Modeling Mortality Rates with Environmental Variables

Xuân Duong Fernandez, Katherine Davis, Anthony Weishampel, Austin McMillan

Part 1: Introduction

The dataset our group chose is a well-known one from 1973 by G.C. McDonald and R.C. Schwing on the relationship between various environmental and demographic variables on mortality rates in metropolitan regions throughout the continental United States. Specifically, we chose to look at environmental variables and their relative predictive values when examined against mortality. The question that guided us was, which environmental variables were the best predictors of mortality?

Part 2: Data Collection

Data: McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

Specifically looking at the environmental factors (listed below) in the SMSA and total age-adjusted mortality rate per 100,000 in the SMSA.

Cases: 60 Standard Metropolitan Statistical Areas (populated regions, defined in the original report)

Variables: relative hydrocarbon pollution potentials, relative nitric oxide potentials, relative sulfur dioxide potentials, precipitation, mean January temperature, mean July temperature, mean annual humidity, and age-adjusted mortality rate (deaths per 100,000 people in population).

(The pollution potentials are the products of the tons emitted of the specified pollutant per day per square kilometer and of a dispersion factor –which takes into account mixing height, wind speed, and dimension of the SMSA.)

Sample: The sample is the 60 SMSAs, and their measures of the variables above

Population: All urban areas in the continental U.S.

Parameters of interest: Relationship between these environmental variables and mortality rates.

The data for the environmental variables and mortality rates were compiled from U.S. Department of Commerce reports and two other papers (see references 30, 29, 1, and 4 in McDonald and Schwing), with the greatest gap in data collection across sources being 9 years (1962 and 1971). In this project, we assume that there is no significant difference in atmospheric measurements and mortality rates across this gap. Since this analysis relies on observational data, and there is no random assignment, a causal relationship cannot be inferred. Furthermore, the SMSAs were chosen based on availability of data (a convenience sample), so our results are not generalizable.

Part 3: Exploratory Data Analysis

We began by creating scatter plots of each of the environmental variables and mortality in order to see the relationship between each explanatory variable and mortality (the response variable), and to check for outliers that may be influential, meaning that they would influence the regression lines created by our models and therefore confound our analysis. Figures 1 through 7 show these initial plots.

We noted outliers that appeared to be both high-leverage and influential in the variables hydrocarbon pollution potential and nitric oxide pollution potential (see figures 5 and 6). We identified the highest-leverage outlier (the one with the highest hydrocarbon pollution potential, which was the same case that had the highest nitric oxide pollution potential) and removed it from the dataset, then recreated the scatter plots for those two variables to determine if removing that outlier was sufficient to make the regression line appear to match the main body of the data. Both regression lines still showed significant deviation from the majority of the data due to the remaining outliers, so we repeated the process until the regression lines appeared to match the body of the data, at which point we had removed a total of four outliers. We were unable to determine which SMSAs these outliers were, as the dataset we were using did not include SMSA names. We then re-plotted mortality vs each environmental variable, using the new dataset with outliers removed (new plots are shown in figures 8 through 14).

Variable	Multiple R^2	p-value
Precipitation	0.1323	0.00586
January temperature	0.03482	0.169
July temperature	0.02762	0.221
Hydrocarbons	0.1774	0.00123
Nitric oxide	0.2787	2.9e-05

Sulfur dioxide	0.2143	0.000327
Humidity	0.0002856	0.902

As shown in the table, the precipitation, hydrocarbon pollution potential, nitric oxide pollution potential, and sulfur dioxide pollution potential variables have high R^2 values, indicating that they explain a substantial portion of variability in mortality, and low p-values, indicating that they are significant predictors. Therefore, we would expect to see these variables in a multiple linear regression model for mortality.

Based on the scatter plots for mortality vs each explanatory environmental variable (figures 8 through 14), it appears that there is a moderate positive linear relationship between mortality and precipitation, a weak positive linear relationship between mortality and January temperature, a weak positive linear relationship between mortality and July temperature, a moderate positive linear relationship between mortality and hydrocarbon pollution potential, a moderate positive linear relationship between mortality and nitric oxide pollution potential, a moderate positive linear relationship between mortality and sulfur dioxide pollution potential, and no relationship between mortality and humidity.

