Statistical Data Mining

Air Pollution and Mortality

Timothy Adams

4/29/2014

Introduction

For this assignment we are looking at a data set, Mortality Data, collected by researchers at General Motors to investigate whether air pollution contributes to mortality. The data set collects a handful of different measurements of demographics, climate, and air quality variables for 60 cities. Its 17 variables can be seen in appendix 1a. Our goal is to investigate this data to determine if there is evidence of a link between pollution and mortality. We will use both R and SAS to perform this investigation, leading to the secondary goal for this paper: to determine the relative strengths and weaknesses of each of these platforms for performing this type of analysis.

Data Exploration

The first step in any analysis is data exploration, so we load our data into our statistical packages to take a look.

Individual Measures: Summary Statistics and Graphical Displays

First we produce a set of summary statistics to get a feel for our data. The results can be seen in appendix 1b and 1c. Looking at the results of these summaries, we can see that, with the exception of city, all of our variables are numeric. City is categorical, but being unique for each row, it is also simply the key for our dataset, so we will not be including it in the actual analysis. Another obvious observation the summary statistics is that both the population and median income columns are missing one value. Revisiting the actual data reveals that both of these 'NA' values are for the same row: Fort Worth, TX; a fact that could be useful later. A quick glance at the data makes it appear that most of the means and medians are closer together, with the means tending to come out a little higher, indicating some right skews to our data, but to better understand the distributions we turn to some visualizations.

R vs SAS – Summaries

R summary is more concise, quartiles and mean. SAS univariate procedure gives a much more detailed breakdown of the variables: variance, standard deviation, quantiles, range, etc. but the detail interferes with the data set at a glance approach. SAS also reports the mode, where R lacks a native mode function.

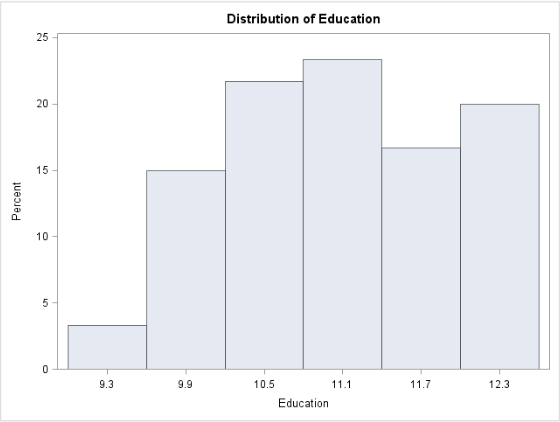
To begin our investigation into variable distributions, we start by producing histograms and boxplots for our data. In the interest of space, the histograms can be seen in the diagonal of the scatterplot matrix in appendix 1d. For the most part our distributions appear normal: income, JanTemp, JulyTemp, PopDensity, Rain, RelHum, and WC. We have a little concern that Education might show a bimodal distribution. We see a heavy right skew to a handful of variables: HCPot, NonWhite, NOx, NOxPot, Pop, PopDensity, and SO2Pot. Finally we see a bit of a left skew in both PopHouse and Mortality. Since mortality is the response variable we are interested in, we take a closer look at it on its own by running a box plot, see appendix 1e. The boxplot of mortality looks perfect: no outliers, centered, and very even ranges. While running boxplots, we also look at the other variables and note that almost all of them show some outliers. Of particular note are HCPot, NOx, and NOxPot, appendix 1f. Each of these variables has outliers so extreme that they reduce the range of the boxplot. After investigating some transformations to deal with the extreme skew to these three variables, we found that using a log function helps make the distribution closer to normal and reduces the outliers. Appendix 1g and 1h show the histograms and boxplots for the transformed variables.

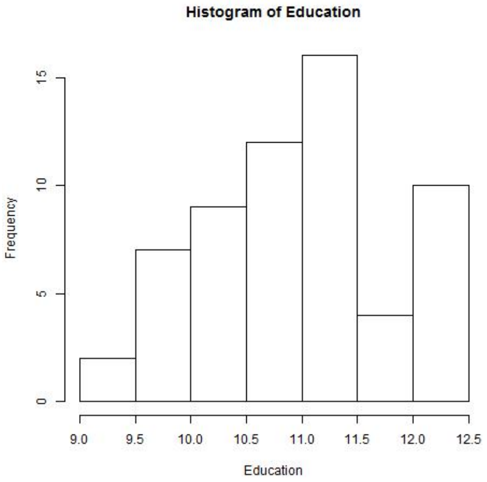
Pairwise Investigations: Correlations and Scatterplots

Once our investigation of individual variables has concluded, we start looking at the relationships between variables, especially how the variables in our dataset relate to mortality, the response variable in which we are interested. A quick and easy way to look at these relationships is to visualize the correlation matrix of the variables, see appendix 1i for a more visual correlation matrix and appendices 1d and 1j for more detail. Our visual correlation matrix (1i) lets us easily see that there are positive relationships between Mortality and pophouse, SO2Pot, PopDensity, Rain, Pop, NOxPot, NOx, JulyTemp, and NonWhite, and negative relationships between RelHum, Education, Income, WC, HCPot, and JanTemp. Furthermore, it indicates that the strongest relationships are between NonWhite, SO2Pot, and PopDensity on the positive side, and Education,

R vs SAS – Single Variable Visualizations

R has the hist() function to easily produce a histogram. SAS, includes a histogram option on proc univariate to produce histograms for each variable along with the summary statistics. From the perspective of having a native approach to produce all summary statistics and visualizations for a dataset in one go, SAS has the edges, but, again, this level of detail can make it difficult to get an overall sense of the data. An interesting note on the histograms: SAS and R have different heuristics to determine where to cut the data as well as different default displays. This can be seen in the education histograms. SAS chooses to use a range of .6 for education and displays the results in percentages (the histogram on the left). R chooses to use a rage of .5 for computing frequencies and displays the results in counts. While the data are the same, and the distributions similar, a quick glance at these might lead one to draw slightly different conclusions.





RelHum, and Income on the negative side. The weakest relationships seem to be between NOxPot, NOx, and HCPot. Our more detailed matrices mostly support these conclusions, except that when looking at the numbers, the relationship between HCPot and Mortality is stronger than the relationship between JanTemp and Mortality. The difference is subtle and the nuance is lost looking at the strictly visual, higher level visualizations of the correlations.

In addition to looking for correlations with the response variable, we also want to look for variables that are highly correlated with one another. From our correlation matrices, we can see a strong relationship between NOx, NOxPot, and HCPot. We can also see a strong relationship between JulyTemp and RelHum. These are things we will need to keep in mind when we begin modeling the data.

R vs SAS – Pairwise Analysis

R has some very powerful packages for finding and visualizing correlations: the Yale toolkit and ellipse package come to mind. However, the SAS proc corr procedure is quite powerful and does a lot. It was easier to transform data with R than SAS. Here R felt more intuitive and easier. This does not even take into account RATTLE, which is a visual R extension that makes it easy to visualize, summarize, and transform data.

However, before we are ready to begin modeling our data, we still want to look at the relationships among our variables. While we have investigated these relationships to see which variables are related, we also want to look at the nature of those relationships. We produce some scatterplots to represent these relationships visually, appendix 1d and 1k. Two things stand out: the relationship between SO2Pot and mortality seems distinctly non-linear (it starts out steep, but levels off) and the scatterplots for NOx, NOxPot, and HCPot show almost no relationship. While this is the first time we have remarked anything about SO2Pot, we have seen the other three variables several times now. When looking at their original distributions, we concluded that they were distinctly right skewed and that a logarithmic transformation helped alleviate that. This is convenient, sine the relationship between SO2Pot and Mortality seems logarithmic. Investigating how transforming these variables affects our data we look at Mortality by the variable, Mortality by the log transformed variable, and then log transformed mortality by the log transformed variable. The results can be seen in appendix 1l. In each case, taking the log of the predictor variable strengthens the relationships and increases the linearity of the relationship with Mortality. However, taking the log of the response variable has little extra effect. With these transformations in mind we recreate the scatterplot matrix as seen in appendix 1m. We can see from the results that the strength of the relationships between Mortality and both NOx and NOxPot has increased significantly with the log transformation, while the relationship between Mortality and HCPot has decreased slightly, but the direction has changed from positive to negative. It appears as if the relationship between Mortality and S02Pot is basically the same as the relationship between Mortality and the log of S02Pot. It should also be noted that we attempted a log transformation on the pop variable to see if it would help strengthen its correlation with Mortality, but it had little effect.

Data Modeling

After the data exploration above we are ready to begin developing some models to represent our data. We will want to take advantage of the computing power of our software platforms to recommend a model for us using automated stepwise regression. We will do this with and without the transformations we have considered. We will also try applying some common sense, along with the lessons we learned through data exploration to build up a parsimonious model to see how that performs.

Automated Stepwise Regression: Without Transformations

The proverbial kitchen-sink model for this data set is:

$$\begin{aligned} Mortality = {} & Intercept + \alpha_1(JanTemp) + \alpha_2(JulyTemp) + \alpha_3(RelHum) + \alpha_4(Rain) + \alpha_5(Education) \\ & + \alpha_6(PopDensity) + \alpha_7(NonWhite) + \alpha_8(WC) + \alpha_9(pop) + \alpha_{10}(pophouse) + \alpha_{11}(income) \\ & + \alpha_{12}(HCPot) + \alpha_{13}(NOxPot) + \alpha_{14}(S02Pot) + \alpha_{15}(NOx) \end{aligned}$$

Running a stepwise regression this model is pared down to the following:

$$\begin{aligned} Mortality = {} & 1131 - 1.213(JanTemp) - 2.547(JulyTemp) + 1.191(Rain) + .009(PopDensity) \\ & + 4.965(NonWhite) - 2.286(WC) - .878(HCPot) + 1.71(NOxPot) \end{aligned}$$

See appendix 1n for the results. The multiple R-squared for this model is .7412, implying that 74% of the variability in Mortality is explained by this model. However, the adjusted R-squared is only .6998, implying that we have some non-essential variables in the model. This is supported by the p-values of the variables.

