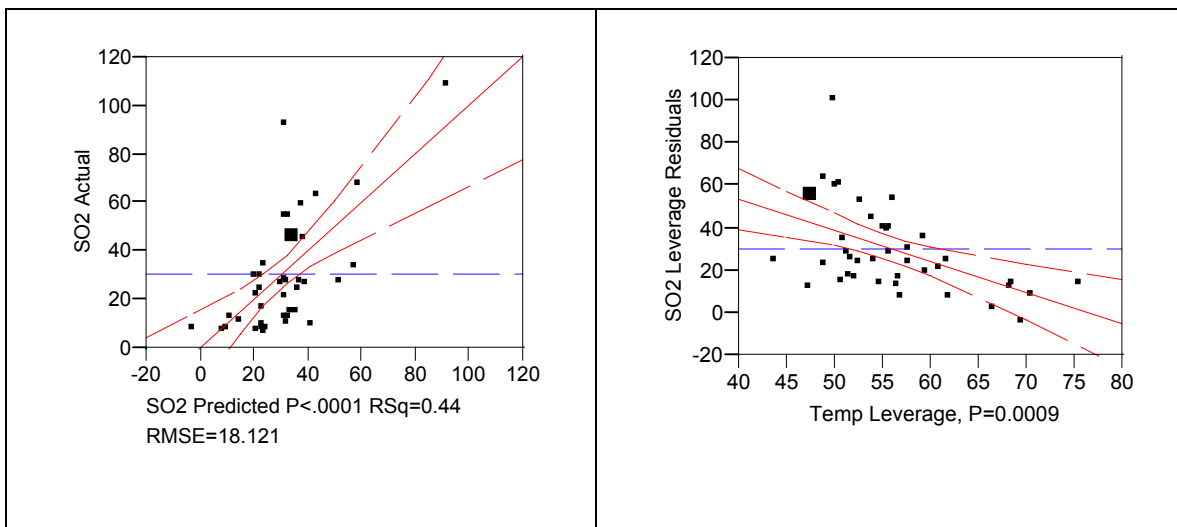
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	4244.5330	12.9254	0.0009
Pop	1	1	4885.3820	14.8769	0.0004

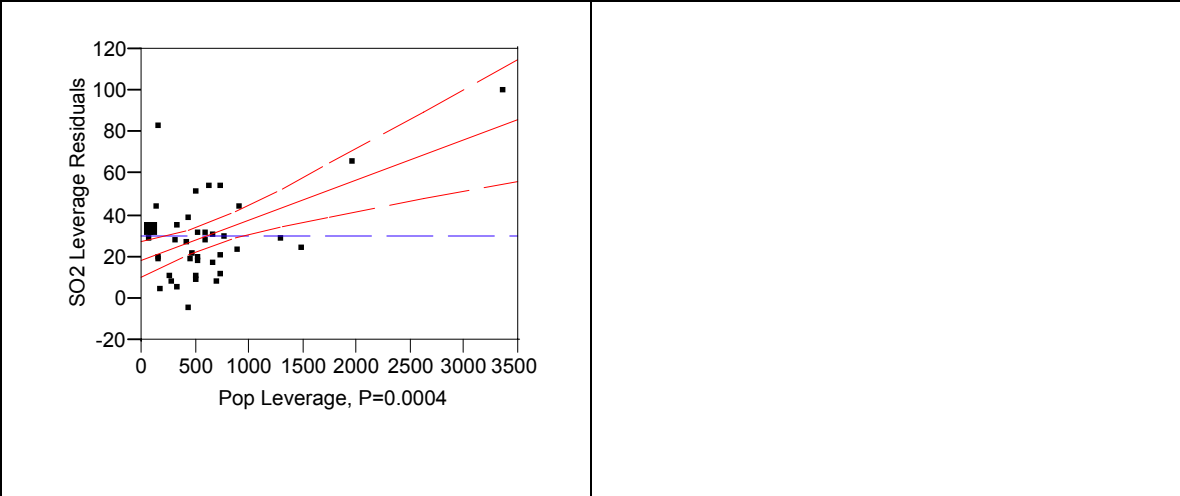
The following plot, Residuals against fitted values, does not shows a clear pattern, but it could be that the variability in the residuals could slightly show a pattern, and like mentioned in the previous regression, a transformation could be helpful to clarify the doubt.



The plots of the residuals against individual predictor variables, Temp and Pop, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

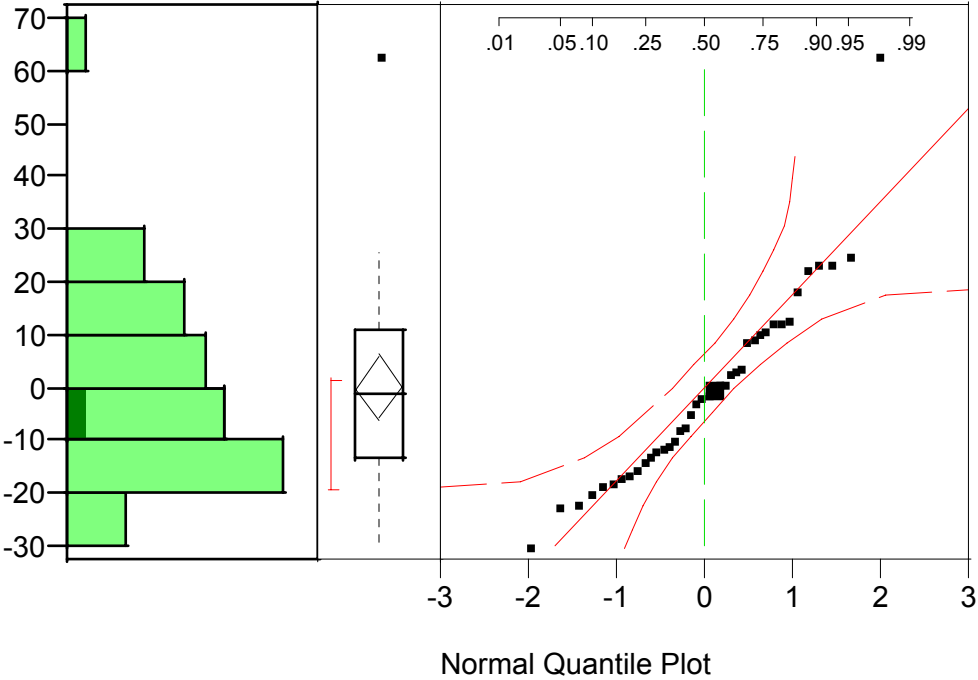
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Regression 4

In this regression the variables taken into the model are the natural logarithm of Tem and Man and $\ln SO_2$. Because the correlation between the predictor variables and the response variable are not good enough a transformation was tried. The transformation used is the power form, where $y = \alpha x_1^\beta x_2^\delta$. The correlation between the predictors and the response variable is 36.5%. Because of the r^2 value, the variable $\ln Man$ appears to be statistically more significant than $\ln Pop$ for this kind of multilinear regression.

Equation: $\ln SO_2 = 14.49 - 3.10 \ln Temp + 0.20 \ln Man$

Summary of Fit

RSquare	0.405761
RSquare Adj	0.37364
Root Mean Square Error	0.56078
Mean of Response	3.162514
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	7.945043	3.97252	12.6323
Error	37	11.635553	0.31447	Prob > F
C. Total	39	19.580596		<.0001

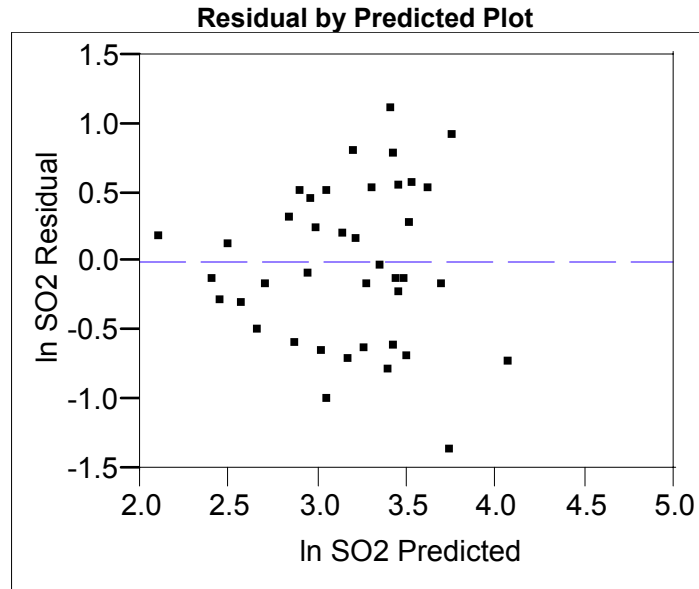
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	14.489273	3.093285	4.68	<.0001
In Temp	-3.102739	0.73966	-4.19	0.0002
In Man	0.2009224	0.093515	2.15	0.0383

