**The significance of climatic factors in the site selection for large polluting enterprises(manufacturers)**

6214 Project

Professor: Yuanzhang Li

Jiayi Zheng, Jiaxi Jiang, Xueqi Zhou

Due Date: 2019/12/14

Table of Contents

[Abstract 3](#_Toc26748795)

[Introduction 3](#_Toc26748796)

[Analysis —Linear Regression 4](#_Toc26748797)

[Step 1: Check collinearity 4](#_Toc26748798)

[Step 2: Check linear relationship 5](#_Toc26748799)

[Step 3: Variable first selection 5](#_Toc26748800)

[Step 4: Diagnose the first model 7](#_Toc26748801)

[(1). Test the normality of the residuals 7](#_Toc26748802)

[(2). Test whether the mean of the residuals equal to 0 7](#_Toc26748803)

[(3). Test whether the residuals have autocorrelation 7](#_Toc26748804)

[(4). Test whether the residuals hold constant variance 8](#_Toc26748805)

[(5). Influential point 8](#_Toc26748806)

[Step 5: Improving model 9](#_Toc26748807)

[Conclusion 11](#_Toc26748808)

[References 13](#_Toc26748809)

# Abstract

With the rapid development of society, the increasing concern on environmental health has become a hotbed for discussion on air pollution, especially on sulfur-dioxide content, since 99% sulfur-dioxide comes from human sources. The main goal for the project is providing suggestions on site selection for potential large air-polluting manufacturers by taking climatic elements, specifically wind-speed and precipitation per year, into consideration to minimize human influence on air condition. In our project, we are going to find the relationship between climatic variables and sulfur dioxide content in the air by constructing a regression analysis. There are 41 observations involved in analysis and we choose Temp, Man, Wind and Rain and Population as our predictor variables. The conclusion should be the list of cities that can minimize the human influence on sulfur dioxide content in the air.