Automated Stepwise Regression: With Transformations

Remembering that we transformed some of our variables during data exploration in order to increase their relationship to Mortality, we try stepwise regression again, this time with transformations. The starting model is:

$$\begin{aligned} Mortality = \; & Intercept + \alpha_1(JanTemp) + \alpha_2(JulyTemp) + \alpha_3(RelHum) + \alpha_4(Rain) + \alpha_5(Education) \\ & + \alpha_6(PopDensity) + \alpha_7(NonWhite) + \alpha_8(WC) + \alpha_9(pop) + \alpha_{10}(pophouse) + \alpha_{11}(income) \\ & + \alpha_{12}(\log(HCPot)) + \alpha_{13}(\log(NOxPot)) + \alpha_{14}(\log(S02Pot)) + \alpha_{15}(\log(NOx)) \end{aligned}$$

The result of the stepwise regression is:

$$\begin{aligned} Mortality = \; & 1031.9491 - 2.0235(JanTemp) + 1.8117(Rain) - 10.7463(Education) + 4.0401(NonWhite) \\ & - 1.4514(WC) + 19.2481 * log(NOxPot) \end{aligned}$$

This new model has a multiple R-squared of .7383, so it explains a little less of the variability of Mortality than the previous model we created without the transformations. However, looking at the adjusted R-squared, our new model comes in at .7081, implying that our new model is a little better than the previous one. Looking at the p-values of the variables in this model, it does appear we have chosen better predictors.

Manual Model Building

Finally, we take everything we have learned so far to try to construct a model manually. We are interested in pollution and its effect on Mortality, so we want to include these variables in our model: HCPot, NOxPot, S02Pot, and NOx. Something we did not touch on earlier: in our correlation matrix, NOx and NOxPot are perfectly correlated, so we do not want to include both in our model. Instead we drop NOx, so that all of our pollution predictors are stated in terms of pollution potential. This consistency will help the interpretability of our model.

After deciding on HCPot, NOxPot, and S02Pot, we move to the predictor with the highest correlation with Mortality: NonWhite. Running a regression on this model we have an R-squared of 61.7% and adjusted R-squared of .5892. Not bad, but we know we can do better. We add the next most highly correlated variable, Education, and try again. This increases our R-squared and adjusted R-squared to 66.9% and .6383, respectively: a good improvement, and our model is still understandable, since it is looking at the effects of race, education, and pollution on mortality. The next most highly correlated variable is Rain, which makes sense in our model, since we suppose rainfall and air pollution could be related. However, adding it to the model only increases our adjusted R-squared to .6451, so this is not much better of a model than the previous one, and the p-value on the Rain variable is .1595, so it is not even considered significant in terms of the model. Reviewing our correlation matrices, we notice that Rain is correlated with HCPot and NOxPot, so we are likely capturing a lot of this information in our model already. So, we pass on this variable. Finally, we look at popHouse as a variable instead. This actually hurts our adjusted R-squared, .6319, so we disregard it.

R vs SAS – Model Development

R and SAS are quite similar for model development. R has the lm and glm functions for linear models and generalized linear models, respectively, where SAS has proc reg and proc glm for the same. One can load the MASS package in R and perform stepwise regression using stepAIC. SAS has a native proc stepwise for performing such a function. They produce the same results given the same input, and use very similar syntax. However, SAS natively produces some additional visualizations to diagnose model fit, which is a nice feature.

The results of this exploration of data yield the simple model:

$$Mortality = 1095.33553 - 18.98703(Education) + 3.53555(NonWhite) - .8688(HCPot) \\ + 1.56252(NOxPot) + .16680(S02Pot)$$

But, we had some transformations on the pollution predictors that proved helpful before, so we investigate if those help the model, by applying a log to the HCPot, NOxPot, and S02Pot variables. The regression on the transformed variables yields and R-square of 71.27% and adjusted

R-square of .6856, so this is the model we want to keep out of this section. The model is shown below and more detail can be seen in appendix 1n.

$$
\begin{aligned}
Mortality = \; & 1168.36270 - 25.12374(Education) + 2.94146(NonWhite) - 28.28948 * log(HCPot) \\
& + 9.21007 * (NOxPot) + 21.83714 * (S02Pot)
\end{aligned}
$$

Investigating Model Assumptions

Once we have chosen some models that we think explain our data, we have to look at how they fit and if they satisfy some of the assumptions of linear regression. We do this by looking at the residuals. Refer to appendix 1o and 1p for details. Looking at histograms of the residuals from our automated untransformed model, automated transformed model, and manual model, we see that the errors are fairly normally distributed. Perhaps the manual model shows some skew. The autocorrelation graphs show no real evidence of autocorrelation. However, the scatterplots of the residuals show some differences between the models. The untransformed automated model shows some extreme errors, and the variance in the errors does not seem consistent. However, the errors for the transformed automated model show a more constant variance. The errors in our manually chosen model seem to be consistently lower for lower values and higher for higher values. That pattern also seems to exist in the transformed automated model, but to a lesser extent. However, these seem to indicate that we have a little more work to do on our models.

R vs SAS – Residuals

R and SAS both provide the same ability to run tests for normality and to create visualizations to look at the distribution of model residuals. As mentioned in the model development section, SAS bundles a handful of these visualizations along with proc reg. In R these visualizations must be created manually. However, R conveniently keeps the residuals along with the developed model, so plotting them is a simple enough exercise.

Unusual and Influential Observations

Creating some influence charts, we can see how outliers may affect our different models. For model 1, our automated non-transformed model, observations 20, York, PA, and 8, Miami-Hialeah, FL, could be influential points. For model 2, our automated transformed model, observation 20, York, PA, appears to be a potentially influential outlier. For model 3, our manually produced model, observation 60, New Orleans, LA, appears to be an outlier . Plots identifying these outliers can be seen in appendices 1q, 1r, and 1s.

Conclusion

I would recommend model 2, our automated, transformed model as the best.

$$Mortality = 1031.9491 - 2.0235(JanTemp) + 1.8117(Rain) - 10.7463(Education) + 4.0401(NonWhite) - 1.4514(WC) + 19.2481 * log(NOxPot)$$

It has the best R-squared and adjusted R-squared of all of our models. Apart from that, the residuals seem to have the least variance and the best distribution. I would conclude that pollution affects mortality. The fact the NOxPot is included in the final model at a level that is significant (not to mention that it showed up in our other models as well) indicates that it is related to mortality. Another important factor seems to be race, although the inclusion of WC (white collar), and Education, leads us to believe there is some class or socio-economic factor at work which is being expressed in the model through the NonWhite predictor.

R vs SAS – Conclusion

After completing this exercise using both R and SAS, we can draw a few conclusions. It seems like SAS has a more powerful base package. It does more, but also requires more training to know where to look for all of the functions. It neatly ties up sets of functionalities in its procedures. R, on the other hand, does less out of the box, but has a very rich set of packages that are easy to find and will do just about anything that one wants. Also, it bears mentioning again, that for simple data exploration and transformation, the RATTLE package is very fast and easy to use. We are not aware of a SAS equivalent. So, with less training, it seems that one can get farther with R. Not to mention the fact that R is free. However, if one has the resources to spend on a SAS license and some SAS training, it really does appear to be a great software package.

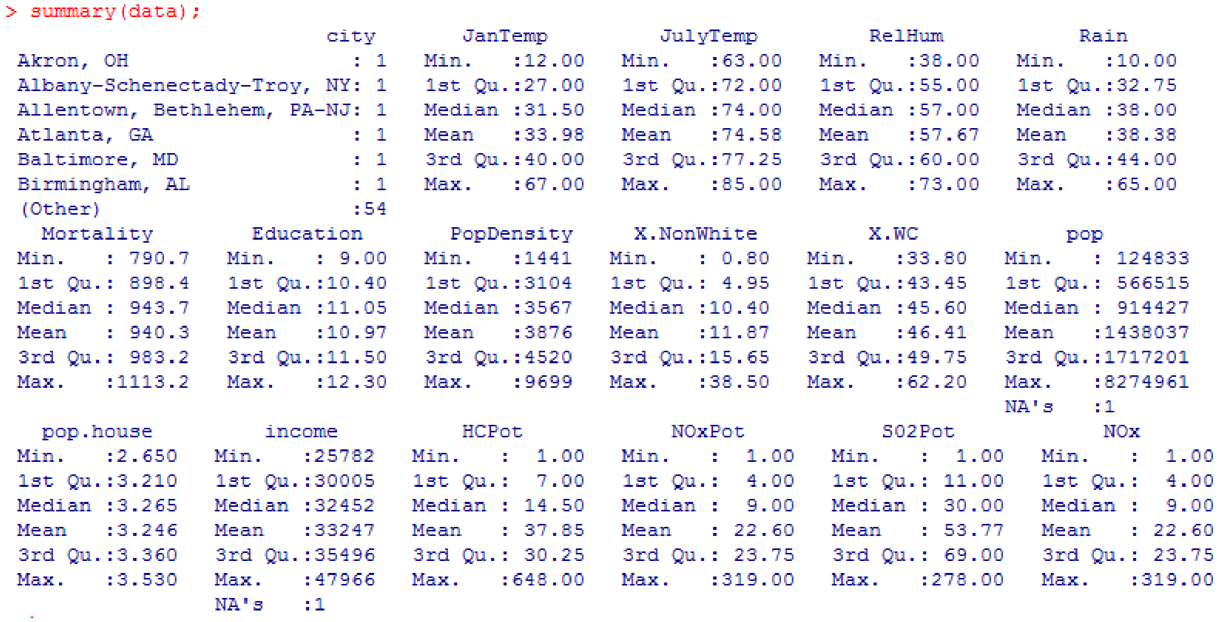
Caveats for these conclusions: we spent all semester working in R, so at this point are much more comfortable with that software package. Also, the citrix based SAS software instance, while free, is far from ideal. It does not store session information and users have limited access to write to all the libraries SAS uses to run in the background. Having a real instance of SAS running on a PC would have made interacting with SAS much better.