Effect Tests

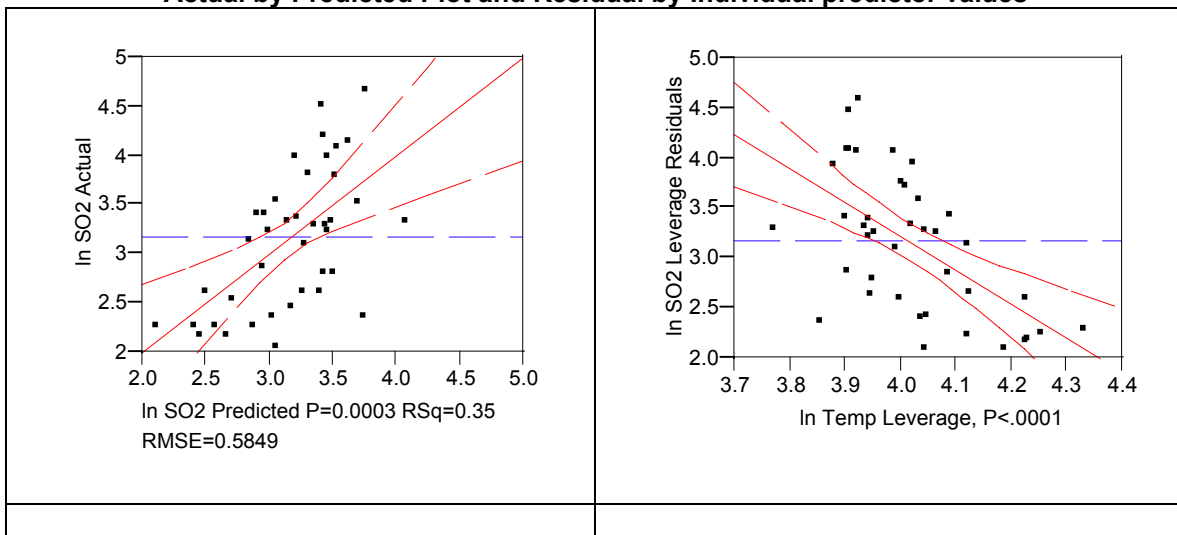
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
In Temp	1	1	5.5336503	17.5965	0.0002
In Man	1	1	1.4517026	4.6163	0.0383

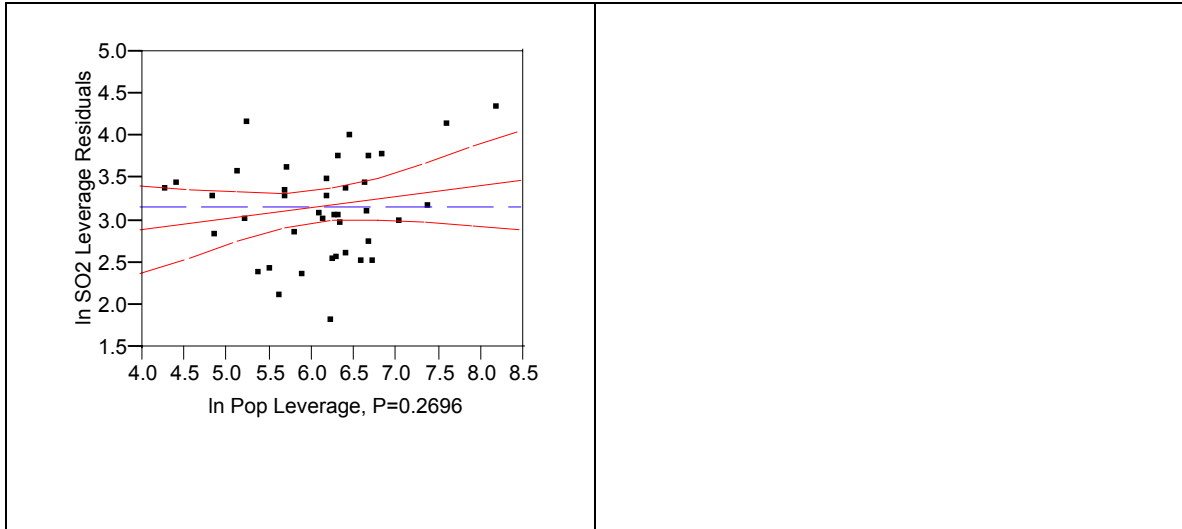
The following plot, Residuals against fitted values, does not shows a clear pattern, but it could be that the variability in the residuals could slightly show a pattern, and like mentioned in the previous regressions, a transformation could be helpful to clarify this doubt. This will be held in the next pages.



The plots of the residuals against individual predictor variables, $\ln(\text{Temp})$ and $\ln(\text{Man})$, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

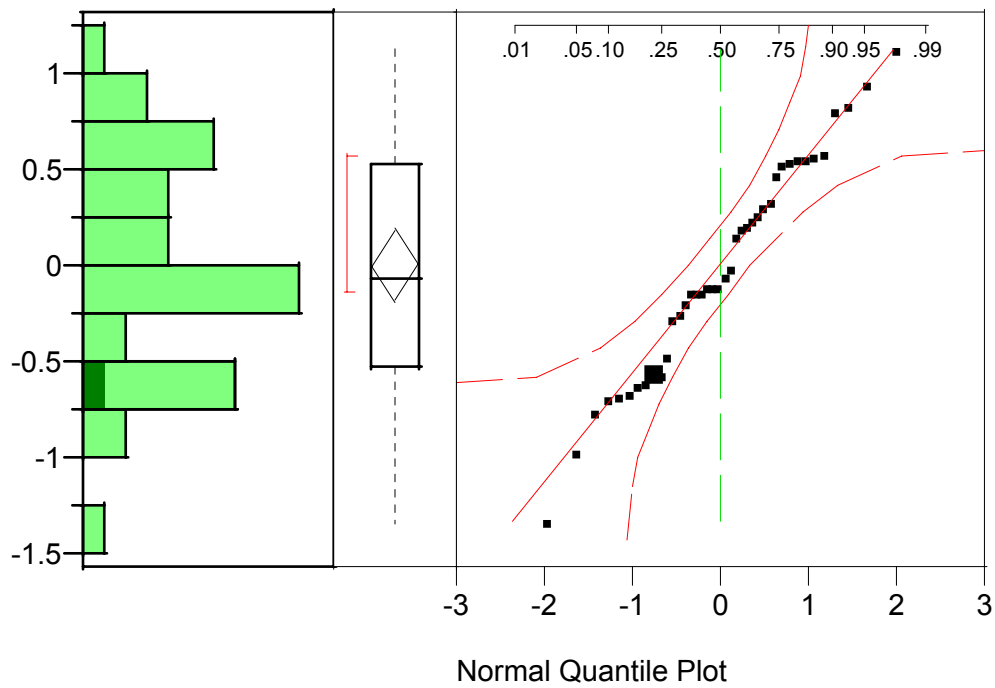
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Regression 5

In this regression candidate the variables taken into the model are $\ln \text{Temp}$ and $\ln \text{Pop}$. The idea is to identify if Man or Pop is the most significant variable for this particular model. The correlation between the predictors and the response variable is 35.4 %. Because of the r^2 value, the variable Man appears to be statistically more significant than Pop for this kind of non linear regression.

