World Pollution

STAT 823: Summer Class Project, 2019

Mark Chintis

Department of Biostatistics and Data Science
University of Kansas, USA
July 26, 2019

Contents

List of Tables

List of Figures

Abstract

Pollution around the world is generally measured by carbon dioxide levels in the atmosphere. The data in this report take a differing approach and measure carbon dioxide emissions per capita. The amount of renewable energy countries produce and consume play a seemingly important role in such emissions. Intuition tells us that as renewable energy consumption increases, carbon dioxide emission will decrease. We would like to be able to determine to what extent of the carbon dioxide per capita is explained by the renewable energy consumption using a linear model. Ninety-three countries were included in a linear regression model and the renewable energy consumption is shown to explain 73% of the variation in the carbon dioxide emission with a p-value $< .05$. The results are lackluster as we determine that even a transformation of the response variable cannot make linear model residual assumptions hold true.

I. Introduction

When one thinks about the world's pollution, he or she probably often thinks about non-renewable energy use and consumption. But what if we could flip that idea and instead measure the world's renewable energy consumption? Would we see that the world's pollution would be significantly linearly dependent on such consumption? Will it reveal intuitive results? This report investigates exactly these linear relationships to identity if any correlation and/or association is present.

A. Study Design.

In a joint effort of data collection, World Bank, Sustainable Energy for All (SE4ALL) database from the SE4ALL Global Tracking Framework led jointly by the World Bank, International Energy Agency, and the Energy Sector Management Assistance Program collaborated to provide the renewable energy consumption for over 200 countries globally. These data are measured as a percent of total final energy consumption for their respective country.

In a separate project, the Carbon Dioxide Information Analysis Center, Environmental Sciences Division, and Oak Ridge National Laboratory collected data on a similar number of countries' carbon dioxide emissions. These data are measured in metric tons per capita. Data included in this report from both data sets are reduced to years 1990 to 2014. For the purposes of this analysis only complete cases are considered.

To build our final data set, the aforementioned data are joined and cleaned to include only data from the following countries:Argentina,Bolivia,Brazil,Chile,Columbia,Ecuador,Peru,Venezuela,Portugal,Germany,Greece,Ukrain Sweden,Finland,Ireland,Czech Republic, Italy,France,Austria,Belarus,Hungary,Bulgaria,Latvia,Poland,Romania,United Kingdom,Spain,Portugal,Estonia,Denmark, Switzerland, Albania, Norway, Netherlands, Algeria, Libya, Egypt, Chad, Niger, Mali, Mauritania, Guinea, Liberia, Senegal, Ghana, Benin, Nigeria, Cameroon, South Sudan, Ethiopia, Somalia, Kenya, Gabon, Angola, Zambia, Zimbabwe, Tanzania, Mozambique, Botswana, Namibia, South Africa, United States, Canada, Mexico, Guatemala, Belize, Nicaragua, Costa Rica, Panama, Cuba, Jamaica, Dominican Republic, China, India, Japan, Vietnam, Cambodia, Thailand, Myanmar, Nepal, Bhutan, Mongolia, Afghanistan, Kazakhstan, Turkmenistan, Pakistan, Iraq, Turkey, Uzbekistan, Saudi Arabia, Oman, Jordan, Tajikistan. For reasons unknown, noteable countries not included in either data set and thus not in the data set for this report are: Iran, Yemen, South Korea, North Korea, Russia, Kyrgyzstan, Sudan.

Table 1: World Pollution

Country	X2	X1	Y
Angola	X1990	72.2552519	0.4317436

Table 1: World Pollution *(continued)*

Country	X2	X1	Y
Benin	X1990	93.7032409	0.1421576
Botswana	X1990	47.5846378	2.1003042
Cameroon	X1990	81.5931351	0.1472392
Algeria	X1990	0.1772283	2.9881098

B. Aims.

This report aims to prove intuitive notion that as renewable energy consumption increases, carbon dioxide emmision globally decreases. Using data from 93 countries, a linear relationship is sought out and the strength of the relationship is explored.

C. Statistical Model.

A multiple regression model is considered in this study. Let

Y_i = Pollution (metric tons per capita)
X_{i1}= Renewable energy consumption (percent of total energy consumption)
X_{i2}= Year, measured from 1990 to 2016

The initial model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim iidN(0, \sigma^2)$, $i = 1, 2, ..., 2639$, and $\beta_0, \beta_1, \beta_2$ and σ^2 are the unknown model parameters.

II. Preliminary Analyses.

A. Bivariate Associations.

Figure 1 gives a scatterplot matrix used to indicate any linear association between all variables. Pearson coefficients are overlaid on the scatterplot matrix for convenience. We see from the scatterplots and the Pearson coefficients that there are no multicollinearity issues to be addressed at this point.

B. Screening of Covariates and Verification of Assumptions

From the results of automatic variable selection methods (Appendix) along with criterion-based statistics, X_2 is excluded.