Appendix

1a. Variables

1	city:	City name
2	JanTemp:	Mean January temperature (degrees Fahrenheit)
3	JulyTemp:	Mean July temperature (degrees Fahrenheit)
4	RelHum:	Relative Humidity
5	Rain:	Annual rainfall (inches)
6	Mortality:	Age adjusted mortality
7	Education:	Median education
8	PopDensity:	Population density
9	%NonWhite:	Percentage of non-whites
10	%WC:	Percentage of white collar workers
11	pop:	Population
12	pop/house:	Population per household
13	income:	Median income
14	HCPot:	HC pollution potential
15	NOxPot:	Nitrous Oxide pollution potential
16	SO2Pot:	Sulfur Dioxide pollution potential
17	NOx:	Nitrous Oxide

1b. Summary in R

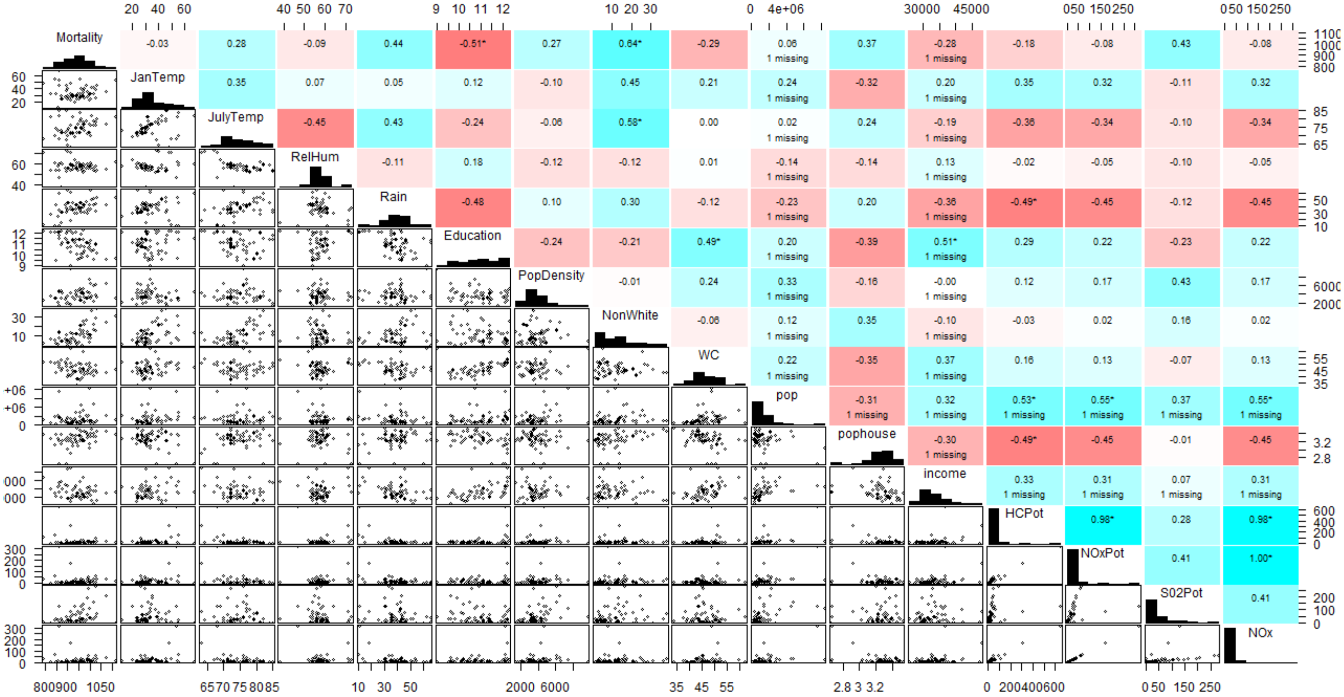


```
> summary(data);
                           city      JanTemp          JulyTemp          RelHum            Rain      
 Akron, OH                  : 1   Min.   :12.00   Min.   :63.00   Min.   :38.00   Min.   :10.00  
 Albany-Schenectady-Troy, NY: 1   1st Qu.:27.00   1st Qu.:72.00   1st Qu.:55.00   1st Qu.:32.75  
 Allentown, Bethlehem, PA-NJ: 1   Median :31.50   Median :74.00   Median :57.00   Median :38.00  
 Atlanta, GA                : 1   Mean   :33.98   Mean   :74.58   Mean   :57.67   Mean   :38.38  
 Baltimore, MD              : 1   3rd Qu.:40.00   3rd Qu.:77.25   3rd Qu.:60.00   3rd Qu.:44.00  
 Birmingham, AL             : 1   Max.   :67.00   Max.   :85.00   Max.   :73.00   Max.   :65.00  
 (Other)                    :54                                                                  
   Mortality        Education        PopDensity      X.NonWhite          X.WC             pop         
 Min.   : 790.7   Min.   : 9.00   Min.   :1441   Min.   : 0.80   Min.   :33.80   Min.   : 124833  
 1st Qu.: 898.4   1st Qu.:10.40   1st Qu.:3104   1st Qu.: 4.95   1st Qu.:43.45   1st Qu.: 566515  
 Median : 943.7   Median :11.05   Median :3567   Median :10.40   Median :45.60   Median : 914427  
 Mean   : 940.3   Mean   :10.97   Mean   :3876   Mean   :11.87   Mean   :46.41   Mean   :1438037  
 3rd Qu.: 983.2   3rd Qu.:11.50   3rd Qu.:4520   3rd Qu.:15.65   3rd Qu.:49.75   3rd Qu.:1717201  
 Max.   :1113.2   Max.   :12.30   Max.   :9699   Max.   :38.50   Max.   :62.20   Max.   :8274961  
                                                                                 NA's   :1        
   pop.house         income            HCPot             NOxPot            SO2Pot             NOx        
 Min.   :2.650   Min.   :25782   Min.   :  1.00   Min.   :  1.00   Min.   :  1.00   Min.   :  1.00  
 1st Qu.:3.210   1st Qu.:30005   1st Qu.:  7.00   1st Qu.:  4.00   1st Qu.: 11.00   1st Qu.:  4.00  
 Median :3.265   Median :32452   Median : 14.50   Median :  9.00   Median : 30.00   Median :  9.00  
 Mean   :3.246   Mean   :33247   Mean   : 37.85   Mean   : 22.60   Mean   : 53.77   Mean   : 22.60  
 3rd Qu.:3.360   3rd Qu.:35496   3rd Qu.: 30.25   3rd Qu.: 23.75   3rd Qu.: 69.00   3rd Qu.: 23.75  
 Max.   :3.530   Max.   :47966   Max.   :648.00   Max.   :319.00   Max.   :278.00   Max.   :319.00  
                 NA's   :1                                                                          
```

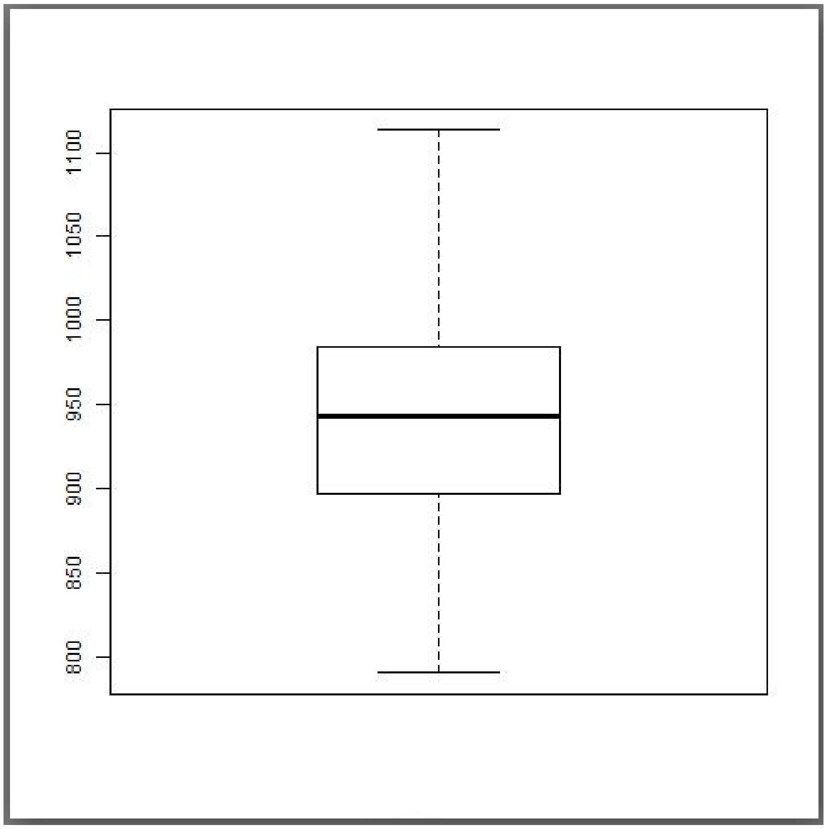
1c. Summary – SAS

```
The UNIVARIATE Procedure
                                        Variable:  JanTemp

                                              Moments

                  N                            60    Sum Weights                 60
                  Mean                 33.9833333    Sum Observations          2039
                  Std Deviation        10.1688985    Variance            103.406497
                  Skewness             0.96071148    Kurtosis            1.08782139
                  Uncorrected SS            75393    Corrected SS        6100.98333
                  Coeff Variation      29.9231933    Std Error Mean      1.31279915


                                    Basic Statistical Measures

                          Location                    Variability

                      Mean     33.98333     Std Deviation           10.16890
                      Median   31.50000     Variance               103.40650
                      Mode     24.00000     Range                   55.00000
                                            Interquartile Range     13.00000

              Note: The mode displayed is the smallest of 2 modes with a count of 5.


                                    Tests for Location: Mu0=0

                         Test           -Statistic-    -----p Value------

                         Student's t    t  25.88616    Pr > |t|    <.0001
                         Sign           M        30    Pr >= |M|   <.0001
                         Signed Rank    S       915    Pr >= |S|   <.0001


                                     Quantiles (Definition 5)

                                      Quantile      Estimate

                                      100% Max          67.0
                                      99%               67.0
                                      95%               54.5
                                      90%               48.5
                                      75% Q3            40.0
                                      50% Median        31.5
                                      25% Q1            27.0
                                      10%               24.0
                                      5%                23.0
                                      1%                12.0
                                      0% Min            12.0
```