Equation: $\ln SO_2 = 15.97 - 3.38 \ln \text{Temp} + 0.13 \ln \text{Pop}$

Summary of Fit

RSquare	0.353567
RSquare Adj	0.318625
Root Mean Square Error	0.584889
Mean of Response	3.162514
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	6.923058	3.46153	10.1186
Error	37	12.657538	0.34210	Prob > F
C. Total	39	19.580596		0.0003

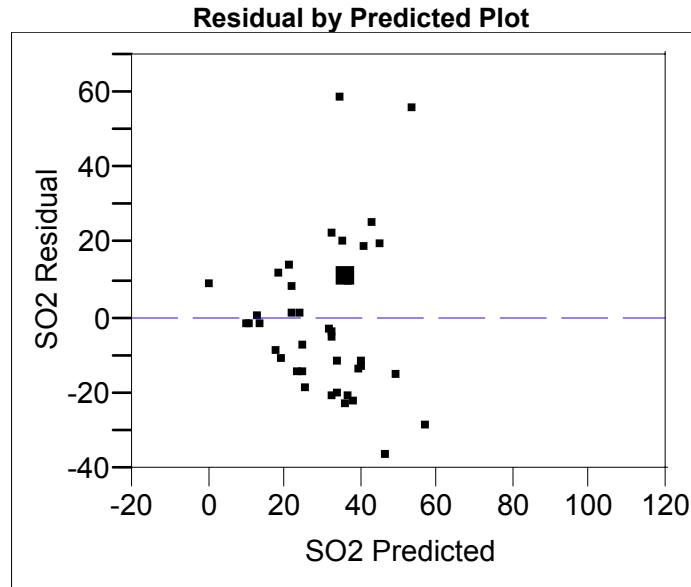
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	15.97116	3.110615	5.13	<.0001
$\ln \text{Temp}$	-3.384314	0.76541	-4.42	<.0001
$\ln \text{Pop}$	0.1297081	0.115731	1.12	0.2696

Effect Tests

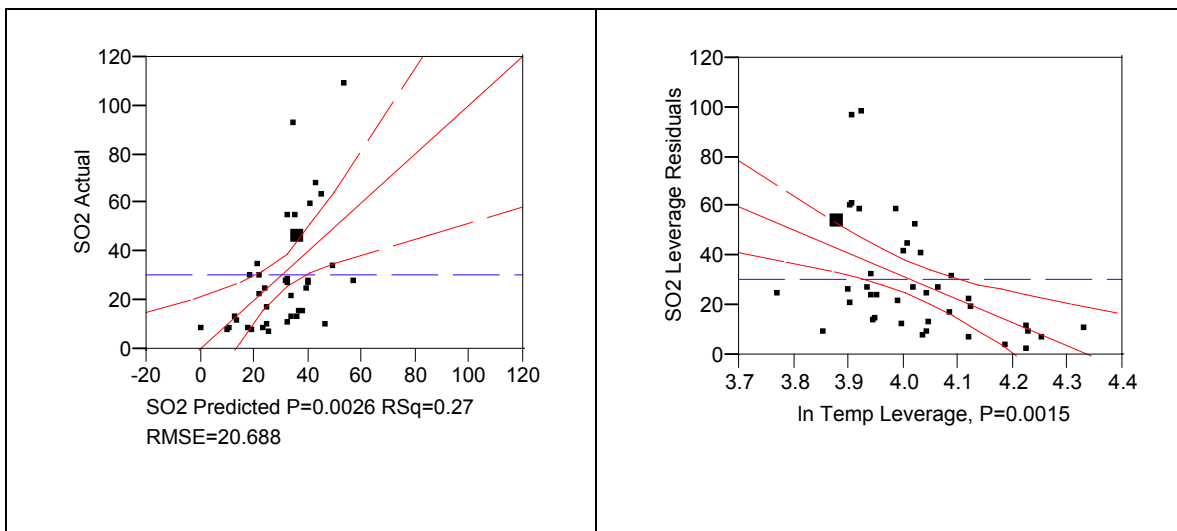
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
$\ln \text{Temp}$	1	1	6.6880688	19.5503	<.0001
$\ln \text{Pop}$	1	1	0.4297177	1.2561	0.2696

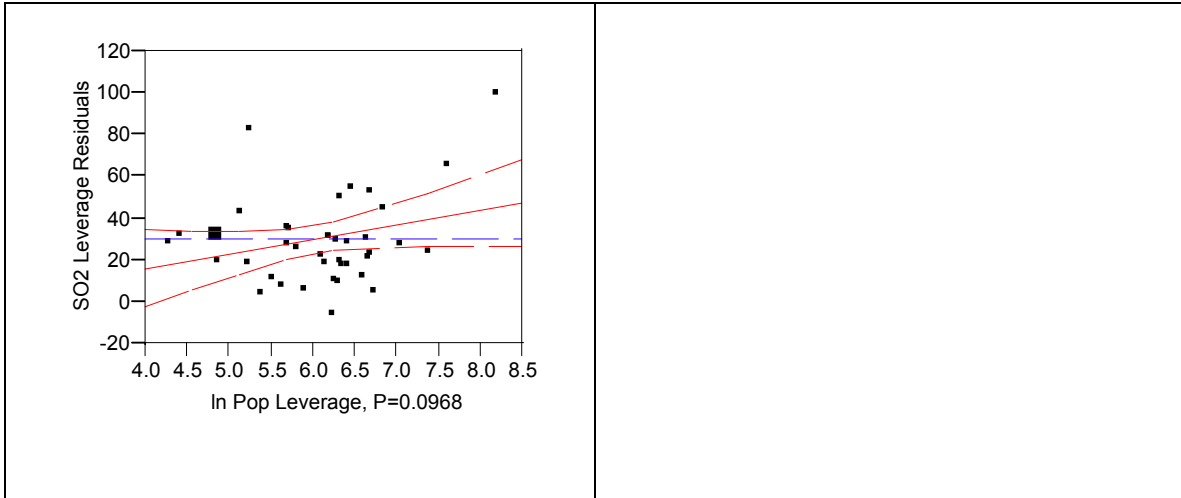
The following plot, Residuals against fitted values, like before does not shows a clear pattern, there could be a possibility that the variability in the residuals could slightly show a pattern, a con kind of pattern, and like mentioned in the previous regressions analysis, a transformation could be helpful to clarify this doubt. This will be held in the next pages.



The plots of the residuals against individual predictor variables, $\ln(\text{Temp})$ and $\ln(\text{Pop})$, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

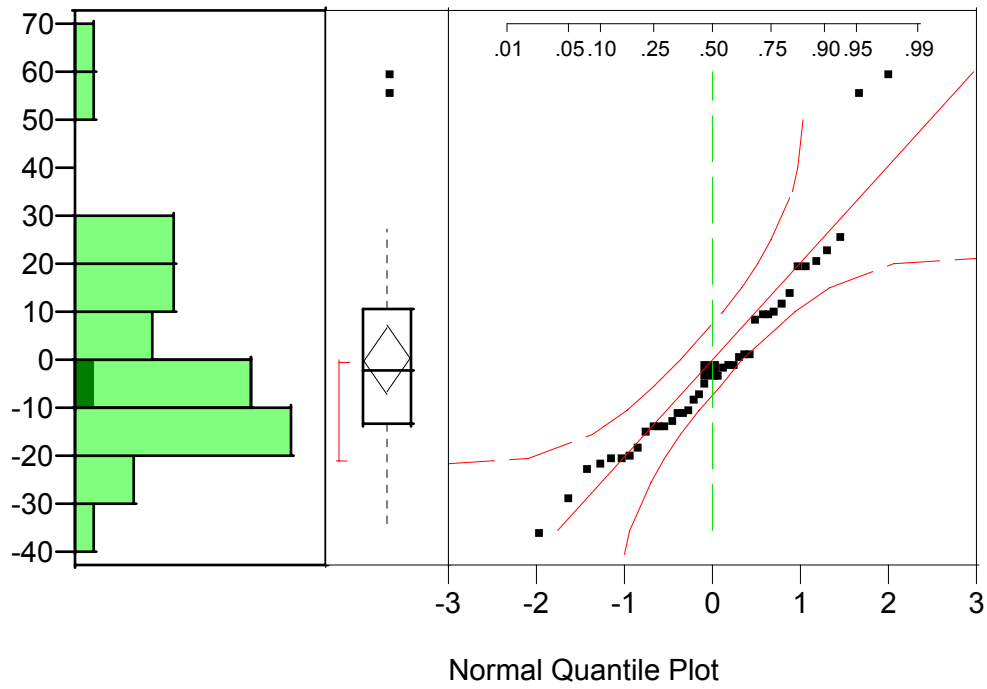
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Regression 6

In this regression the variables taken into the model are the natural logarithm of the response variable SO_2 . The general idea is to verify if this transformation will stabilize the variance. Because of the appearing of a cone shape in the plots of residuals vs. the predicted values, $SD(Y) \propto \mu$ the proper transformation is a logarithmic one. The correlation between the predictors and the response variable is 48.9%.