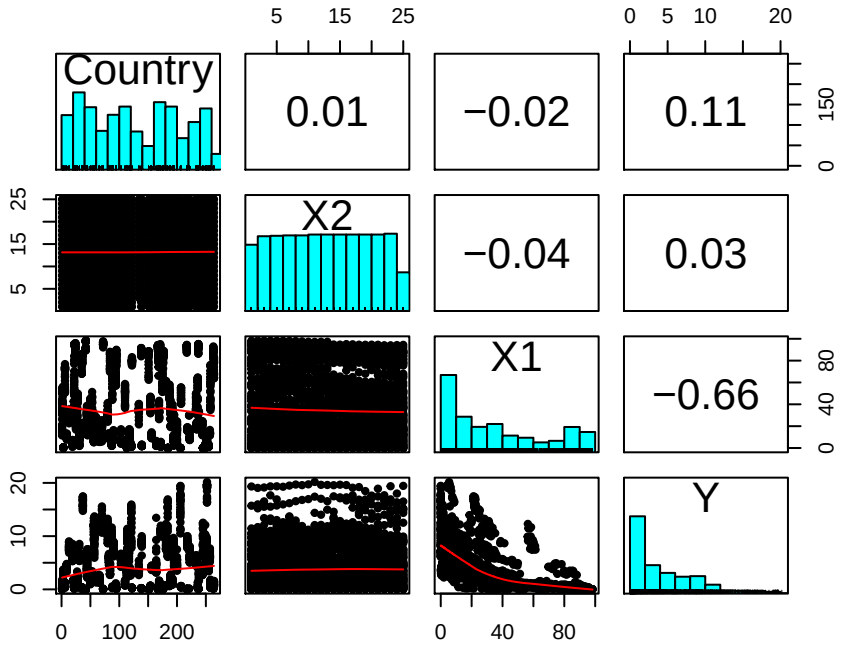


Figure 1: Scatterplot and Correlation Matrix

C. Final Model

The final model is given by

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

where $\varepsilon_i \sim iidN(0, \sigma^2)$, $i = 1, 2, ..., 2639$, and β_0, β_1 and σ^2 are the unknown model parameters.

III. Statistical Analysis.

A. Variable Selection/Model Reduction

Through automatic variable selection techniques described in the Appendix, two models are considered for hypothesis testing and validation. The full model for the hypothesis tests is given by

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_{i2} + \varepsilon_i.$$

The reduced model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i.$$

We wish to test the following hypotheses:

$$H_0: \quad \beta_1 = 0$$
$$H_1: \quad \beta_1 \neq 0.$$

The general linear test approach gives the decision rule to reject the null hypothesis that $\beta_1 = 0$, if $F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F} > F_{.95,1,2216}$. Since $F^* = 0.086 < F_{.95,1,2216} = 3.846$, we fail to reject the null hypothesis that $\beta_1 = 0$, thus we conclude that the incorporating the variable X_2 into the model is unnecessary. Leading us to use the model of that of the reduced model above.

The fitted model is displayed below. The rate of decrease of pollution is, on average, -0.092 (95% CI -0.096, -0.087) for every one percent increase in renewable energy consumption. From this model, the percent of renewable energy consumption explains 43.6% of the variation in global pollution.

```
##
## Call:
## lm(formula = Y ~ X1, data = dat_total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1456 -2.4052 -0.3223  1.2555 13.0714
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.605517   0.106245   71.58   <2e-16 ***
## X1          -0.091744   0.002219  -41.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.292 on 2216 degrees of freedom
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4353
## F-statistic:  1710 on 1 and 2216 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## X1           1  18528 18527.8  1709.9 < 2.2e-16 ***
## Residuals 2216  24012    10.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                   2.5 %      97.5 %
## (Intercept)  7.39716689  7.81386625
## X1          -0.09609445 -0.08739274
```

B. Transformations.

From the residual analysis shown in the Appendix, we find it necessary to transform the response variable, Pollution. We use the transformation by raising the Pollution variable to the one tenth power ($Y_i^{\frac{1}{10}}$). This will be our final model.

C. Final Model.

By consideration of variable selection and variable transformations, the final model of

$$Y_i^{\frac{1}{10}} = \beta_0 + \beta_1 X_{i1} + \epsilon_i.$$

The R summary of this model is shown below.

```
##
## Call:
## lm(formula = (dat_total$Y)^(1/10) ~ dat_total$X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32749 -0.04908 -0.00265  0.03975  0.30283
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.237e+00  2.769e-03  446.93   <2e-16 ***
## dat_total$X1 -4.471e-03  5.781e-05  -77.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08578 on 2216 degrees of freedom
## Multiple R-squared:  0.7297, Adjusted R-squared:  0.7295
## F-statistic:  5981 on 1 and 2216 DF,  p-value: < 2.2e-16
```