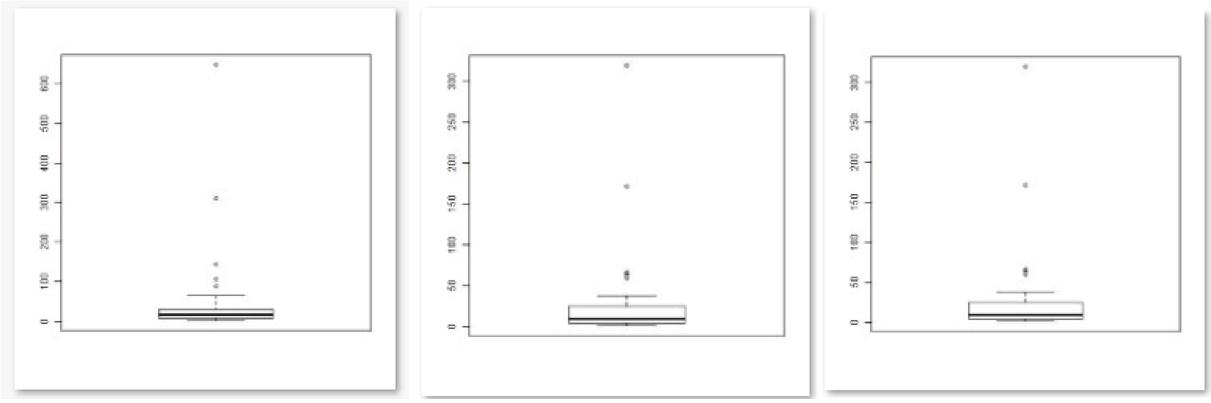
1d. Scatterplot Matrix - R



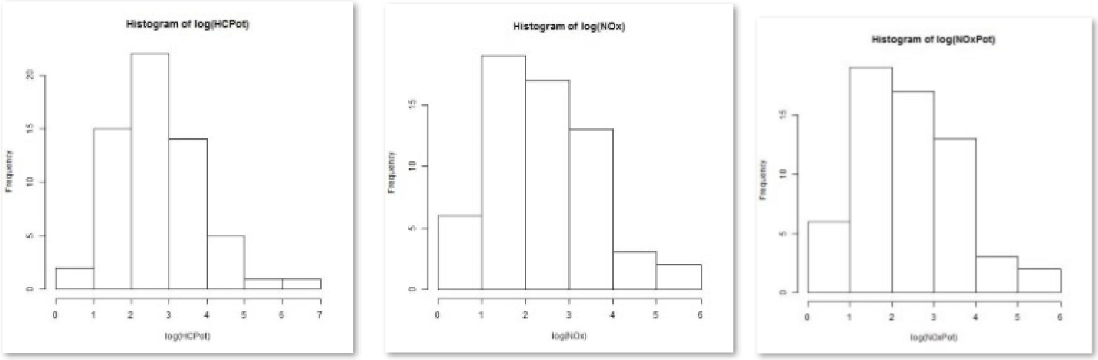
1e. Boxplot of Mortality – R



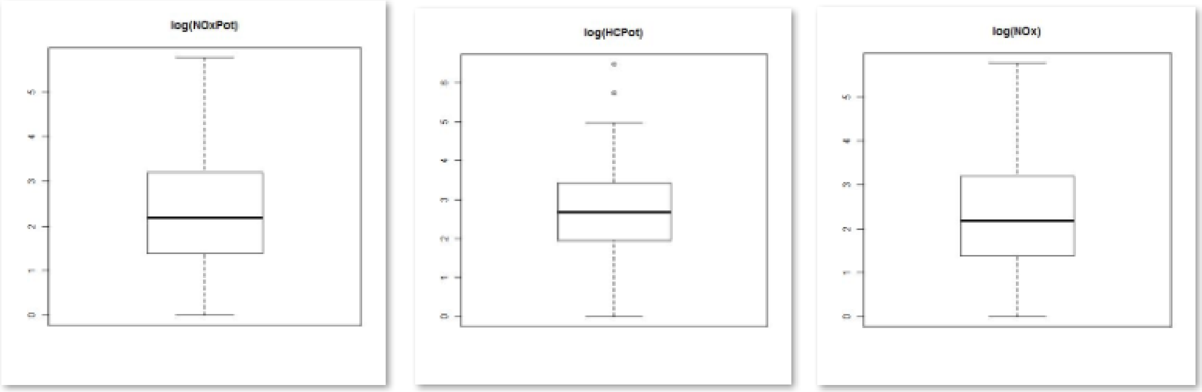
1f.Boxplot HCPot, NOx, NOxPot – R



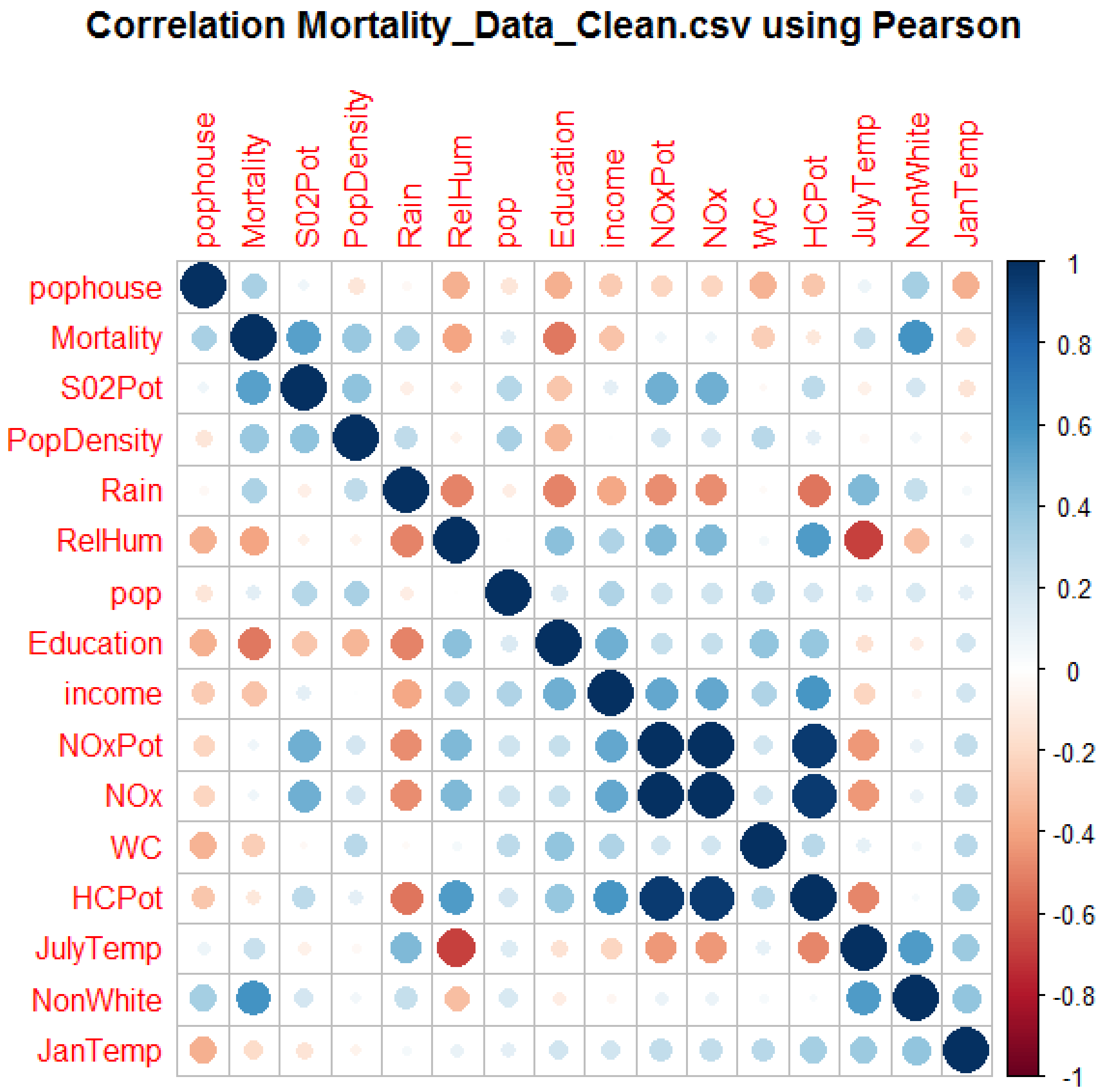
1g. Histograms log(HCPot), log(NOx), log(NOxPot) – R