Equation: $\ln SO_2 = 5.77 - 0.05Temp + .0005Man$

Summary of Fit

RSquare	0.488619
RSquare Adj	0.460977
Root Mean Square Error	0.520217
Mean of Response	3.162514
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9.567451	4.78373	17.6765
Error	37	10.013145	0.27063	Prob > F
C. Total	39	19.580596		<.0001

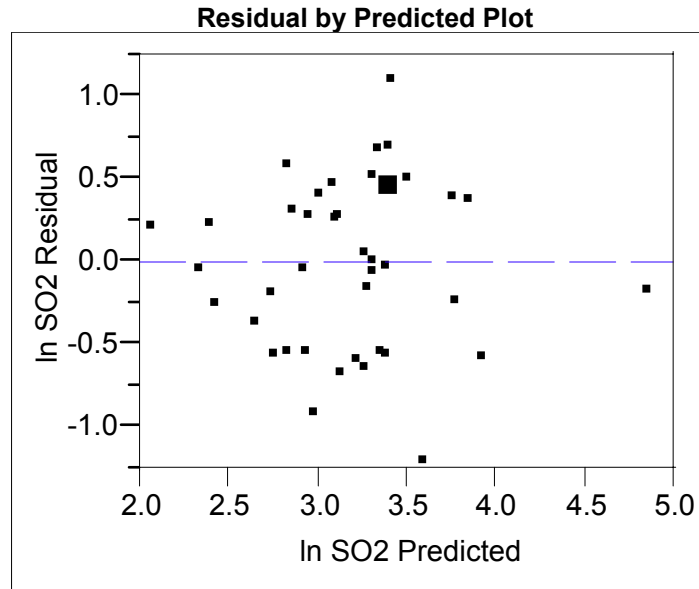
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	5.7663717	0.686955	8.39	<.0001
Temp	-0.050501	0.011886	-4.25	0.0001
Man	0.0004886	0.000149	3.29	0.0022

Effect Tests

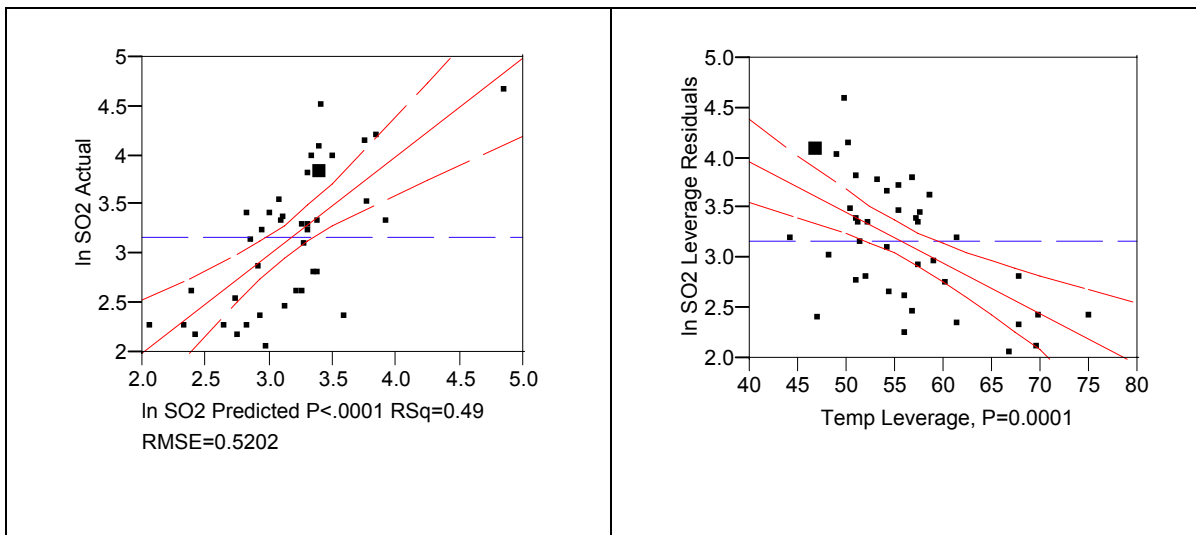
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	4.8852309	18.0516	0.0001
Man	1	1	2.9218159	10.7965	0.0022

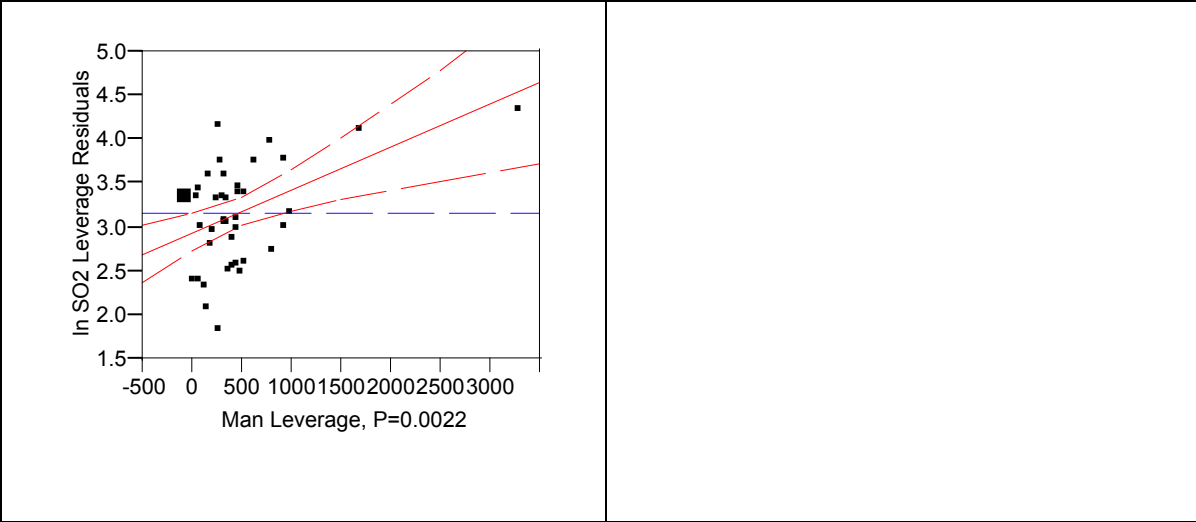
The following plot, Residuals against fitted values, like before does not shows a clear pattern, and even more, the logarithm transformation does seem to help to stabilize the variance. Hence the transformation is useful in this means. Now, comparing the correlation between the response and the predictor variables a new try will be held using this transformation, but using the predictor Pop instead of Man.