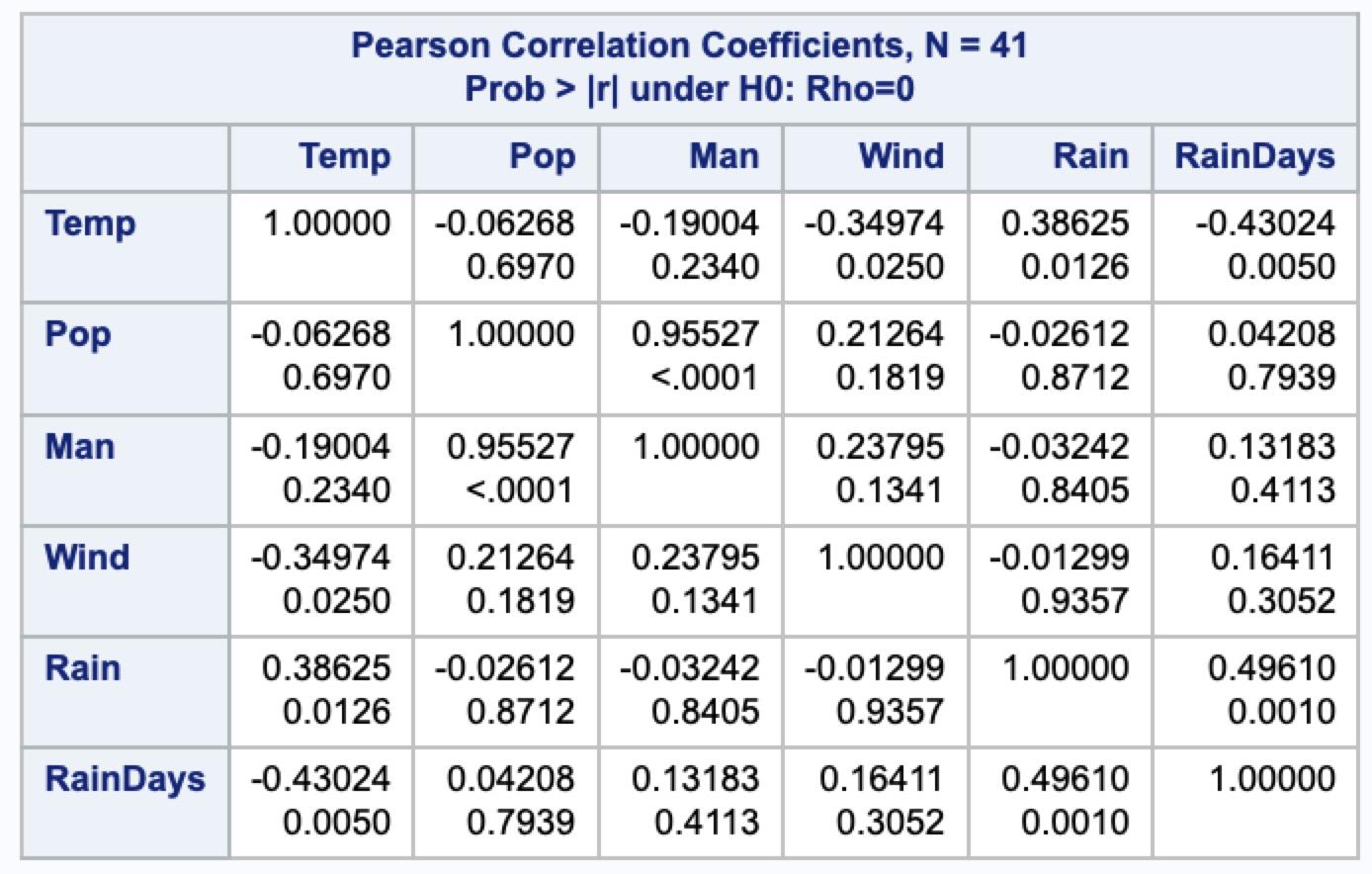
# Introduction

Nowadays, air pollution has already drawn a lot of attentions from different countries, since it will not only cause environmental issues, but also be hazardous for human health. The World Health Organization estimates that 7 million people die in 2012 because of air pollution (WHO, 2014), which means one person dies for every five seconds. Although there are many various kinds of gases contributing to the air pollution, we will focus on sulfur dioxide in this article, which is mainly produced by human activity. The main sources of sulfur dioxide are fossil fuel combustion and natural volcanic activity. High concentrations of sulfur dioxide can cause irritation of the respiratory system and inflammation, and even destroy lung function. People define 1.0-3.0 ppm as unhealthy level for SO2. It is significant to figure out potential factors on sulfur dioxide concentration. The purpose of this study is to determine the possible relationship between climatic elements, wind and precipitation, and sulfur dioxide concentration, and to find how these variables affect the sulfur dioxide concentration. This article focuses on linear regression and contains the process of deriving relatively adequate model.

# Analysis —Linear Regression

## Step 1: Check collinearity

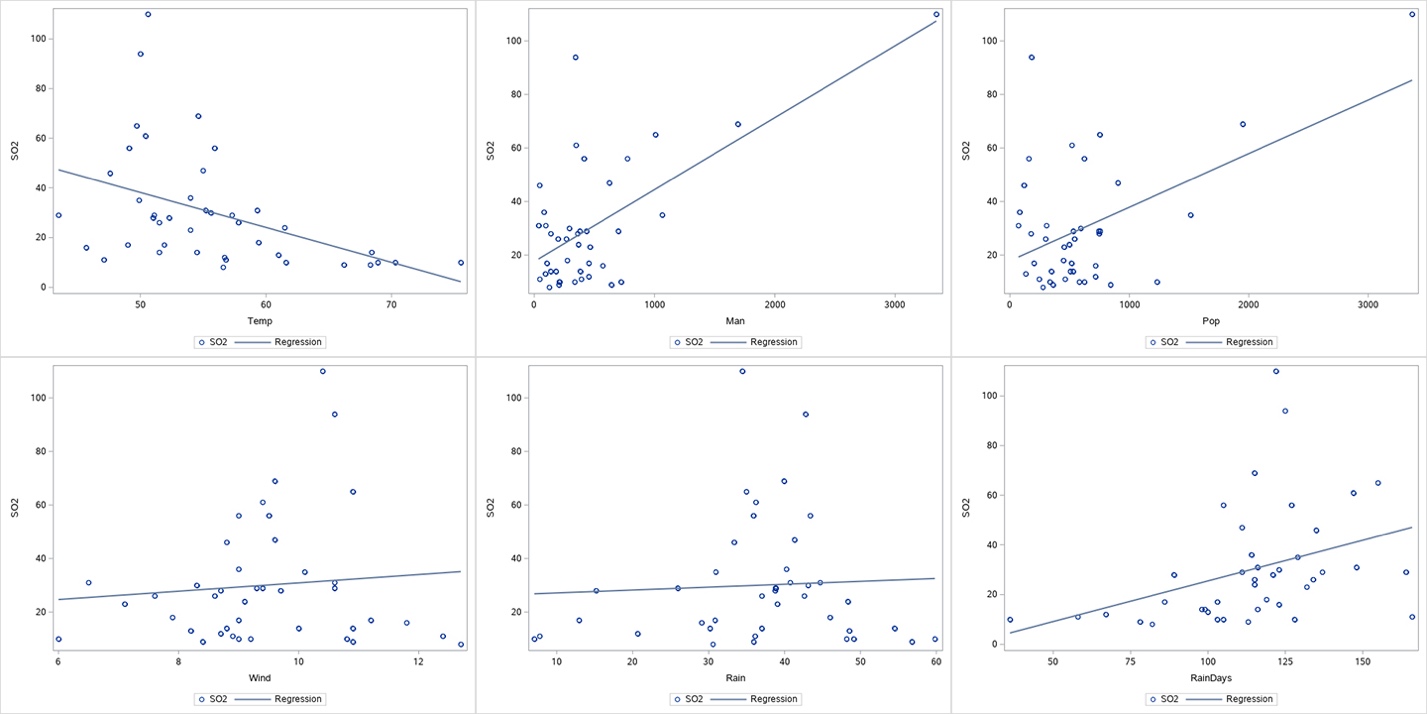
Instead of doing the variable selection, we want to know whether there exists collinearity between predictors. Table 1 shows the Pearson collinearity coefficient table. We can tell that the correlation coefficient between Man and Pop is 0.95527, which indicates the strong collinearity issue between Pop and Man.