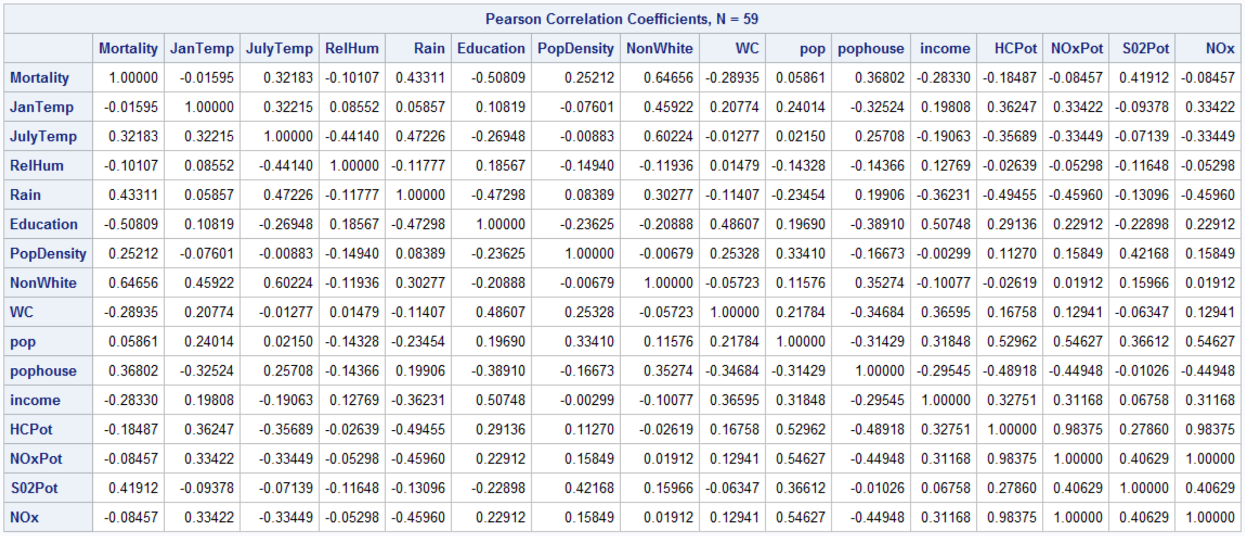
1h. Boxplots log(HCPot), log(NOx), log(NOxPot) – R



1i. Correlation Matrix - R



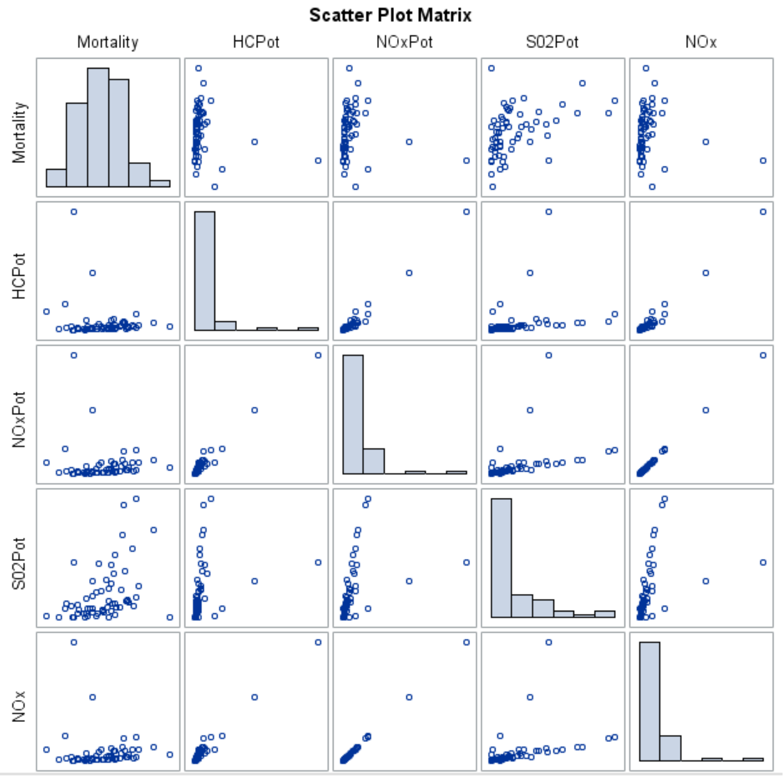
1j. Correlation Matrix – SAS



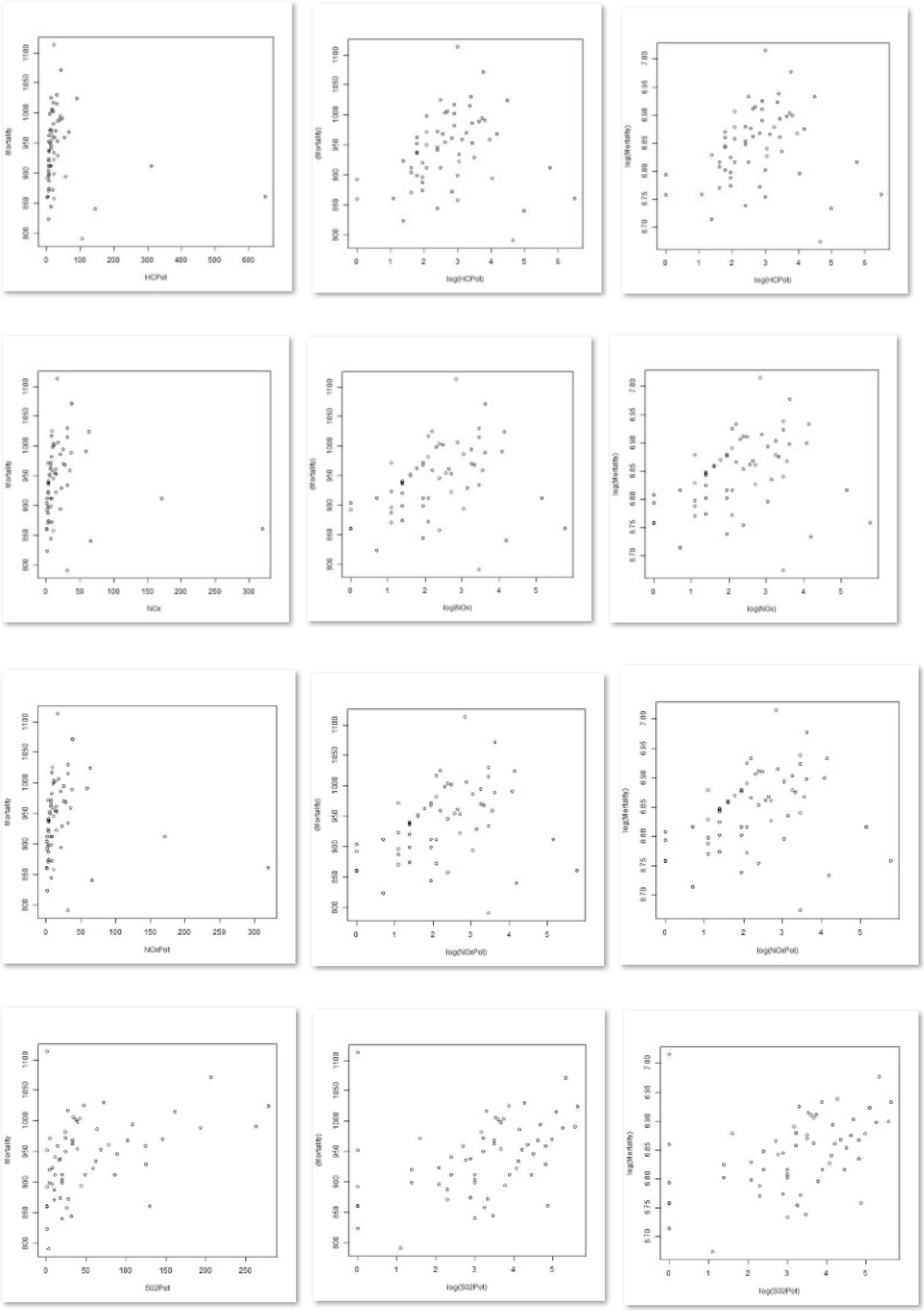
Pearson Correlation Coefficients, N = 59