Now, in the final model, the rate of decrease of the square root of pollution is, on average, -0.004 (95% CI -0.005, -0.004) for every one percent increase in renewable energy consumption. From this model, the percent of renewable energy consumption explains 73% of the variation in global pollution.

```
##
## Call:
## lm(formula = Y^(1/10) ~ X1, data = proj_valid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11630 -0.05768 -0.02624  0.02581  0.29148
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2410866  0.0211950   58.56  < 2e-16 ***
## X1          -0.0044431  0.0004409  -10.08 1.33e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08632 on 33 degrees of freedom
## Multiple R-squared:  0.7548, Adjusted R-squared:  0.7473
## F-statistic: 101.6 on 1 and 33 DF,  p-value: 1.333e-11
```

```
## Analysis of Variance Table
##
## Response: Y^(1/10)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 0.75675 0.75675  101.56 1.333e-11 ***
## Residuals 33 0.24588 0.00745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To validate the predition capability of our final model, a random sample of 35 observations from the project data set are chosen to comprise a validation set. The observed mean square prediction error $MSPE = 0.007$ (shown above). This is very close to the $MSE = 0.007$, therefore giving an appropriate, unbiased indication of the predictive ability of the model.

IV. Summary of Findings.

The final model provides us with a quite strong correlation and negative association between our predictor variable, percentage of renewable energy consumption and carbon dioxide emissions, seen in Figure 2. This confers with the intuitive notion that the greater percentage of renewable energy consumed, the less carbon dioxide emission. Unfortunately, our model fails to pass the normality assumptions, leading to further question the appropriateness of a linear model at all and pushes towards other regression models, such as an exponential model.

V. Appendix

A. Diagnostics of Predictors

Boxplots, histograms, and summaries of each predictor variable are provided in Figure 3, Figure 4, and Figure 5, respectively.

We see that the predictor, renewable energy consumption is highly skewed left. This is likely explained by the number of countries with very little renewable energy production or availability in general. This should not impact the analysis as there are not immediate recognizable outliers.

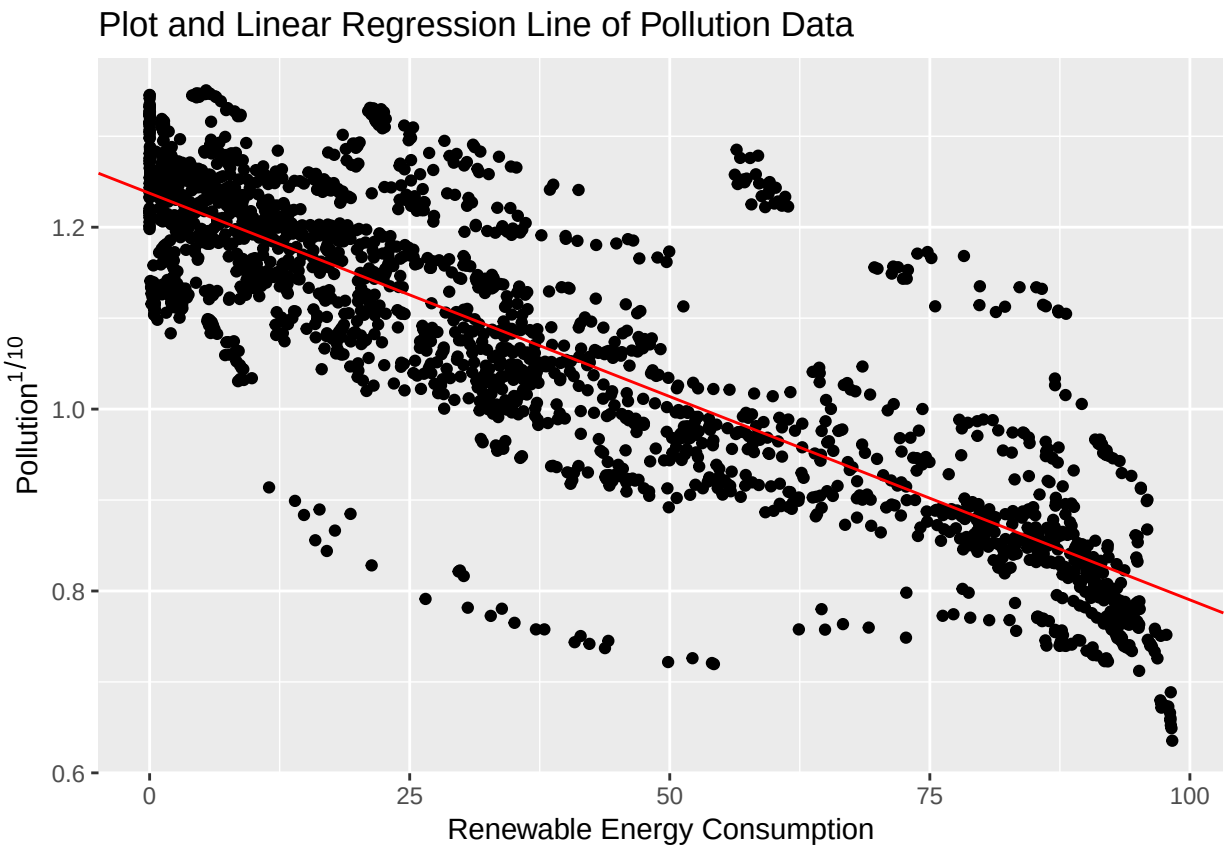


Figure 2: Scatterplot and Least-Square Regression Line of Pollution Data

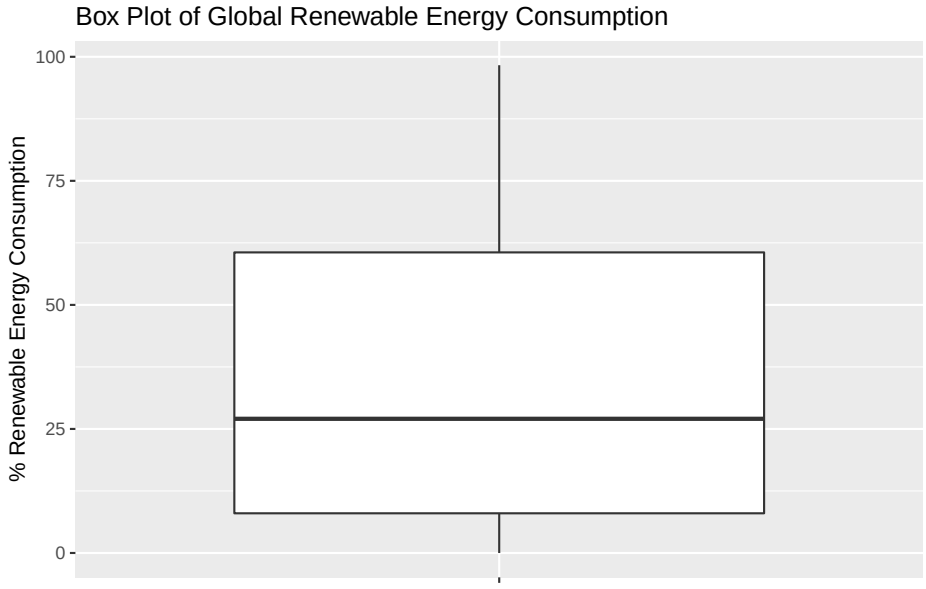


Figure 3: Boxplot of Renewable Energy Variable

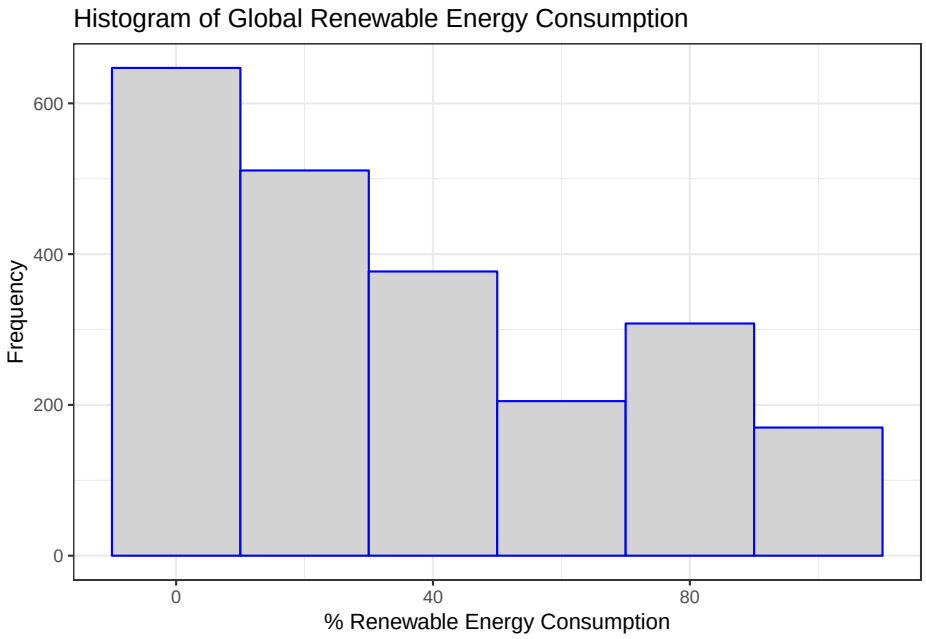


Figure 4: Histogram of Renewable Energy Variable

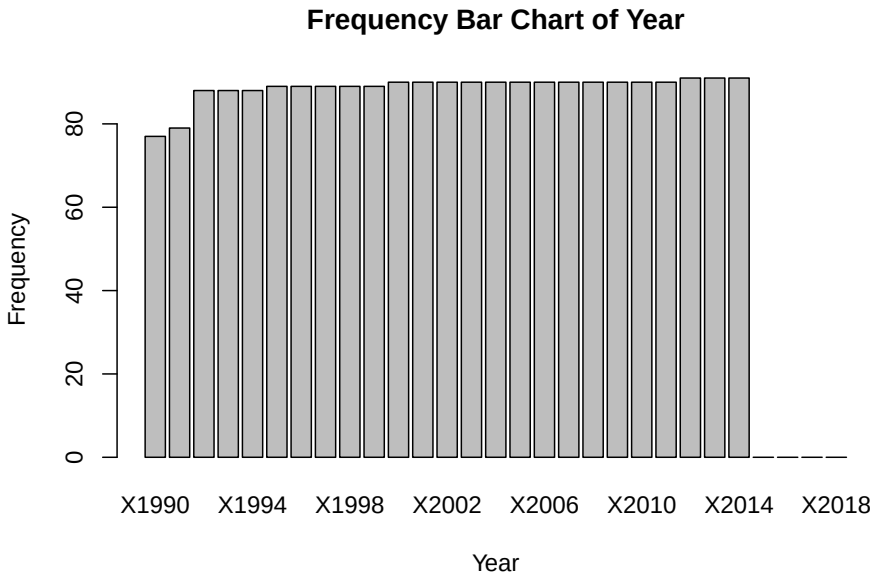
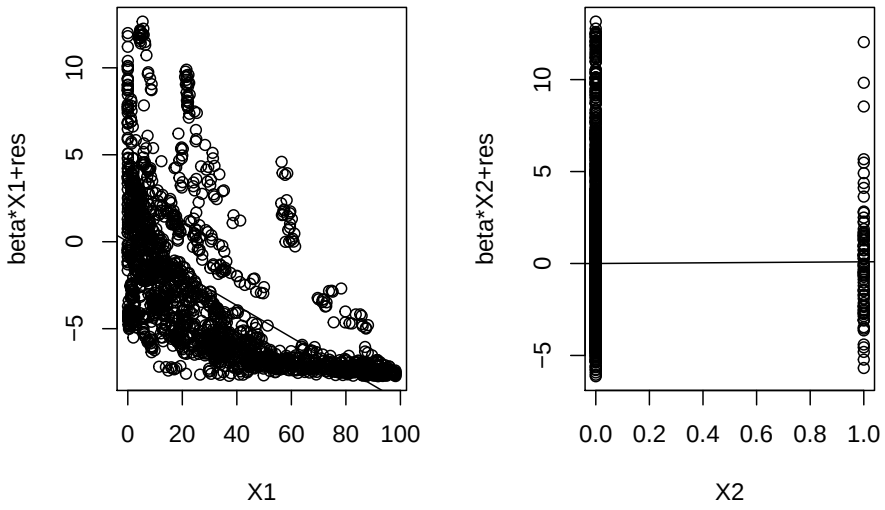


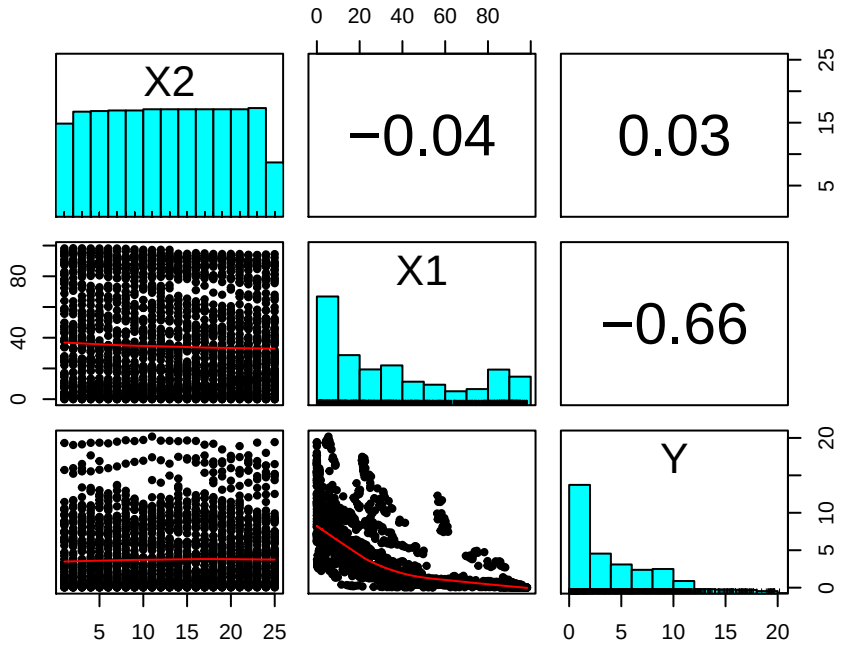
Figure 5: Frequency Bar Chart of Year Variable

B. Screening of Predictors

1. Below are added variable plots for each variable. From these plots we can begin to establish that the "Year" variable does not add much value since the slope of the linear relationship with the outcome is nearly zero. On the other hand, the "Renewable" variable look to have stong linear associations.