Part 4: Methodology

A stepwise multi-variable linear regression model was calculated through the process of backward elimination strategy to construct a model that best predicts the mortality of the SMSA through the explanatory environmental variables. This process begins with the inclusion of all of the variables and then eliminates all of the insignificant variables, and results in a model that is determined by only the significant variables. The resulting linear regression model is the model that has the highest possible value of the adjusted R^2. This process is an effective method for determining if there are any significant relationships between mortality rates and the environmental variables. The two hypotheses that are being tested in this analysis are the null hypothesis, that the variables are not significant determinants of mortality, and the alternate hypothesis, that the variables are significant determinants of mortality.

H_0: $\beta_i = 0$ when all other explanatory variables are included in the model

H_{ALT}: $\beta_i \neq 0$ when all other explanatory variables are included in the model

The p-value for each slope is used to figure out if there is enough evidence to reject the null hypothesis and include the explanatory variable in the model.

Part 5: Results

The first model (Equation 1) includes all of the environmental variables. The application of the backward elimination strategy in the calculation of the stepwise model eliminated all of the insignificant variables. If all of the variables were significant determinants of the mortality then the final model would be that of equation 1; however, this was not the case. The resulting model (Equation 2) consisted of two significant explanatory variables - precipitation and nitric oxide. Both of these variables were significant at the significance level of 0.05. The p-value of for the nitric oxide is 2.14 x 10^{-6}, and the p-value for the precipitation is 0.000354. Because both of the p-values for these variables are significant we can reject the null hypotheses for both of these variables.

$$(1) \quad \widehat{Mortality} = \beta_0 + \beta_1(Prec) + \beta_2(JanTemp) + \beta_3(JulTemp) + \beta_4(HC) + \beta_5(NOx) + \beta_6(SO)$$

$$(2) \quad \widehat{Mortality} = 2.3262(NOx) + 2.9707(Prec) + 798.5399$$

The final model (Equation 2) provided the regression model with the highest value of $R^2_{adj.}$ The resulting adjusted R-square value is 0.413, and the multiple R^2 is 0.4343, meaning that the model (with explanatory variables nitric acid and precipitation) explains 43.43% of variability in mortality per 100,000 people. The correlation coefficient for the model is 0.6426. This correlation coefficient suggests that there is a moderately strong linear correlation in the prediction of mortality based on nitric oxide and precipitation.

This model suggests that for every increase of 1 in the relative pollution potential of nitric oxide, there is an expected increase in mortality of 2.3262 deaths. The slope of the precipitation from the model concludes that for every inch increase in the mean annual precipitation, there is an expected increase in mortality of 2.9707 deaths. The intercept of the model suggests that if the mean annual precipitation was 0 inches and there was no relative pollution potential of oxides of nitrogen, than the expected mortality is 798.5399 deaths per 100,000.

Figure 15 illustrates the plane that the final model predicted. Nitric oxide is represented along the x-axis, and precipitation is on the y-axis. The expected mortality is on the z-axis. This plane illustrates that there are expected positive relationships with mortality and nitric oxide, and mortality and precipitation. These expected positive relationships explain the increase of mortality as a given point moves further away from the origin.

Model Diagnostics:

In order to appropriately use multiple regression methods using the model $\hat{y}= \beta_0 + \beta_{11}x_1 + \beta_2x_2 +\ldots.\beta_px_p$, the following assumptions must be satisfied (note that all diagnostics were done using the revised dataset with outliers removed, as discussed in part 3, because this was the dataset used for the models):