	Mortality	JanTemp	JulyTemp	RelHum	Rain	Education	PopDensity	NonWhite	WC	pop	pophouse	income	HCPot	NOxPot	S02Pot	NOx
Mortality	1.00000	-0.01595	0.32183	-0.10107	0.43311	-0.50809	0.25212	0.64656	-0.28935	0.05861	0.36802	-0.28330	-0.18487	-0.08457	0.41912	-0.08457
JanTemp	-0.01595	1.00000	0.32215	0.08552	0.05857	0.10819	-0.07601	0.45922	0.20774	0.24014	-0.32524	0.19808	0.36247	0.33422	-0.09378	0.33422
JulyTemp	0.32183	0.32215	1.00000	-0.44140	0.47226	-0.26948	-0.00883	0.60224	-0.01277	0.02150	0.25708	-0.19063	-0.35689	-0.33449	-0.07139	-0.33449
RelHum	-0.10107	0.08552	-0.44140	1.00000	-0.11777	0.18567	-0.14940	-0.11936	0.01479	-0.14328	-0.14366	0.12769	-0.02639	-0.05298	-0.11648	-0.05298
Rain	0.43311	0.05857	0.47226	-0.11777	1.00000	-0.47298	0.08389	0.30277	-0.11407	-0.23454	0.19906	-0.36231	-0.49455	-0.45960	-0.13096	-0.45960
Education	-0.50809	0.10819	-0.26948	0.18567	-0.47298	1.00000	-0.23625	-0.20888	0.48607	0.19690	-0.38910	0.50748	0.29136	0.22912	-0.22898	0.22912
PopDensity	0.25212	-0.07601	-0.00883	-0.14940	0.08389	-0.23625	1.00000	-0.00679	0.25328	0.33410	-0.16673	-0.00299	0.11270	0.15849	0.42168	0.15849
NonWhite	0.64656	0.45922	0.60224	-0.11936	0.30277	-0.20888	-0.00679	1.00000	-0.05723	0.11576	0.35274	-0.10077	-0.02619	0.01912	0.15966	0.01912
WC	-0.28935	0.20774	-0.01277	0.01479	-0.11407	0.48607	0.25328	-0.05723	1.00000	0.21784	-0.34684	0.36595	0.16758	0.12941	-0.06347	0.12941
pop	0.05861	0.24014	0.02150	-0.14328	-0.23454	0.19690	0.33410	0.11576	0.21784	1.00000	-0.31429	0.31848	0.52962	0.54627	0.36612	0.54627
pophouse	0.36802	-0.32524	0.25708	-0.14366	0.19906	-0.38910	-0.16673	0.35274	-0.34684	-0.31429	1.00000	-0.29545	-0.48918	-0.44948	-0.01026	-0.44948
income	-0.28330	0.19808	-0.19063	0.12769	-0.36231	0.50748	-0.00299	-0.10077	0.36595	0.31848	-0.29545	1.00000	0.32751	0.31168	0.06758	0.31168
HCPot	-0.18487	0.36247	-0.35689	-0.02639	-0.49455	0.29136	0.11270	-0.02619	0.16758	0.52962	-0.48918	0.32751	1.00000	0.98375	0.27860	0.98375
NOxPot	-0.08457	0.33422	-0.33449	-0.05298	-0.45960	0.22912	0.15849	0.01912	0.12941	0.54627	-0.44948	0.31168	0.98375	1.00000	0.40629	1.00000
S02Pot	0.41912	-0.09378	-0.07139	-0.11648	-0.13096	-0.22898	0.42168	0.15966	-0.06347	0.36612	-0.01026	0.06758	0.27860	0.40629	1.00000	0.40629
NOx	-0.08457	0.33422	-0.33449	-0.05298	-0.45960	0.22912	0.15849	0.01912	0.12941	0.54627	-0.44948	0.31168	0.98375	1.00000	0.40629	1.00000

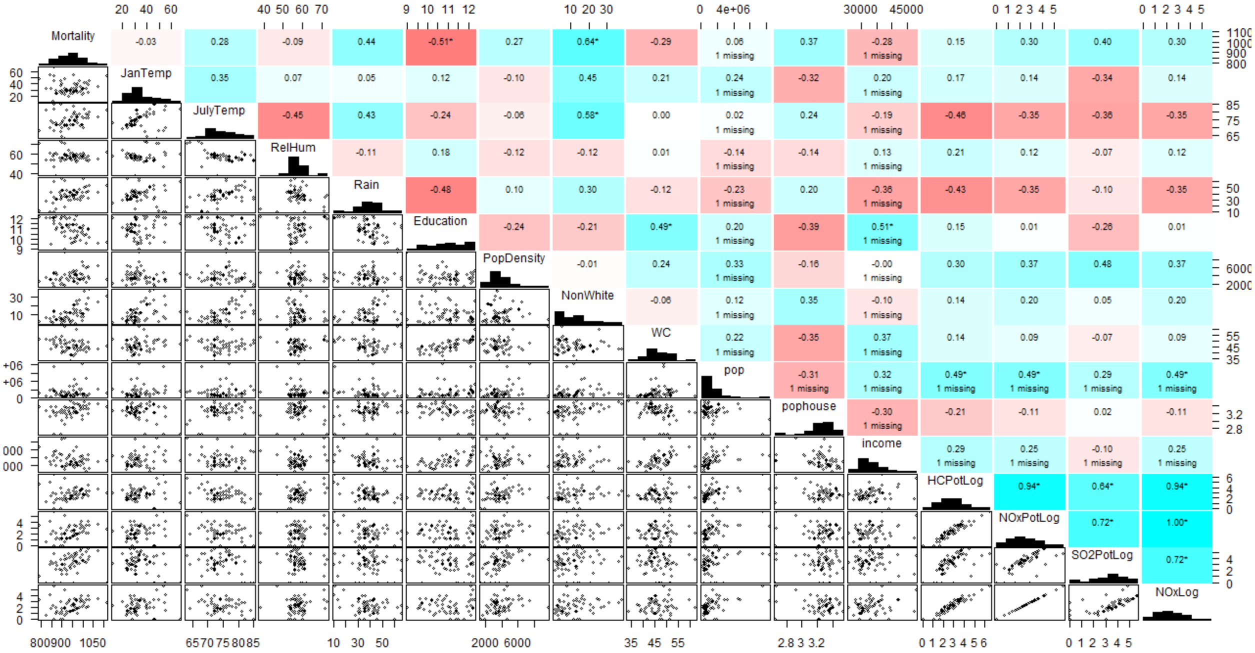
1k. Scatterplot Matrix – SAS



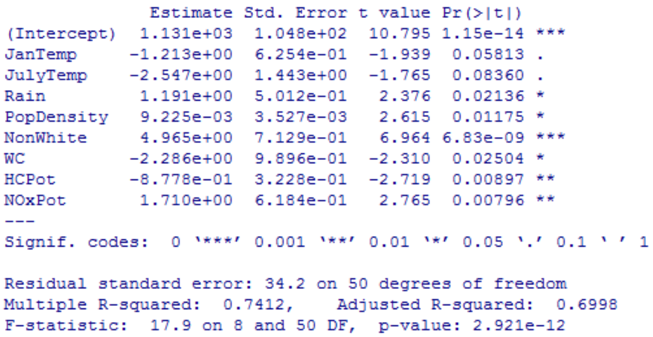
1l. Transformed Scatterplots, HCPot, NOx, NOxPot, SO2Pot – R



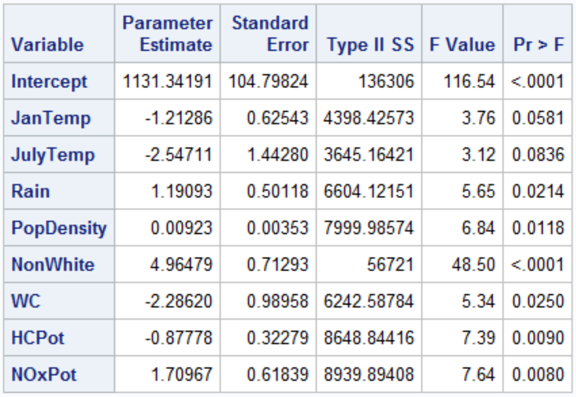
1m. Scatterplot Matrix with Transformations – R



1n. Results of Stepwise Regression – R & SAS

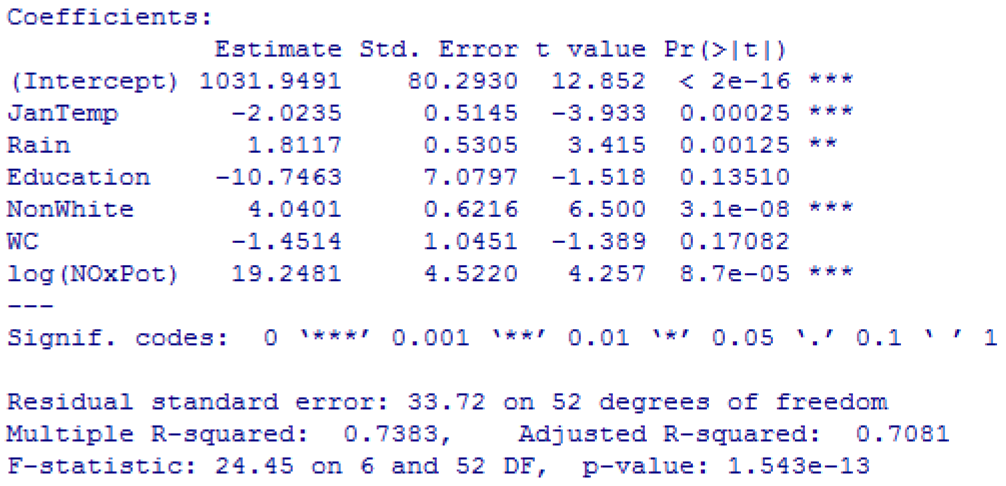


```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.131e+03  1.048e+02  10.795 1.15e-14 ***
JanTemp      -1.213e+00  6.254e-01  -1.939  0.05813 .  
JulyTemp     -2.547e+00  1.443e+00  -1.765  0.08360 .  
Rain          1.191e+00  5.012e-01   2.376  0.02136 *  
PopDensity    9.225e-03  3.527e-03   2.615  0.01175 *  
NonWhite      4.965e+00  7.129e-01   6.964 6.83e-09 ***
WC           -2.286e+00  9.896e-01  -2.310  0.02504 *  
HCPot        -8.778e-01  3.228e-01  -2.719  0.00897 ** 
NOxPot        1.710e+00  6.184e-01   2.765  0.00796 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.2 on 50 degrees of freedom
Multiple R-squared:  0.7412,	Adjusted R-squared:  0.6998 
F-statistic:  17.9 on 8 and 50 DF,  p-value: 2.921e-12
```