2. Multicollinearity is addressed with the scatterplot matrix and Pearson coefficients table shown below. It indicates very little multicollinearity in among the variables.



3. Automatic variable selection is used to begin the elimination process of variables, especially those with high multicollinearity. No variables are forced to stay in or out of the model. The results are shown in the R output below.

Table 2: Automatic Variable Selection Criteria

AdjustedR2	Cp	BIC
0.435	-23.943	-1253.052
0.435	-22.193	-1245.601
0.435	-20.396	-1238.102
0.435	-18.560	-1230.564
0.434	-16.761	-1223.063
0.434	-14.946	-1215.546
0.434	-13.091	-1207.989
0.434	-11.248	-1200.443

In a separate plots of all of the criteria, R^2_{adj}, Bayes Information Criterion (BIC), and Mallow's C_p statistic are prepared below.

With the best number of parameters highlighted in red in each plot, we determine that the model with only the "Renewable" variable, X1, to be included in the final model.

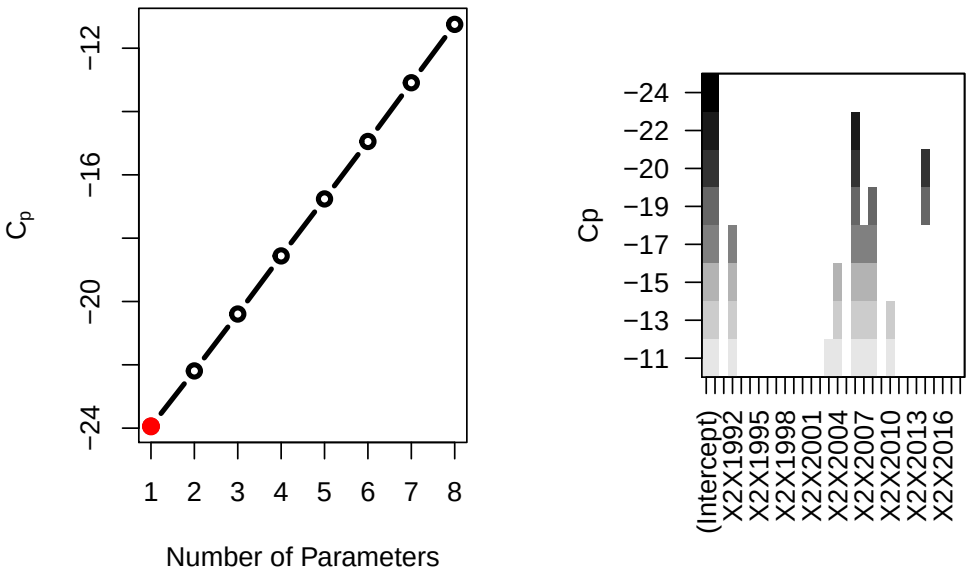


Figure 6: Cp variable Selection

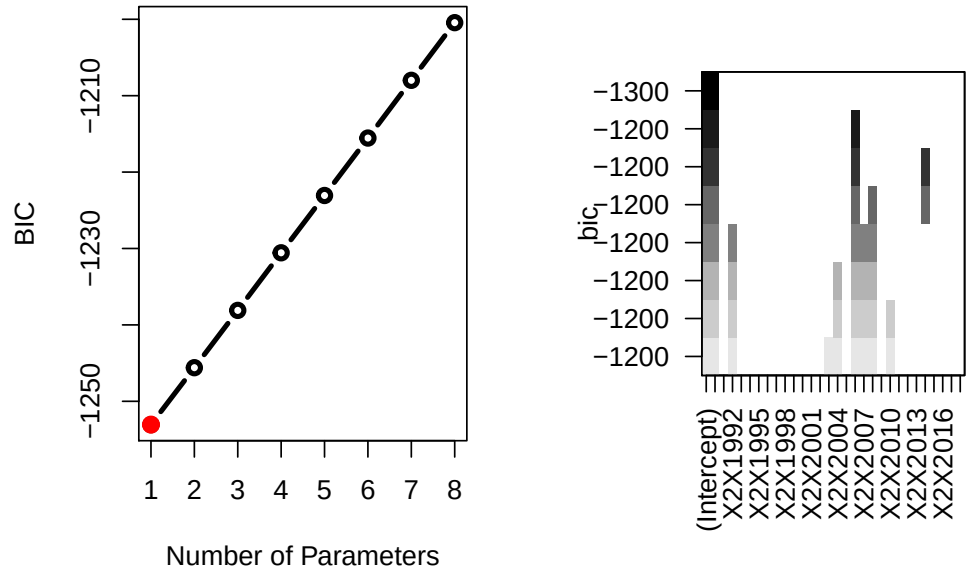


Figure 7: BIC variable Selection

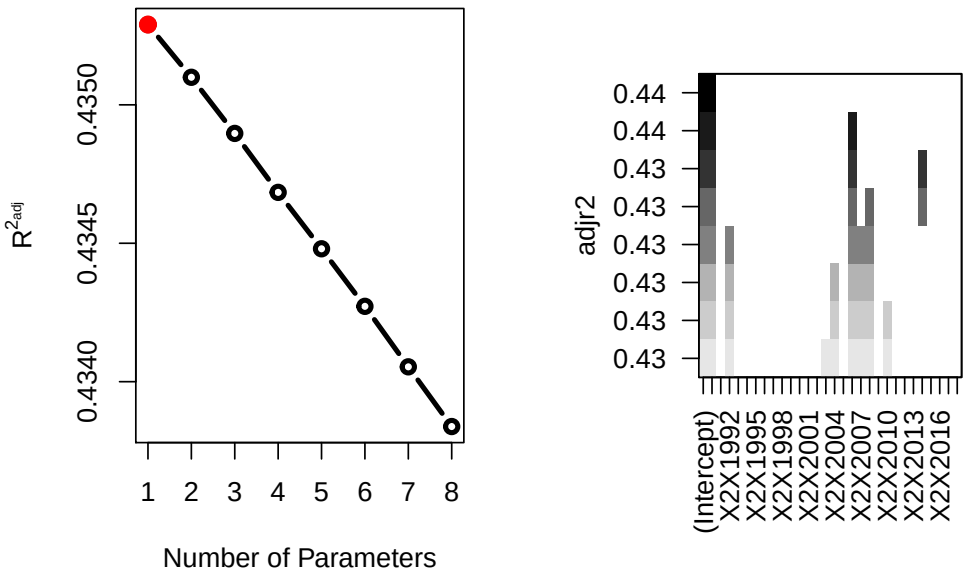


Figure 8: R squared adjusted variable Selection

4. Variable Inflation Factors. Variable inflation factors are shown for the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ in the R output below. Since each inflation factor is less than the rule of thumb of 10, none present any immediate issues.