1. Nearly normal residuals. Nearly normal residuals mean that the residuals are normally distributed around 0, with about half above 0 and half below. To ensure that this condition is reasonable, the residuals of the data must resemble a nearly straight line on a normal probability plot. The normal probability plot of the residuals (figure 17) showed a nearly straight line, although with a few deviating points at either extremes, so we concluded that the residuals were nearly normal.
2. Nearly constant variability in the residuals. This condition was assessed by plotting the residuals against the fitted or predicted values of the mortality rate (figure 16). Given the fact that the majority of the residuals were centered around the line y = 0, and that there was no relationship seen between fitted values and residuals, it is reasonable to say that there is constant variance in the residuals of the data; i.e. there were no major deviations from constant variance in the plot of the residuals against the predicted values of the response variable.
3. Independent residuals. This condition was checked by plotting the residuals in order of their data collection (figure 18). Again, with this plot it is apparent that the majority of the residuals are centered around the line at y = 0. Independence is considered to be reasonable given the fact that

there is no noticeable relationship in the residuals, either positive or negative. Considering a graph of the residuals plotted in order of their data collection is useful because it assists in identifying any connection between observations that were collected close to one another.

4. Each predictor variable is linearly related to the outcome. This was assessed with graphs of each predictor variable plotted against the response variable (figures 8-14), as discussed above in part 3. It is reasonable to say that each variable is linearly related to the outcome because each plot consisted of a linear relationship between the explanatory variable and the response variable.

Part 6: Conclusion

From the statistical analyses it can be concluded that the best model for determining mortality per 100,000 in a SMSA given the environmental factors is a linear regression model with the mean annual precipitation and the pollution potential of nitric oxide as two significant explanatory variables. Even though it is surprising to have precipitation as a significant factor for the mortality rate of an SMSA, there are probably some untested confounding variables that would help explain this correlation. We cannot account for the confounding variables because they were not collected. It should also be mentioned that correlation does not prove causation and these were just observed data, so it cannot be concluded that mortality rates are causally influenced by precipitation and nitric oxide pollution. Only if the data was collected through a randomized experiment can causation be determined. Furthermore, because of the convenience sample bias demonstrated in the selection of the SMSAs, there is a limit to how much our results can generalize to all of urban continental U.S. Lastly, the data used in this report is already almost 40 years old, and thus it is severely limited in its applicability to current situations. We chose this 70s dataset in full understanding of this limitation, because of the scope and thoroughness of the measurements, and because we believed that value could still be found from studying past conditions alongside current ones. To sum up: random sampling, the collection of additional, up-to-date data for cases, and the inclusion of more environmental variables to explore possible confounds would help create a more accurate model for determining the mortality rate, and allow us to generalize our results with greater confidence.

Figures:

Figure 1:

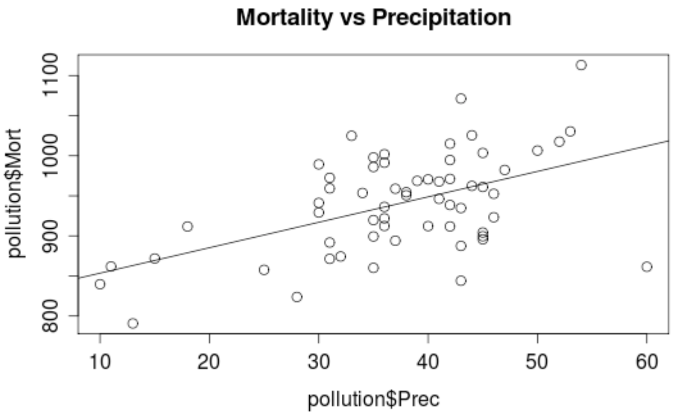


Figure 2:

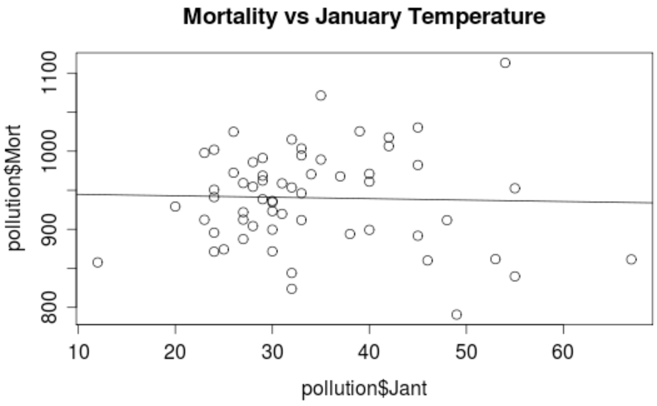


Figure 3:

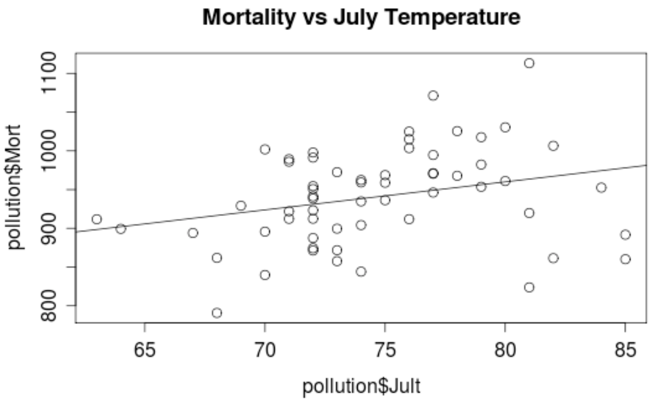


Figure 4:

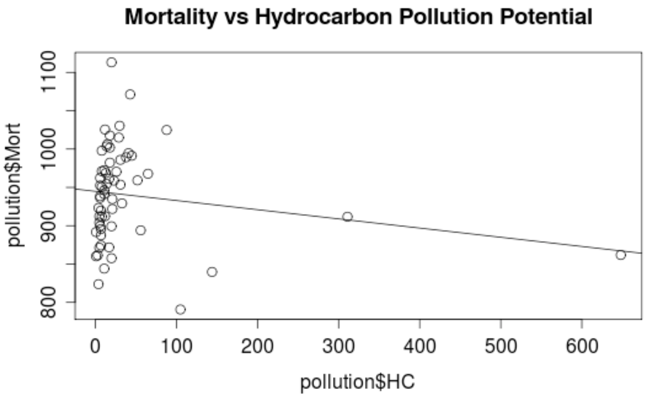


Figure 5:

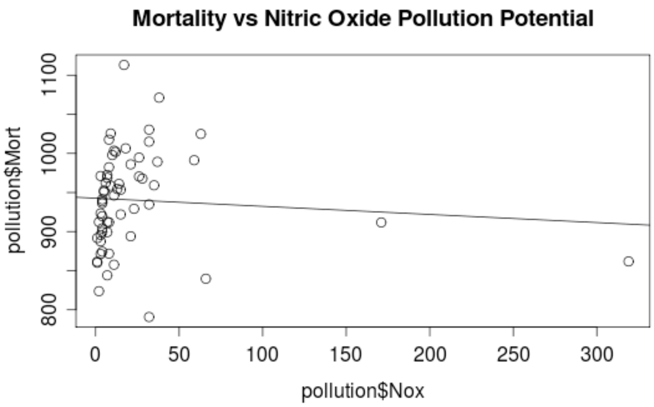


Figure 6:

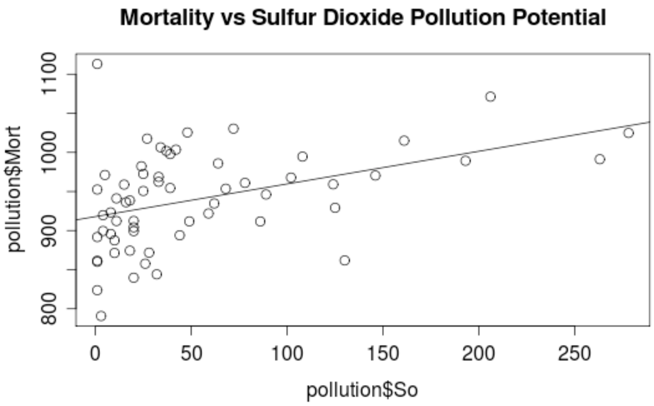


Figure 7:

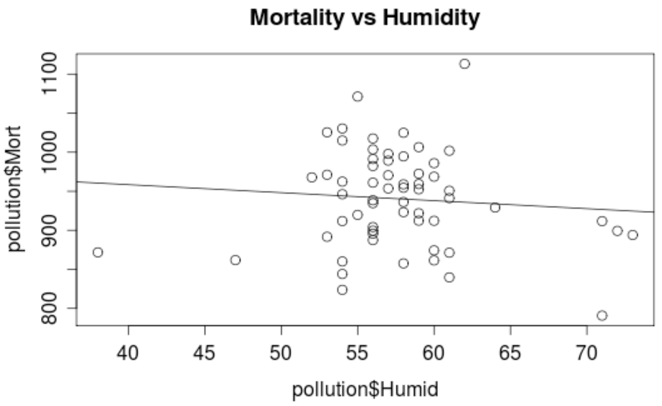

Mortality vs Humidity

Figure 8:

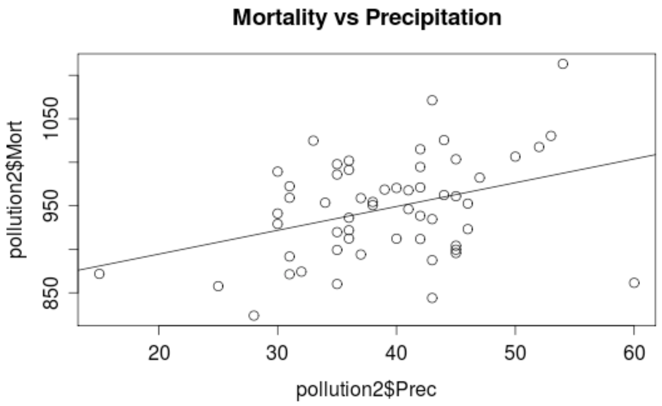

Mortality vs Precipitation

Figure 9:

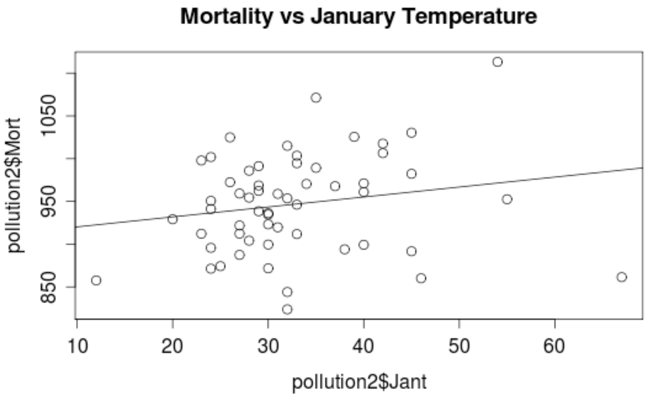

Mortality vs January Temperature

Figure 10:

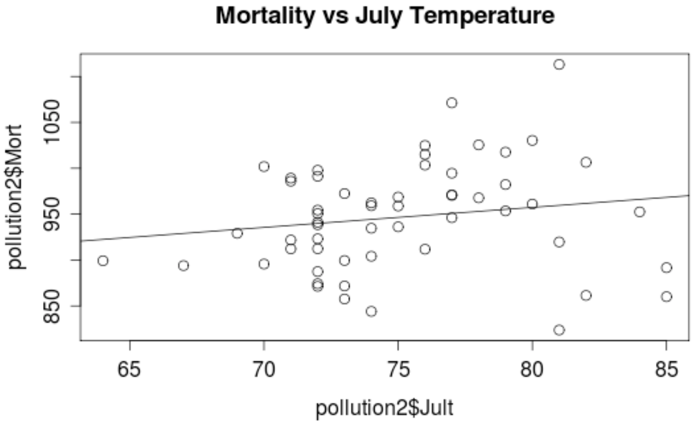

Mortality vs July Temperature

Figure 11:

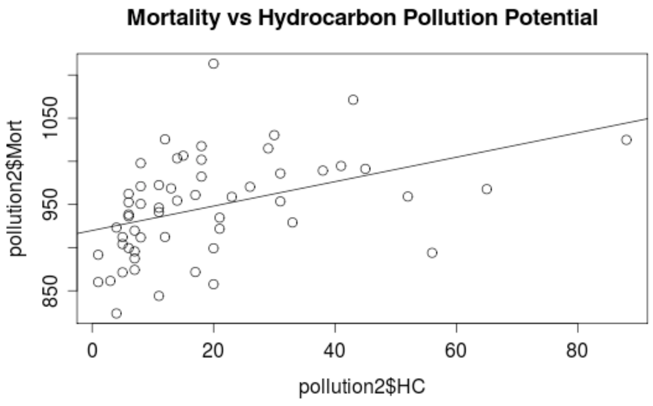

Mortality vs Hydrocarbon Pollution Potential

Figure 12:

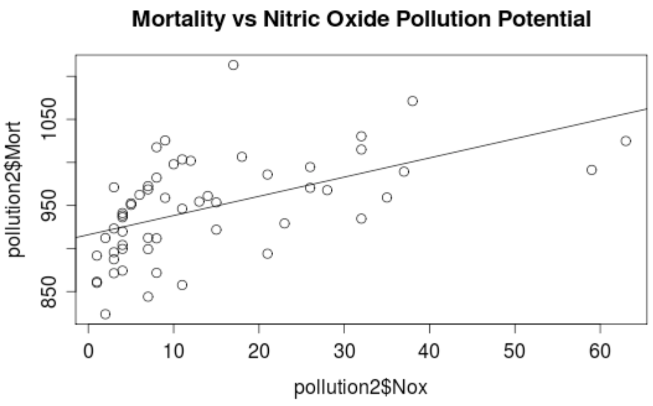

Mortality vs Nitric Oxide Pollution Potential

Figure 13:

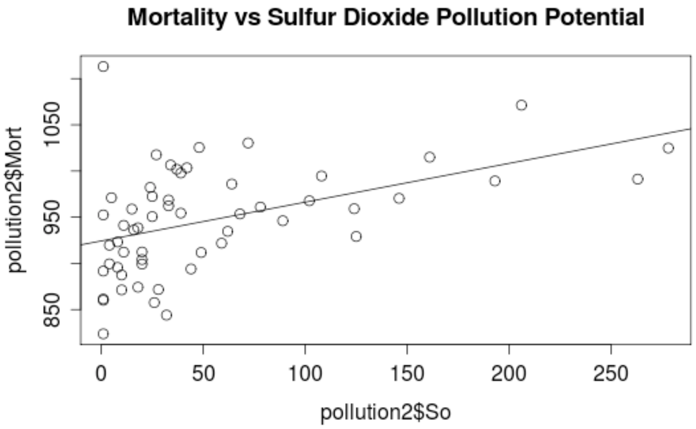


Figure 14:

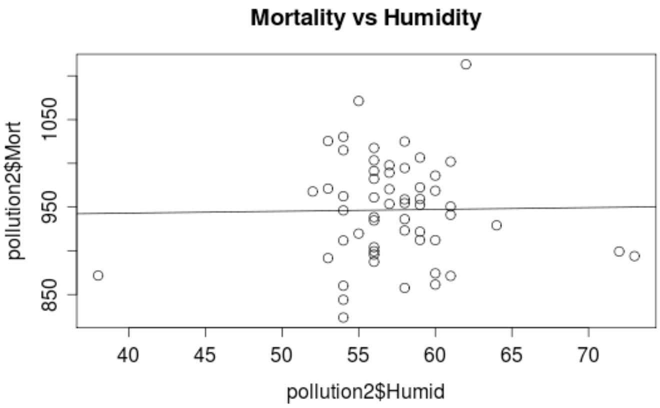


Figure 15:

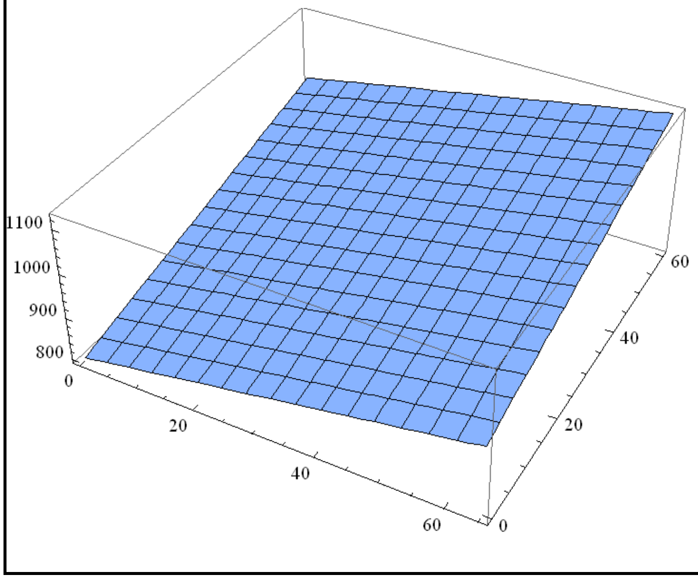


Figure 16:

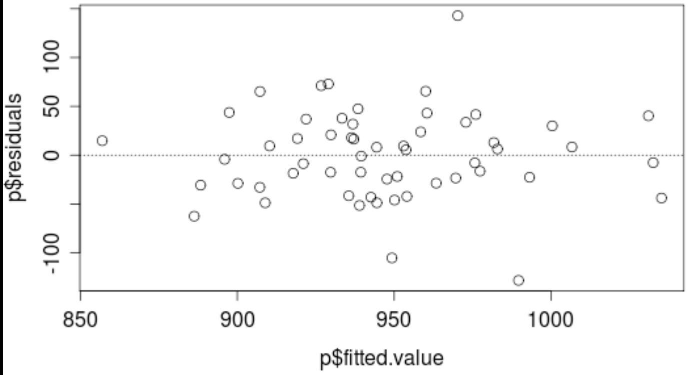


Figure 17:

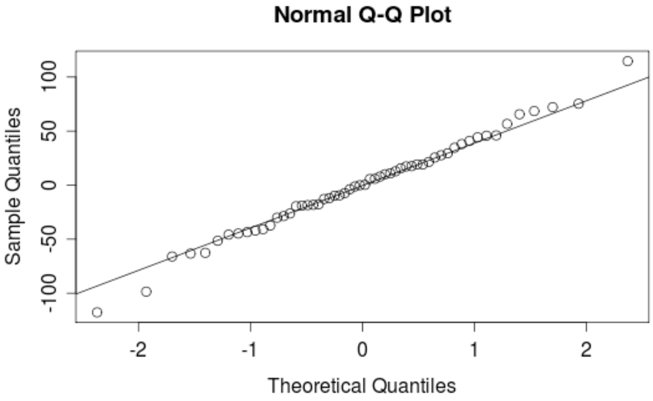
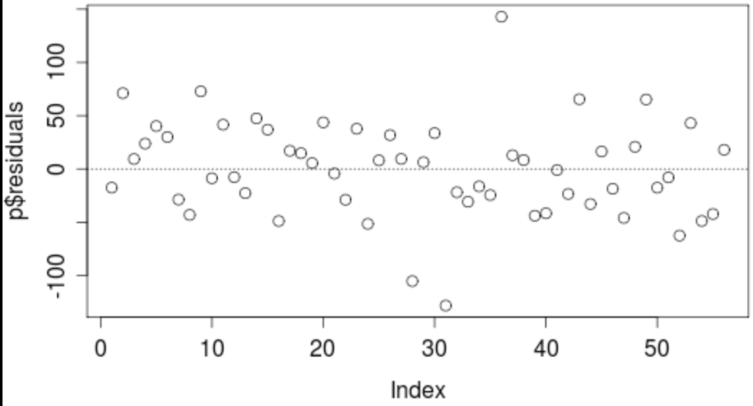


Figure 18:


p$residuals
Index