The plots of the residuals against individual predictor variables, Temp and Man, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

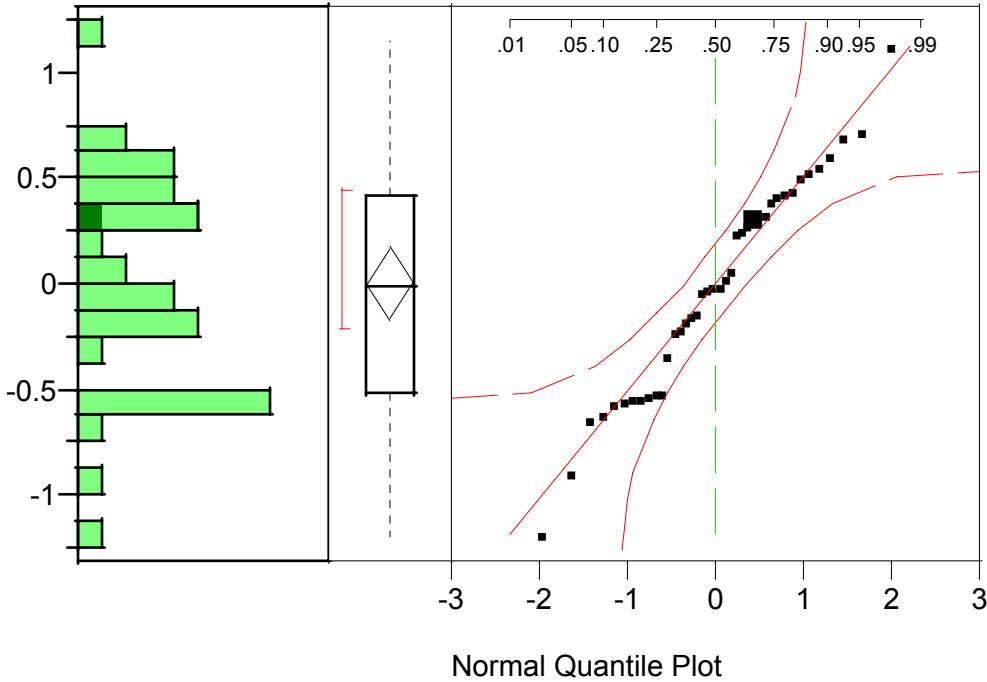
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Regression 7

In this regression the variables taken into the model are the natural logarithm of the response variable SO_2 . The general idea is to verify if this transformation will stabilize the variance. Because of the appearing of a cone shape in the plots of residuals vs. the predicted values, $SD(Y) \propto \mu$ the proper transformation is a logarithmic one. The correlation between the predictors, Temp and Pop, and the response variable is 43.8%. This shows ones more that Man is a better predictor than Pop.

Equation: $\ln SO_2 = 6.07 - 0.06Temp + .0004Pop$

Summary of Fit

RSquare	0.438341
RSquare Adj	0.407981
Root Mean Square Error	0.545191
Mean of Response	3.162514
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.582978	4.29149	14.4381
Error	37	10.997618	0.29723	Prob > F
C. Total	39	19.580596		<.0001

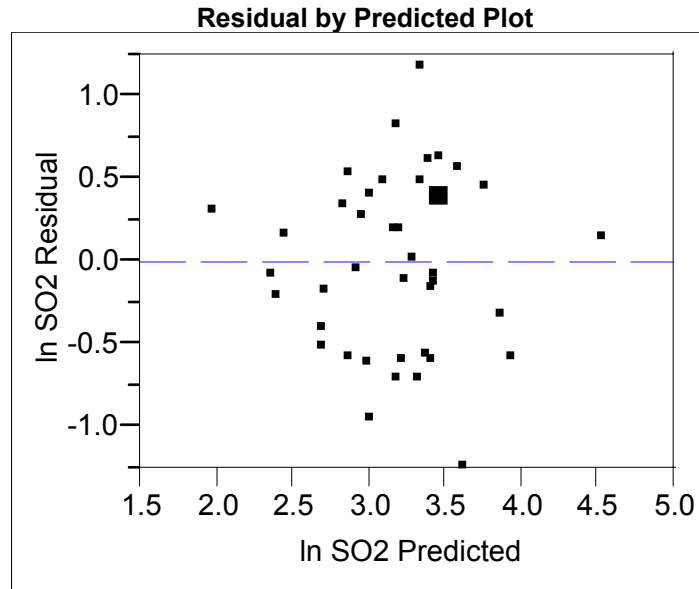
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	6.0715084	0.702805	8.64	<.0001	4.6474911	7.4955256	.
Temp	-0.056052	0.012255	-4.57	<.0001	-0.080882	-0.031221	1.0033139
Pop	0.0003808	0.000149	2.55	0.0149	0.0000786	0.0006831	1.0033139

Effect Tests

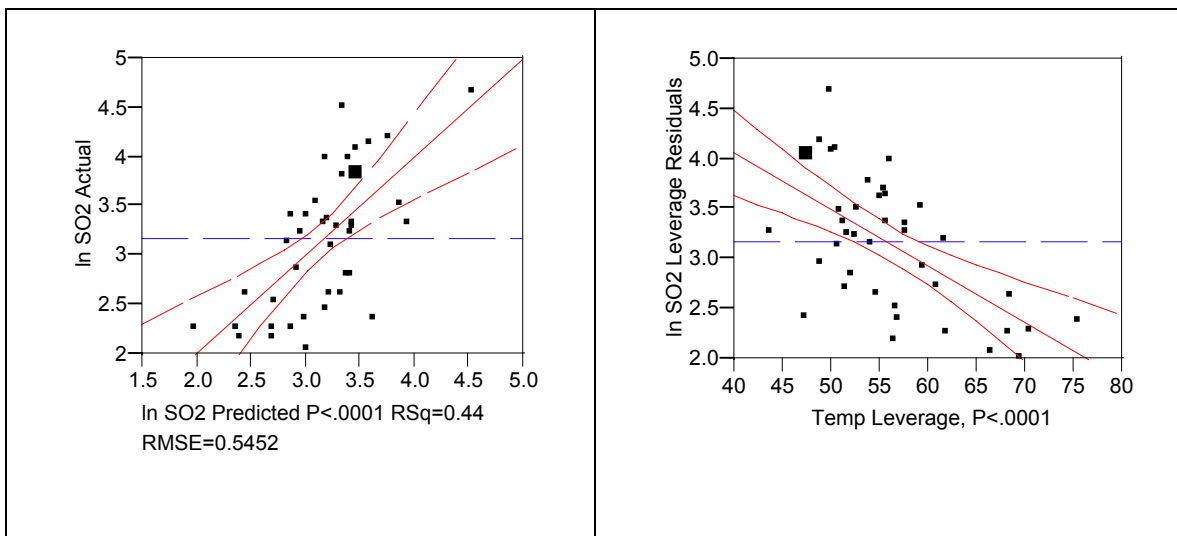
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	6.2183343	20.9207	<.0001
Pop	1	1	1.9373431	6.5179	0.0149

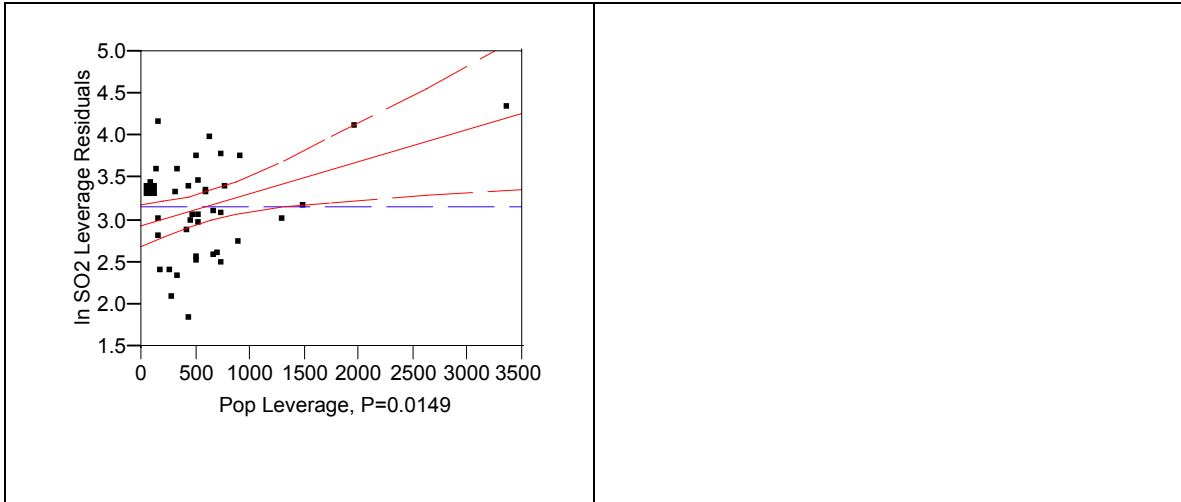
The following plot, Residuals against fitted values, like before does not shows a clear pattern, and even more, the logarithm transformation, like in the previous regression, seems to help to stabilize the variance.