(Table 1)

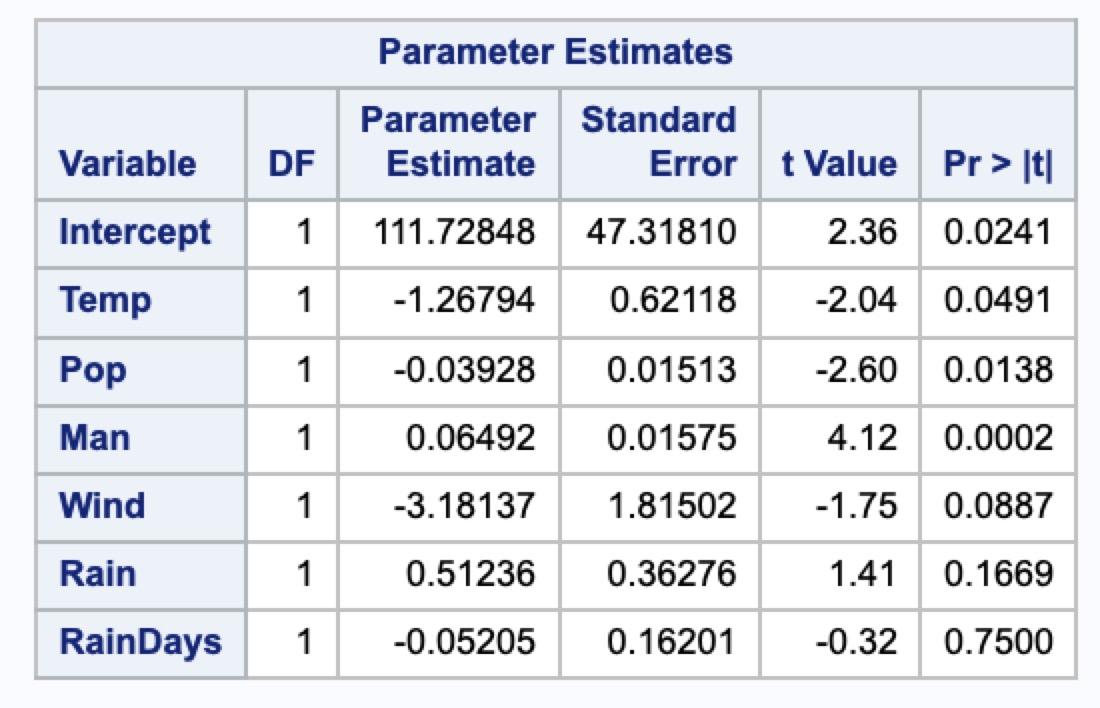
## Step 2: Check linear relationship

We can see from the graph 1 below, which shows that the response variable - SO2 and six predictors have linear relationships. Also, the R-square of full model is 0.6695, which supports the linearity.