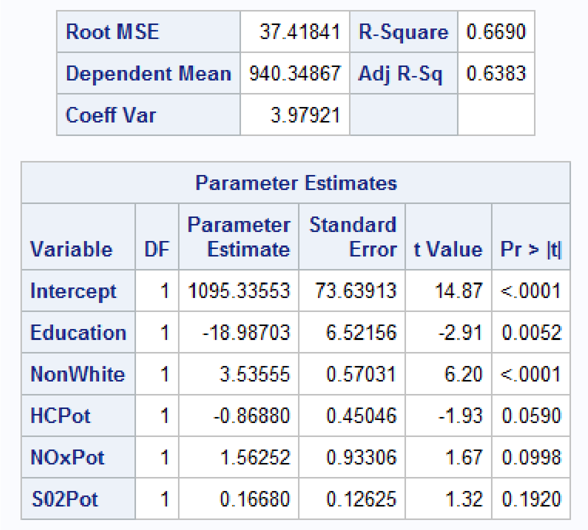
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1131.34191	104.79824	136306	116.54	<.0001
JanTemp	-1.21286	0.62543	4398.42573	3.76	0.0581
JulyTemp	-2.54711	1.44280	3645.16421	3.12	0.0836
Rain	1.19093	0.50118	6604.12151	5.65	0.0214
PopDensity	0.00923	0.00353	7999.98574	6.84	0.0118
NonWhite	4.96479	0.71293	56721	48.50	<.0001
WC	-2.28620	0.98958	6242.58784	5.34	0.0250
HCPot	-0.87778	0.32279	8648.84416	7.39	0.0090
NOxPot	1.70967	0.61839	8939.89408	7.64	0.0080