The plots of the residuals against individual predictor variables, Temp and Pop, shows that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

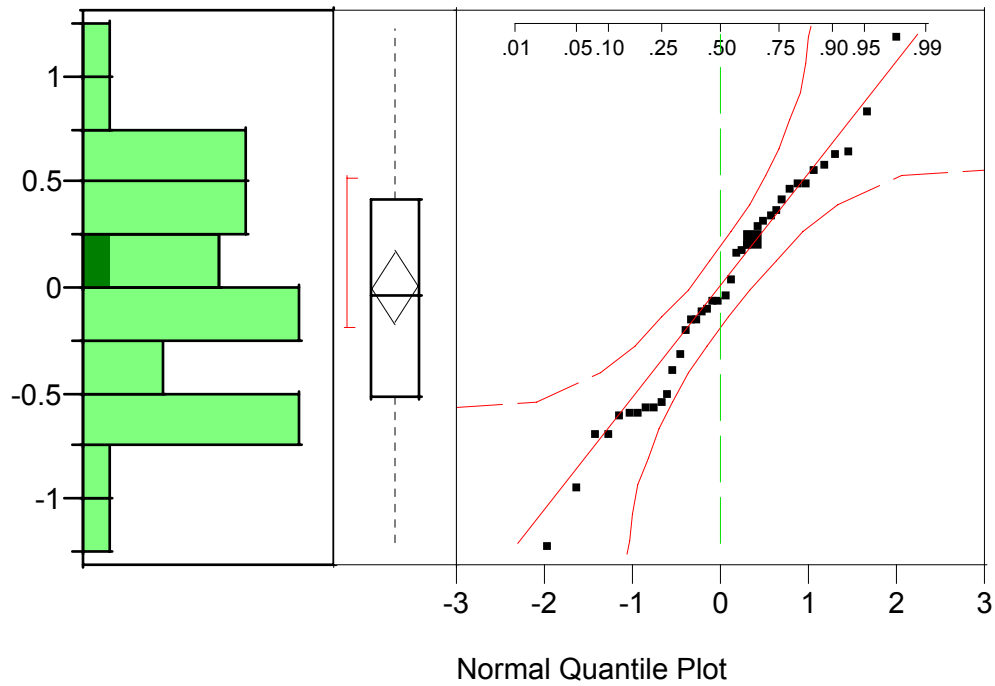
Actual by Predicted Plot and Residual by Individual predictor values





The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



Selection of the Final Model

In this step the competing models are compared by cross validation of them against the test data. The models with the best characteristics, not violating the assumptions are going to be selected. Then the model with a smaller SSE or r-square is going to be the better predictive model. The final selection of the model is going to be based in several considerations. These considerations include residual plots and outliers among others. By this there are two models that accomplish all of the assumptions. As discussed before Regressions 6 and Regression 7, are the best predictors. A comparison between the outliers from both regressions is shown to verify the best characteristics from one to the other model.

From the table of outliers and influential data, there can be seen that both regressions have 2 outliers, which will not be eliminated, because there is no evidence of erroneous measurements. In fact the outliers correspond to the cities of Miami and Chicago, two of the biggest cities in United States. However looking at the influential data regression 7 has 3 influent observations, one more than regression 6. Therefore by this means regression 7 would be a better choice. Plus the r-square value for regression 6 is also better than for regression 7. This corroborates the past analysis stating that Man is a better predictor than Pop.

With all the analysis results, it can be stated that the best model founded in this examination is regression 6.

Regression 6 equation:

$$\ln SO_2 = 5.77 - 0.05Temp + .0005Man$$

with an r-square = 0.48.9, that states a correlation between the predictor variables Temp and Man of 48.9 % with respect to the response variable Ln SO₂. The results actual prediction of this model is not good. The reason why this is not a good model might lay on the data. A bigger number of observations would it make a clearer trend. In addition it is also a fact that the data used in this examination was taken over the years 1969-1971, more than 30 years ago, with lower technology, and fewer knowledge in air pollution, since during the past decades there is been great development in the area.

Outliers and Influential Data

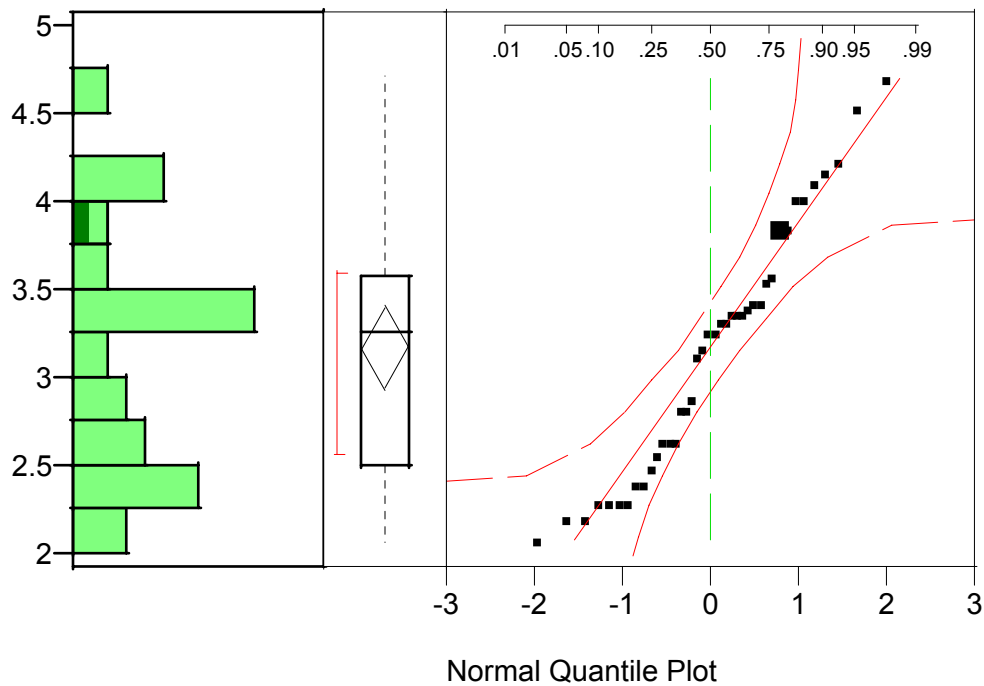
No	City	REGRESSION 6		REGRESSION 7	
		Studentized Residual	h_{ii}	Studentized Residual	h_{ii}
1	Phoenix	-0.03626	0.12784752	-0.0984594	0.12790263
2	Little Rock	-0.325095	0.04459896	-0.2594549	0.05269732
3	San Francisco	-1.2448614	0.02523696	-1.266056	0.02619782
4	Denver	-1.0442799	0.03390878	-0.9806164	0.03443841
5	Hartford	1.06020349	0.05097861	1.22593877	0.06636654
6	Wilmington	0.99144916	0.04082962	0.95586973	0.04850413
7	Washington	0.55038993	0.02583567	0.40786658	0.02768063
8	Jacksonville	0.52900983	0.10380287	0.38723153	0.10226829
9	Miami	0.53834591	0.21631909	0.69595052	0.21858682
10	Atlanta	0.66261205	0.04015834	0.68251234	0.04066598
11	Chicago	-0.4915966	0.68128377	0.5288945	0.60076093
12	Indianapolis	0.05997967	0.03392112	-0.1714139	0.03307442
13	Des Moines	-1.0138919	0.06721692	-1.0775332	0.0641437
14	Wichita	-1.7398522	0.03389017	-1.7255364	0.03313471
15	Louisville	0.58582026	0.02760871	0.40931236	0.02510326
16	New Orleans	-0.4452805	0.10142038	-0.3553667	0.10361537
17	Baltimore	1.08365708	0.02733962	0.96353621	0.0320028
18	Detroit	-0.4192524	0.06521326	-0.5715712	0.10156746
19	Minn-St. Paul	-1.1043746	0.10407987	-1.0646013	0.10434028
20	Kansas City	-1.0935494	0.0270098	-1.0612343	0.02699778
21	St. Louis	1.37489749	0.03300185	1.5794532	0.02502478
22	Omaha	-1.2096782	0.04512354	-1.2698802	0.04114356
23	Albuquerque	-1.0242879	0.03855618	-1.0881455	0.03491484
24	Albany	0.89340335	0.08475065	0.72928035	0.08165794
25	Buffalo	-2.3518932	0.06840754	-2.2982448	0.06758477
26	Cincinnati	-0.2524519	0.0271122	-0.1521458	0.02901927
27	Cleveland	0.84524117	0.06174297	1.13177393	0.04607718
28	Columbus	-0.073457	0.04088923	-0.2473125	0.03583499
29	Philadelphia	0.82858005	0.14568295	0.96140035	0.15990662
30	Pittsburgh	1.41603112	0.04407677	1.24860787	0.04182284
31	Providence	2.23346419	0.04675151	2.28042973	0.05873754
32	Memphis	-1.0158551	0.04083487	-1.0371497	0.04085574
33	Nashville	-0.0206856	0.032249	-0.0415336	0.0322778
34	Dallas	-1.0842556	0.08633817	-0.9293195	0.08336943
35	Houston	-0.6917393	0.12547127	-0.7464621	0.14403282
36	Salt Lake City	0.14690526	0.05066424	0.09859461	0.05306595
37	Norfolk	1.20651981	0.03854345	1.063297	0.03640372