```
##       X1  X2X1991  X2X1992  X2X1993  X2X1994  X2X1995  X2X1996  X2X1997
## 1.001795 1.953816 2.058333 2.058300 2.058310 2.069746 2.069786 2.069873
##  X2X1998  X2X1999  X2X2000  X2X2001  X2X2002  X2X2003  X2X2004  X2X2005
## 2.069961 2.069932 2.081259 2.081375 2.081415 2.081518 2.081552 2.081646
##  X2X2006  X2X2007  X2X2008  X2X2009  X2X2010  X2X2011  X2X2012  X2X2013
## 2.081720 2.081932 2.081891 2.081839 2.081931 2.082064 2.093503 2.093393
##  X2X2014
## 2.093524
```

C. Residual Diagnostics

Potentially influential points are identified using Cook's distances and plotted with a half-normal plot. These unusually large or small values are shown in Figure 9 below and values of observations 1621 and 1571 are extracted, respectively.

```
##      Country    X2       X1        Y
## 4039   Gabon X2005 85.24018 3.515238
## 3775   Gabon X2004 85.77171 3.461695
```

A fit of the model without observations 387 and 361 shows that our model is not changed

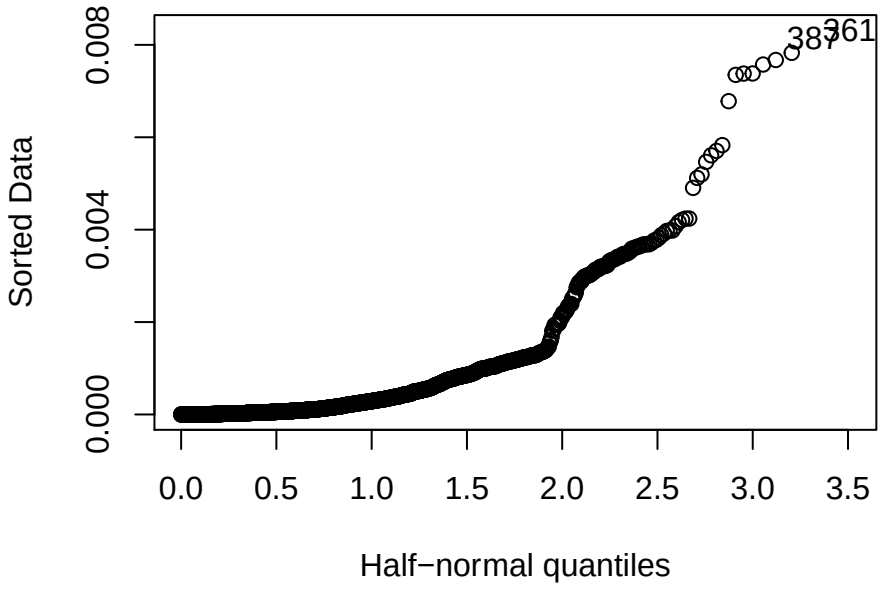


Figure 9: Cook's Distance in Half Normal Plot

drastically with the removal of these two points. Therfore, they will remain included in the final model.