1m. Results of Stepwise Regression – R



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1031.9491    80.2930  12.852  < 2e-16 ***
JanTemp       -2.0235     0.5145  -3.933  0.00025 ***
Rain           1.8117     0.5305   3.415  0.00125 ** 
Education    -10.7463     7.0797  -1.518  0.13510    
NonWhite       4.0401     0.6216   6.500  3.1e-08 ***
WC            -1.4514     1.0451  -1.389  0.17082    
log(NOxPot)   19.2481     4.5220   4.257  8.7e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.72 on 52 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7081 
F-statistic: 24.45 on 6 and 52 DF,  p-value: 1.543e-13
```

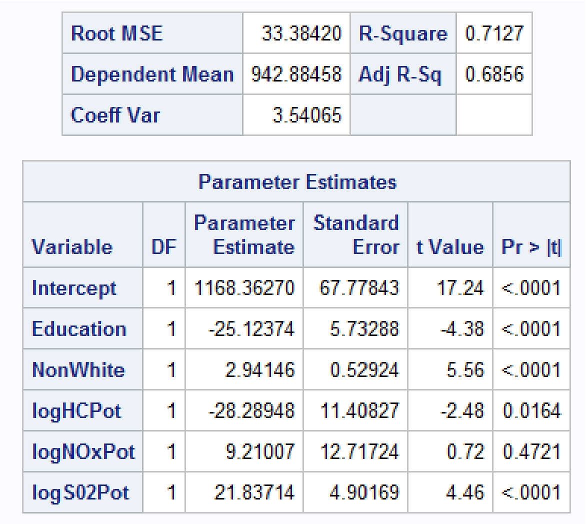
1n. Manually Model Building Results - SAS



Root MSE	37.41841	R-Square	0.6690
Dependent Mean	940.34867	Adj R-Sq	0.6383
Coeff Var	3.97921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1095.33553	73.63913	14.87	<.0001
Education	1	-18.98703	6.52156	-2.91	0.0052
NonWhite	1	3.53555	0.57031	6.20	<.0001
HCPot	1	-0.86880	0.45046	-1.93	0.0590
NOxPot	1	1.56252	0.93306	1.67	0.0998
S02Pot	1	0.16680	0.12625	1.32	0.1920

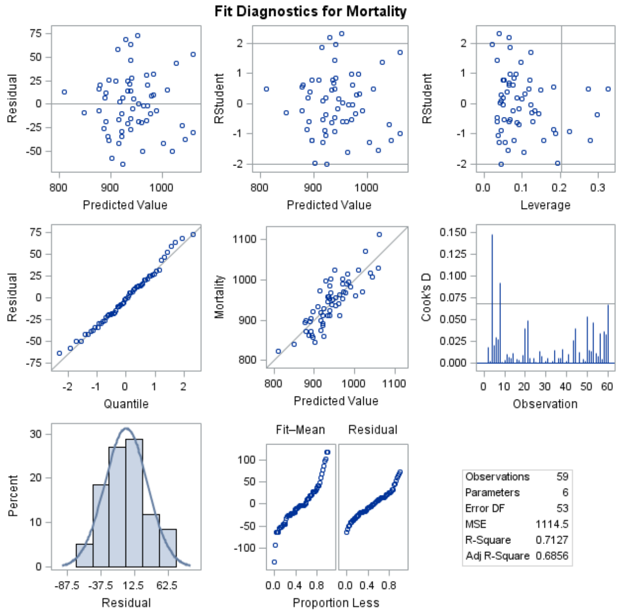


Root MSE	33.38420	R-Square	0.7127
Dependent Mean	942.88458	Adj R-Sq	0.6856
Coeff Var	3.54065		

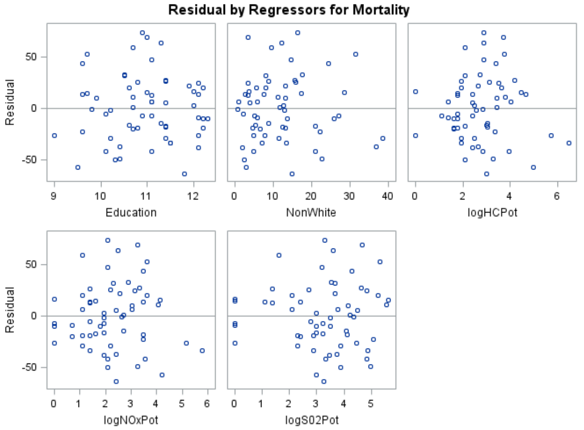
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1168.36270	67.77843	17.24	<.0001
Education	1	-25.12374	5.73288	-4.38	<.0001
NonWhite	1	2.94146	0.52924	5.56	<.0001
logHCPot	1	-28.28948	11.40827	-2.48	0.0164
logNOxPot	1	9.21007	12.71724	0.72	0.4721
logS02Pot	1	21.83714	4.90169	4.46	<.0001

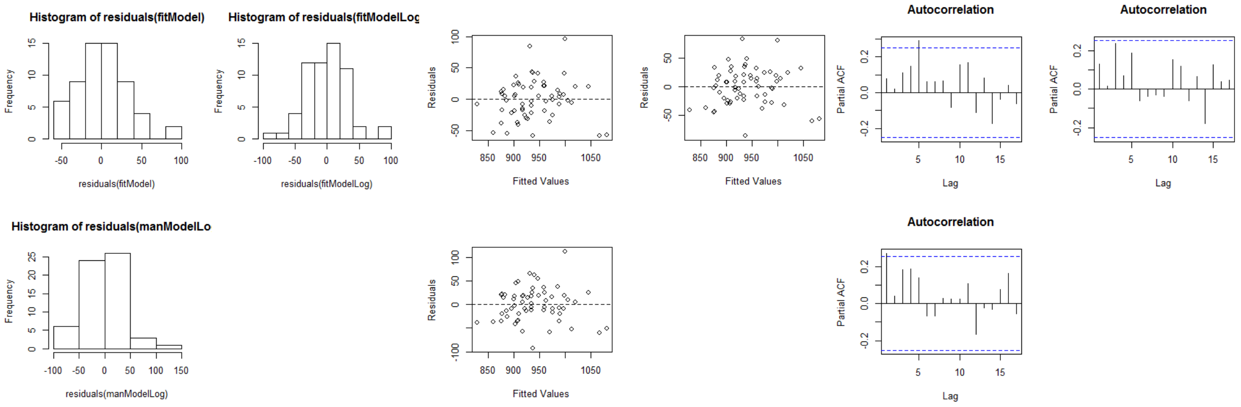
1o. Residuals - SAS



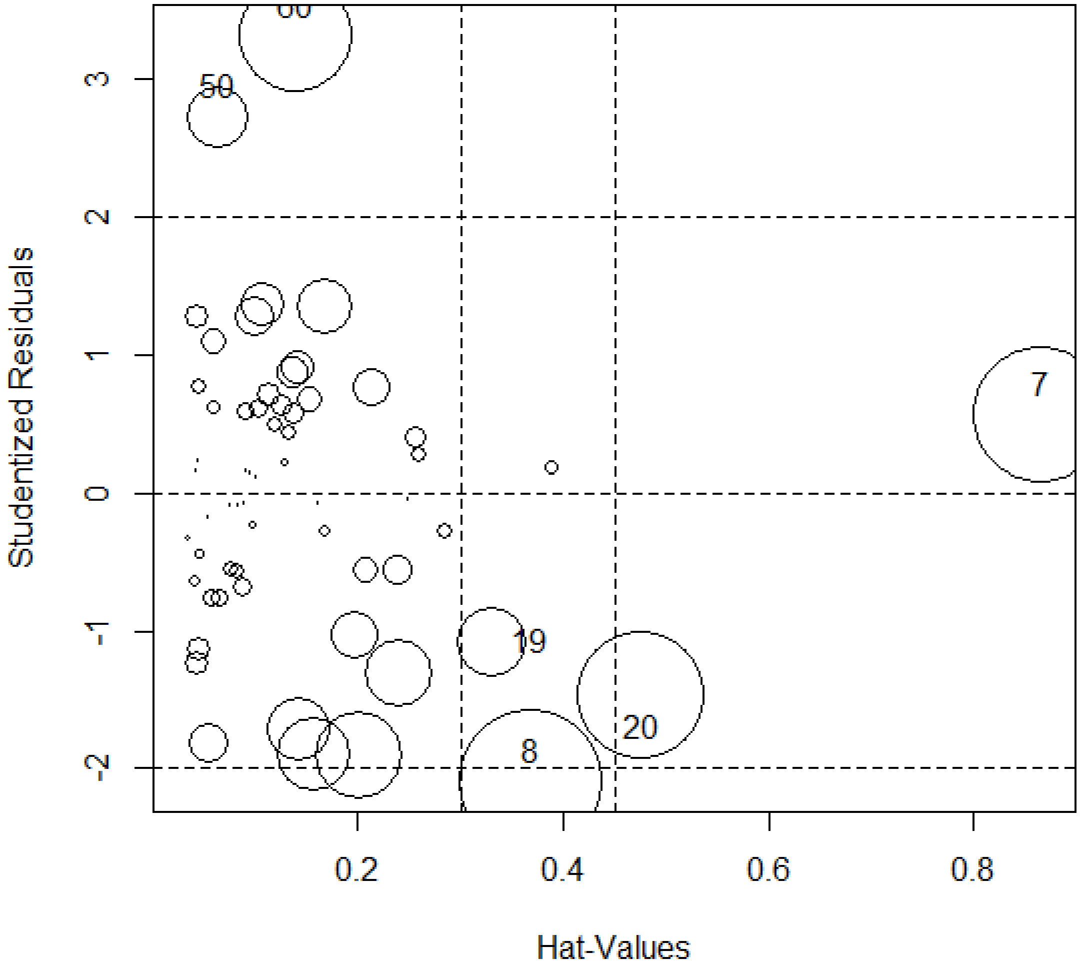
```
Observations      59
Parameters         6
Error DF          53
MSE           1114.5
R-Square      0.7127
Adj R-Square  0.6856
```



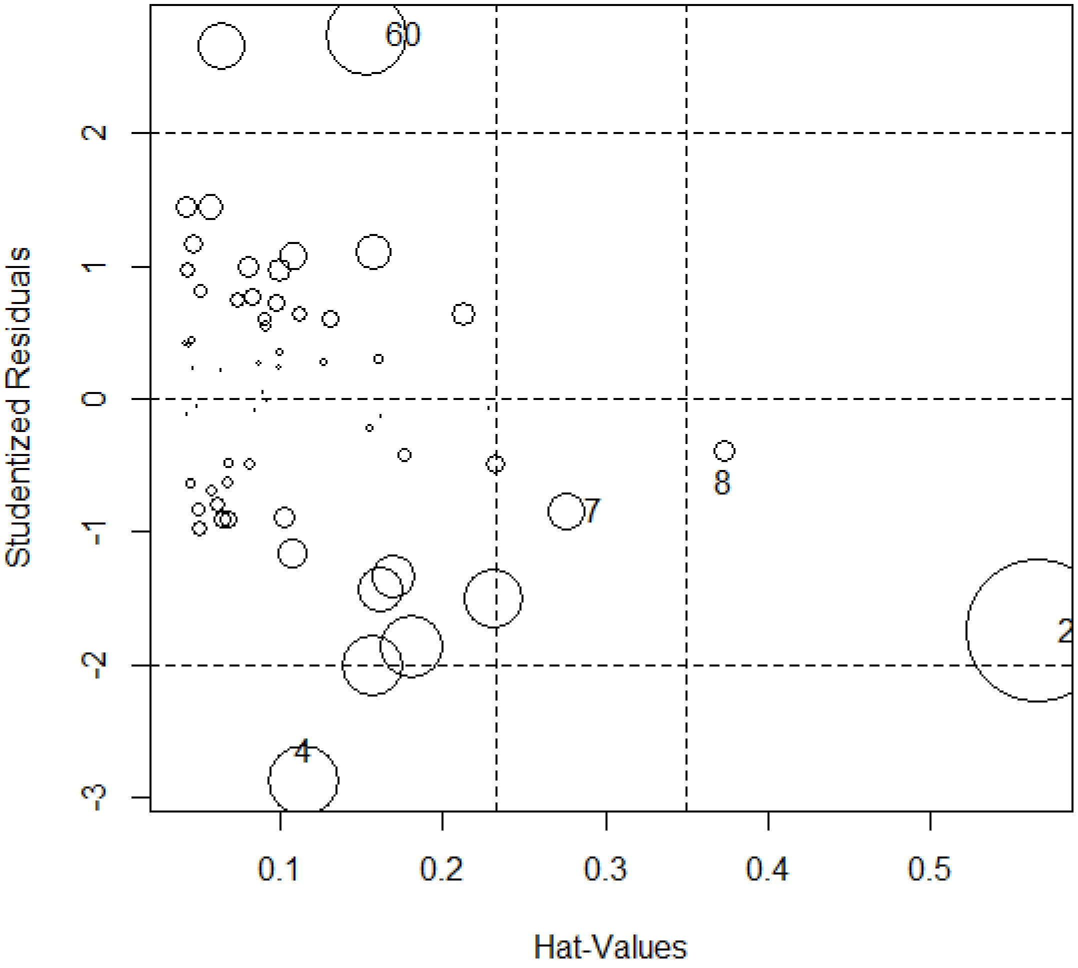
1p. Residuals – R



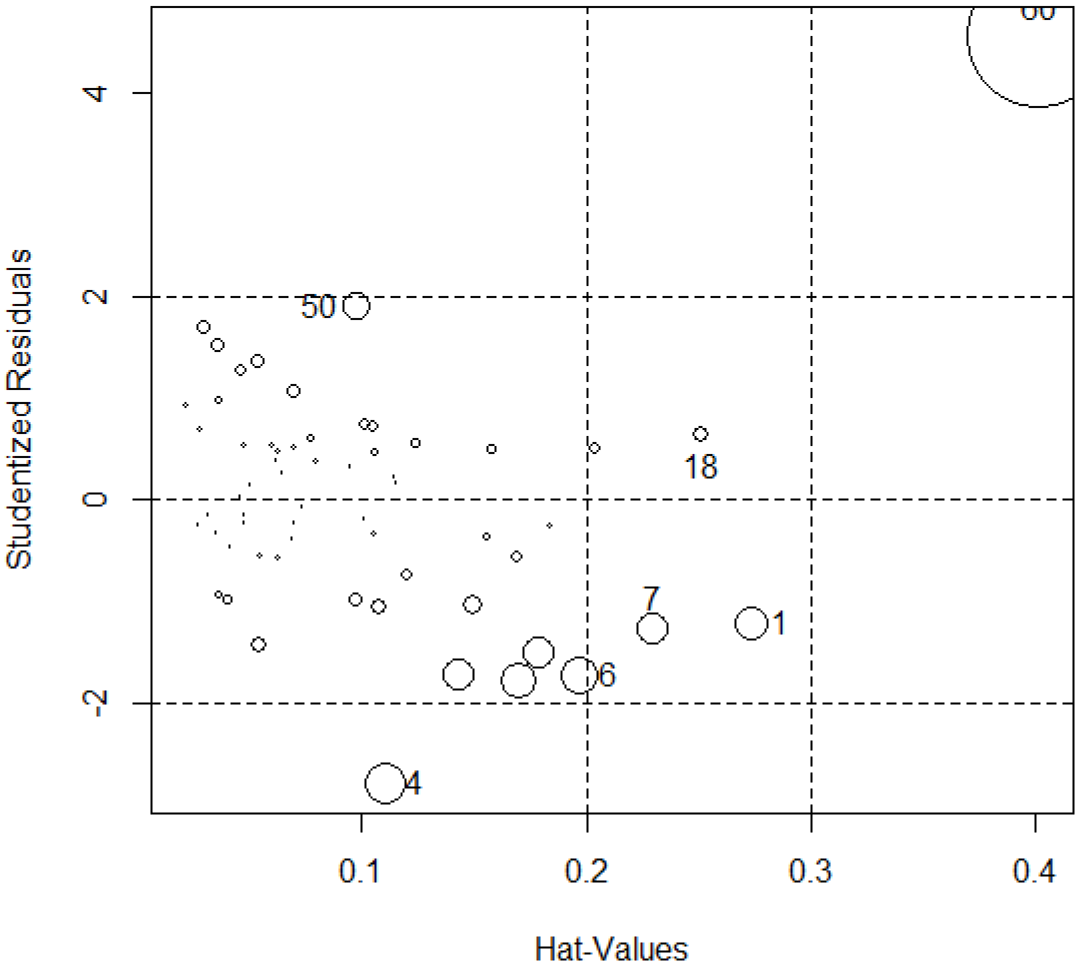
1q. Influence Plot – Model 1



1r. Influence Plot – Model 2



1s. Influence Plot – Model 3


Studentized Residuals
Hat-Values