On the table above the values highlighted in yellow correspond to outliers, in the case the studentized residuals of the selected regressions are greater than 2, or to influential observations, when the values of the hat column of the regressions are greater than 0.15 (value calculated with the equation that follows after the paragraph).

$$\text{Where } h_{ii} > 2 \frac{(k+1)}{n} = 2 \frac{(2+1)}{40} = 0.15 \text{ is influential}$$

The purpose of the following normal plot is to show the normalization of the values of the response variable.

Normal plot for $\ln SO_2$

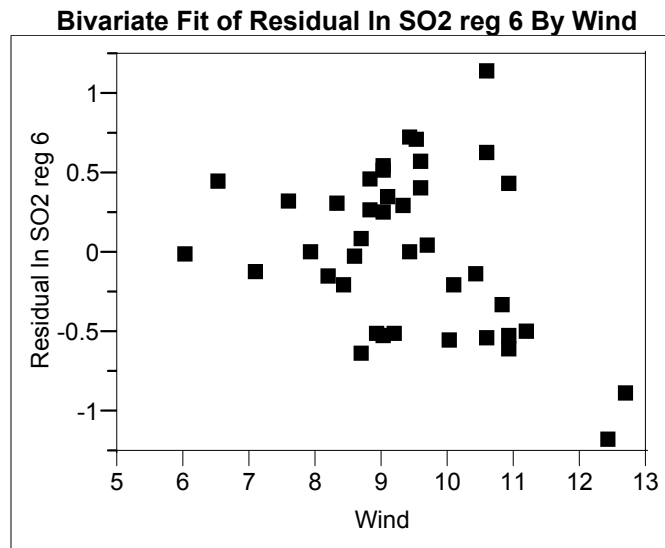


Checking for Omitted Variables

When checking for omitted variables plots of the residuals against any omitted predictor variables are to be made. The omitted variables are Pop, Wind, Rain and Rain days. The variable Pop will not be included because of collinearity with Man. Hence the next plots are searched for systematic pattern.

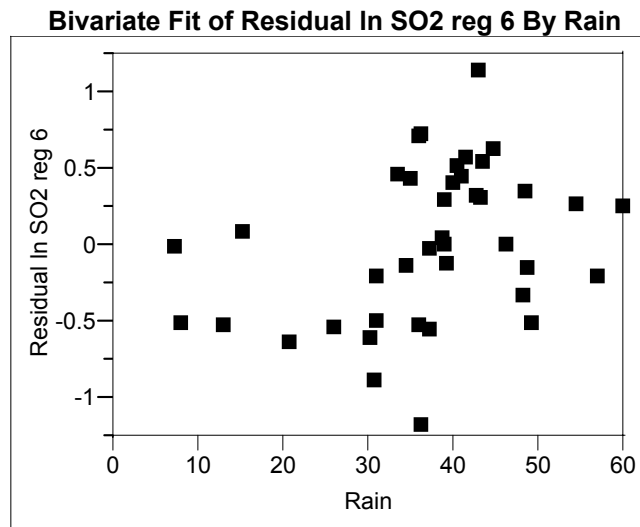
Omitted variable Wind

This plot appears to show a systematic pattern and will be included in the model.



Omitted variable Rain

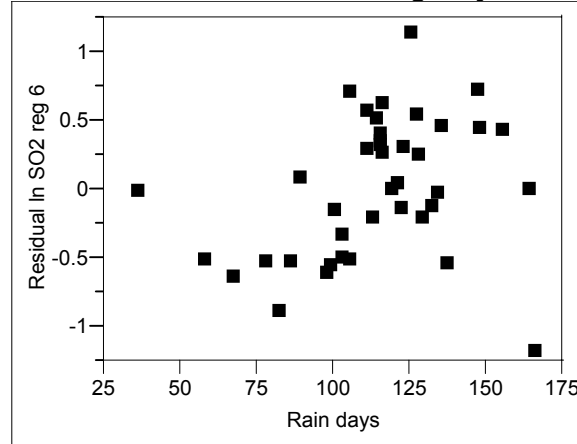
This plot does not appear to show any systematic pattern, it should not be included in the model.



Omitted variable Rain days

This plot does not appear to show any systematic pattern, it should not be included in the model.

Bivariate Fit of Residual In SO2 reg 6 By Rain days



Final Model

After the adding of the predictor variable Wind, the next model has been chosen to be the predictor model for the concentration of SO₂ in the forty cities of the study.

Final Model Regression Equation:

$$\ln SO_2 = 7.48 - 0.0558Temp + .0006Man - 0.141Wind$$

The r-square value for this fit is 0.56, which is much better than the one omitting the predictor variable Wind.

Summary of Fit

RSquare	0.555576
RSquare Adj	0.51854
Root Mean Square Error	0.491655
Mean of Response	3.162514
Observations (or Sum Wgts)	40

From the ANOVA table F is statistically significant and the p-value is less than 0.05.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	10.878505	3.62617	15.0012
Error	36	8.702091	0.24172	Prob > F
C. Total	39	19.580596		<.0001

As shown in the parameters estimates output table, there is no collinearity between predictor variables, the CI are rejecting the null hypothesis, they do not contain zero, and the p-value for all of the variables is smaller than 0.05. Plus the VIF values are smaller than 10.

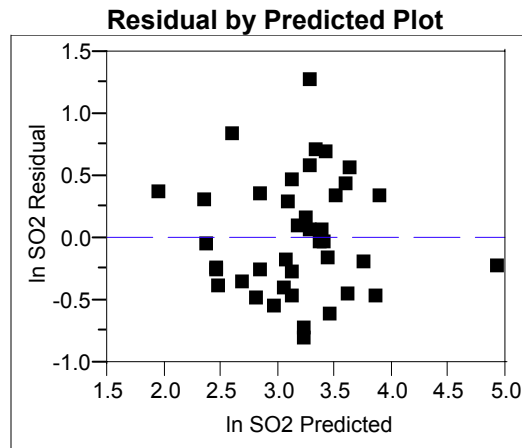
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%	VIF
Intercept	7.4812227	0.981684	7.62	<.0001	5.4902752	9.4721701	.
Temp	-0.058052	0.011692	-4.97	<.0001	-0.081765	-0.03434	1.1230655
Man	0.000553	0.000143	3.86	0.0005	0.0002625	0.0008435	1.0768607
Wind	-0.140814	0.060464	-2.33	0.0256	-0.26344	-0.018187	1.1486869