```
##
## Call:
## lm(formula = Y^(1/10) ~ X1, data = dat_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32736 -0.04893 -0.00249  0.03973  0.30336
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.238e+00  2.757e-03   448.9   <2e-16 ***
## X1          -4.484e-03  5.763e-05   -77.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08541 on 2214 degrees of freedom
## Multiple R-squared:  0.7322, Adjusted R-squared:  0.7321
## F-statistic:  6053 on 1 and 2214 DF,  p-value: < 2.2e-16
```

Figure 10 shows the Studentized deleted residuals plotted against the expected values to

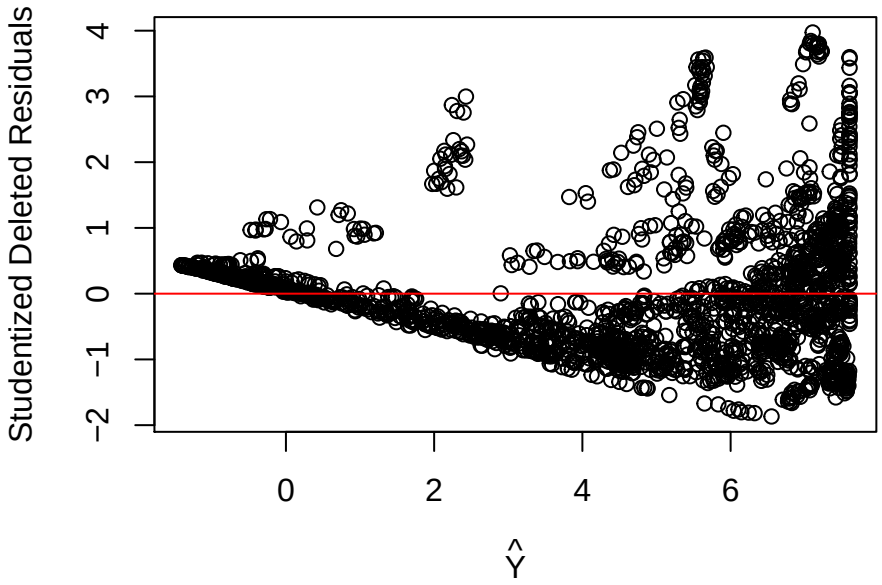


Figure 10: Studentized Residuals vs. Predicted Pollution Values

validate the normality assumption. We can see that the residuals are heavily fanned. This suggests a transformation is in order.

Similarly, the normality assumption of residuals is explored in Figure **??** shows that the assumption is violated, again suggesting a transformation. In addition, we use the Shapiro-Wilks test to formally determine normality. With a p-value < 0.05, we reject the hypothesis that the data are normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit_untrans)
## W = 0.90811, p-value < 2.2e-16
```

To determine the proper transformation, the boxcox() function in R is used. The results are shown in Figure 11. With the highest likelihood of obtaining the lowest SSE coming at a λ value of approximately 0.1, we will transform the response variable, Y (Pollution), by raising it to the 0.1 power.

With this transformation, we now observe the studentized residual plot and QQ-Plot in Figure 12 Figure 13, respectively. This transformation appears to provide a much more constant, normal distribution of residuals. However, with the Shapiro-Wilks test we again reject the hypothesis that the residuals are normal.

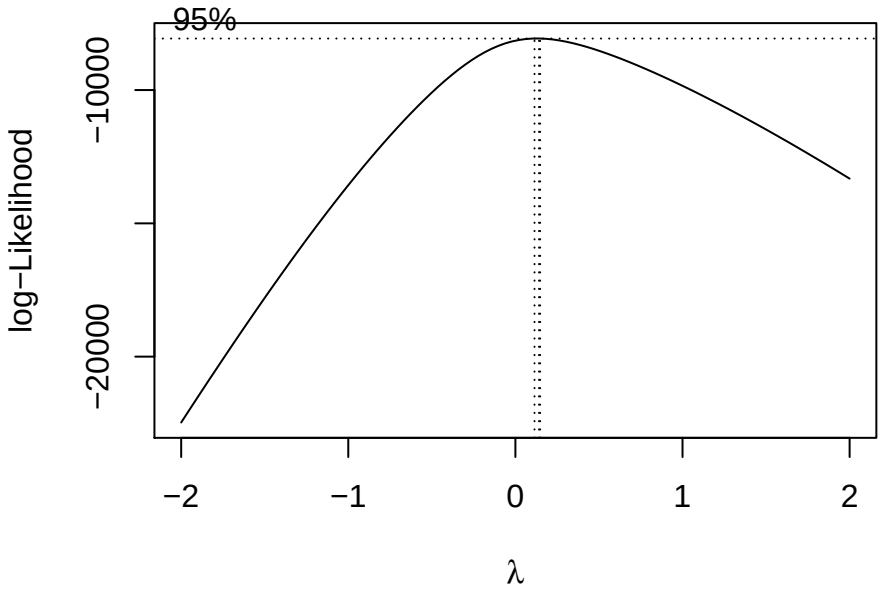


Figure 11: Box Cox Results

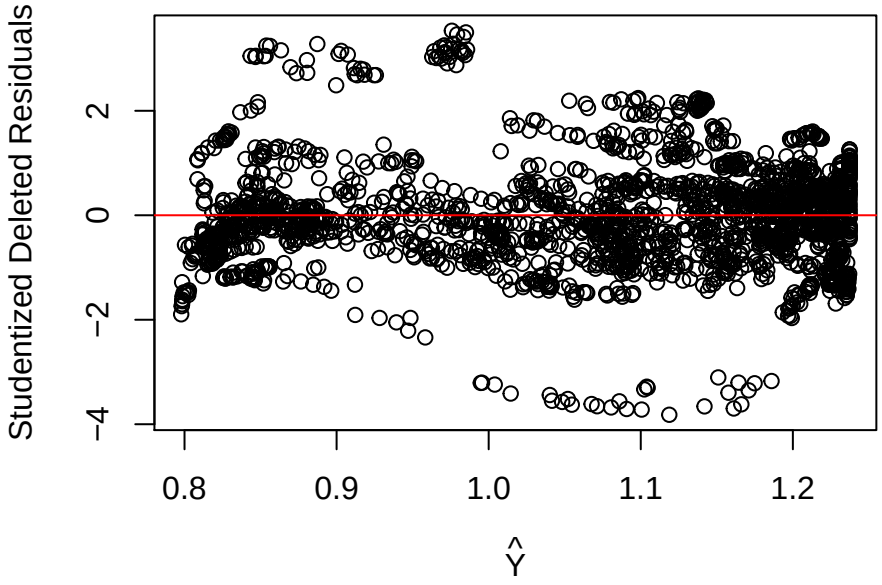


Figure 12: Studentized Residual Plot for Transformed Model

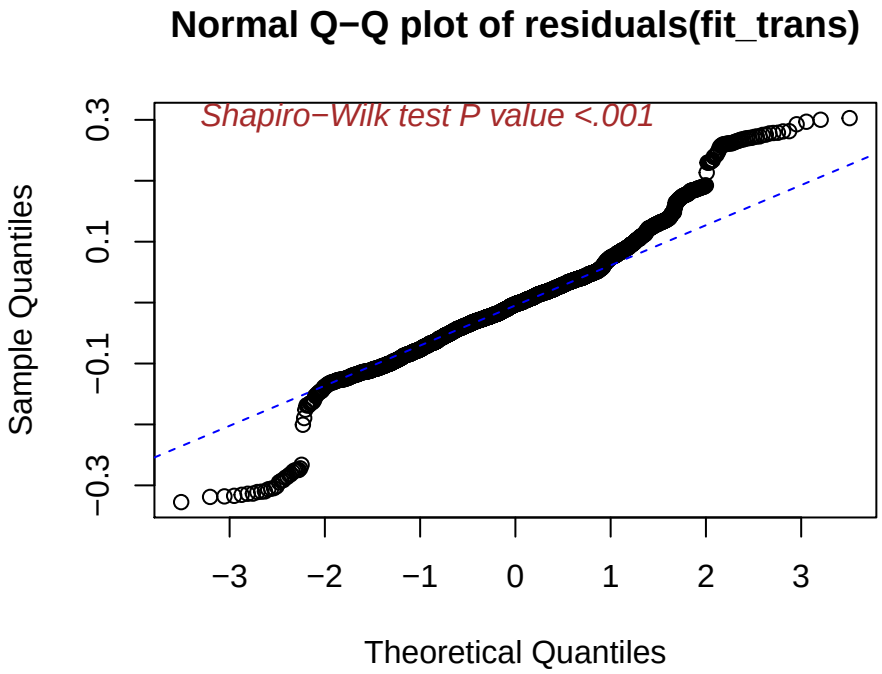


Figure 13: Q-Q Plot for Residuals of Transformed Linear Model