Effect Tests

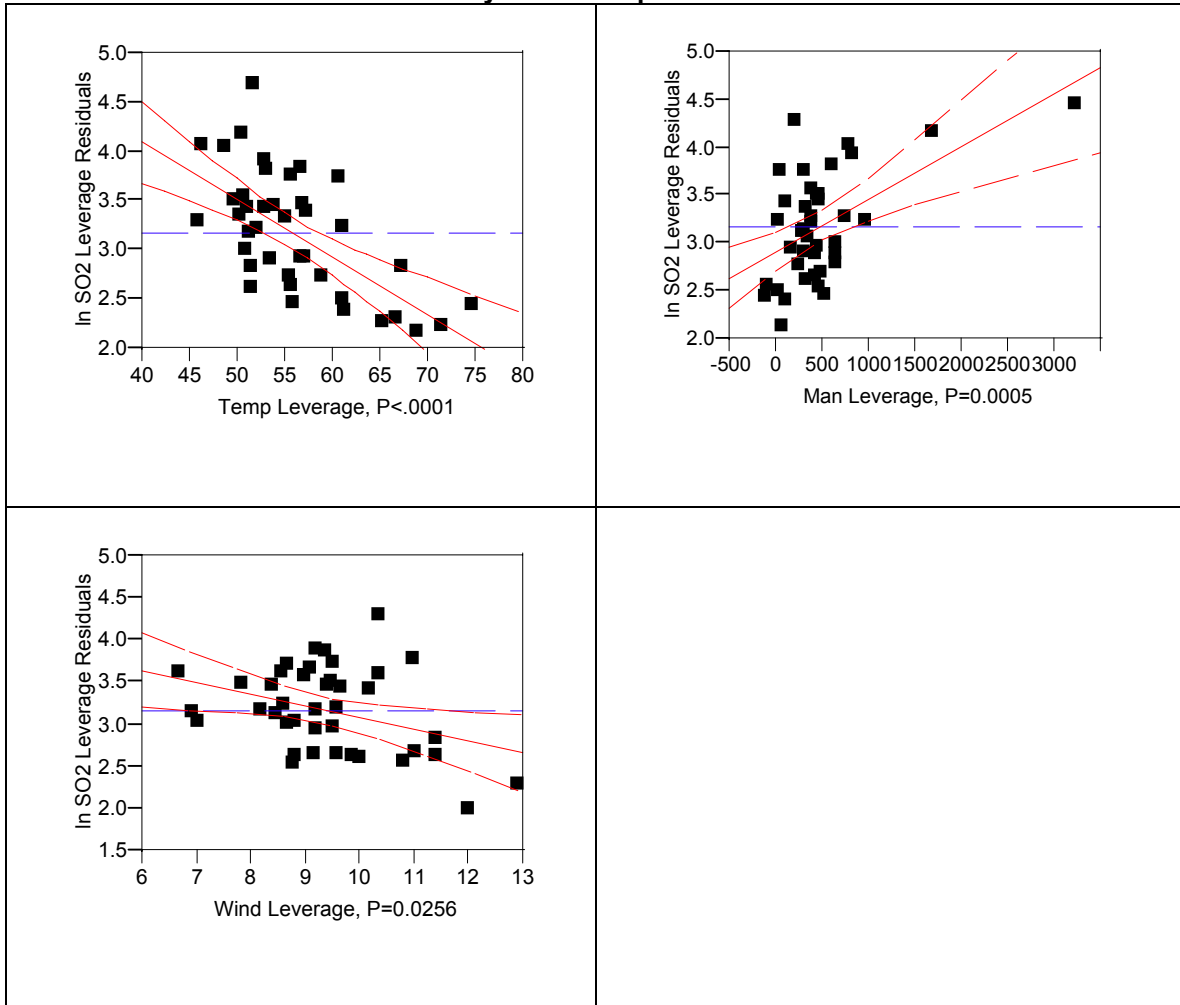
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	5.9589418	24.6518	<.0001
Man	1	1	3.6034500	14.9072	0.0005
Wind	1	1	1.3110539	5.4237	0.0256

The residual against the predicted values plot does not show any systematic pattern, hence the variance is stabilized.



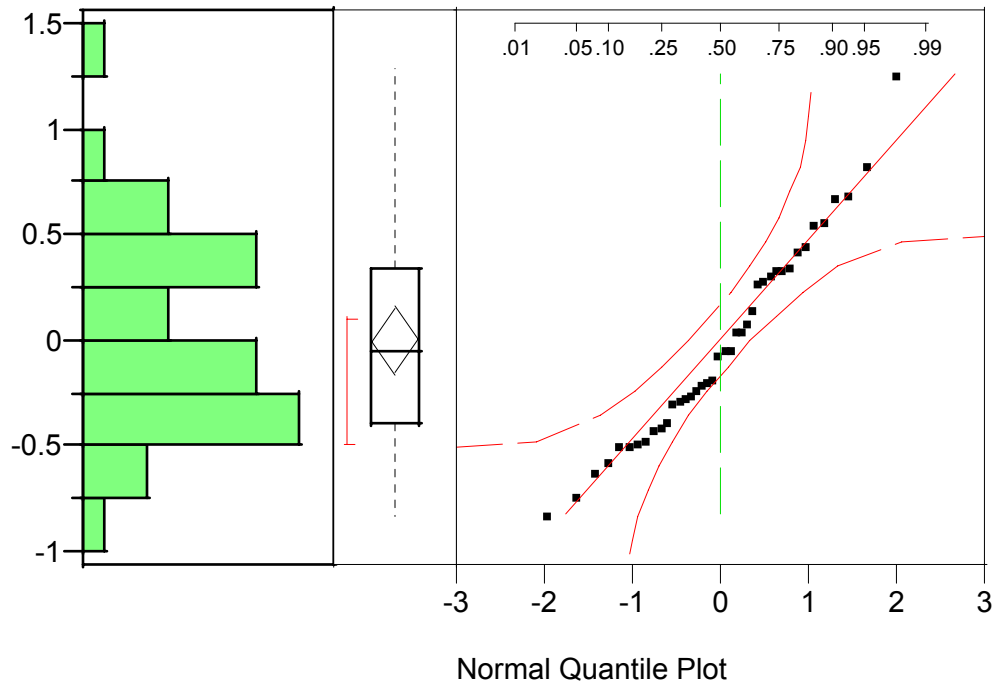
The plots of the residuals against individual predictor variables, Temp, Man and Wind, show that there is no clear systematic pattern. By this it appears to be no violation from the assumption of linearity.

Residual by Individual predictor values



The residual plot appears not to be violating the assumption of normality.

Normal Plot of the residual



There is only one outlier in the final regression (Providence) and it will not be taken out for the reasons explained before. On the other hand the influential observations are four (Phoenix, Miami, Chicago and Wichita), two more than regression 6.

Outliers and Influential Data

No	City	FINAL REGRESSION	
		Studentized Residual	h_{ii}
1	Phoenix	-0.8547107	0.22280942
2	Little Rock	-0.5658686	0.05307608
3	San Francisco	-1.5091232	0.03152618
4	Denver	-1.2842277	0.03940285
5	Hartford	0.91001189	0.05911908
6	Wilmington	0.95652128	0.04236878
7	Washington	0.58120542	0.02583591
8	Jacksonville	0.6288655	0.10458612
9	Miami	0.82543179	0.22542241
10	Atlanta	0.7161538	0.04019834
11	Chicago	-0.8287344	0.68662904
12	Indianapolis	0.11048155	0.0343145
13	Des Moines	-0.6105954	0.10602568
14	Wichita	-0.913343	0.21914147
15	Louisville	0.32344899	0.04364563
16	New Orleans	-0.534461	0.10208038
17	Baltimore	1.17144623	0.02744965
18	Detroit	-0.4107733	0.06539953
19	Minn-St. Paul	-1.0386386	0.10693849
20	Kansas City	-0.9940587	0.03192539
21	St. Louis	1.44457134	0.03302045
22	Omaha	-0.8832381	0.07479722
23	Albuquerque	-1.1583664	0.03952662
24	Albany	0.69623681	0.09556864
25	Buffalo	-1.8269142	0.16819963
26	Cincinnati	-1.0081444	0.1137746
27	Cleveland	1.17640614	0.07472075
28	Columbus	-0.3545579	0.05420067
29	Philadelphia	0.74643851	0.14840534
30	Pittsburgh	1.43030617	0.04490636
31	Providence	2.65972621	0.06029125
32	Memphis	-1.025107	0.04127696
33	Nashville	-0.3811318	0.05470958
34	Dallas	-0.5732804	0.14554527
35	Houston	-0.1278904	0.18516958
36	Salt Lake City	-0.0818986	0.06048847
37	Norfolk	1.76568664	0.07524942

Where $h_{ii} > 2 \frac{(k+1)}{n} = 2 \frac{(3+1)}{40} = 0.20$ is influential

Providence is a city that was an outlier for SO_2 at the beginning of the analysis, since it has high concentration of SO_2 . Even though Providence has a small population, 179,000 people, it has SO_2 levels of big cities like Chicago, $110 \mu\text{g} / \text{m}^3$.