(Graph 1)

## Step 3: Variable first selection

****

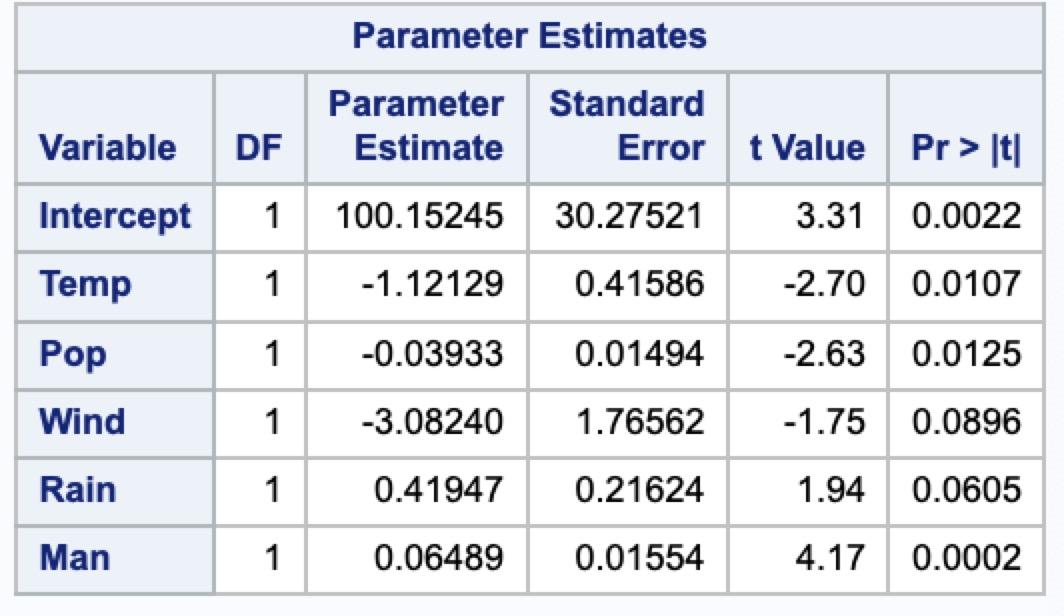
(Table 2)

From the table 2, we can tell that some variables are insignificant. We will consider deleting some variables by variable selection method. We want to get a relatively good linear model, and we named it Model 1. By using r-square procedure in SAS, model fit for full and reduced models will be calculated according to various predictors. The result is partially shown in table 2. By using (CP, PRESS, AIC, BIC and R-square) as criteria, we find the relatively best model, which contains predictors Temp, Pop, Wind, Rain and Man.

Model 1:

****

(Table 3)

****

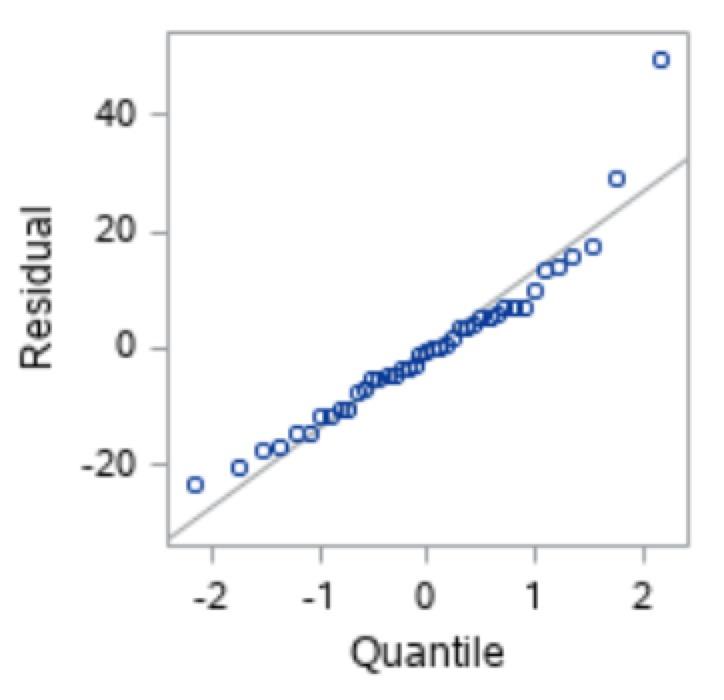
(Table 4)

The table 4 shows the regression coefficient for model 1. We can see that all of the p-values are either close to 0.05 or smaller than 0.05, which means the parameters are all significant relatively.

## Step 4: Diagnose the first model

### (1). Test the normality of the residuals

The QQ plot will give us a basic idea on normality in graph 2. The quantile plot shows that although there may exist a few outliers, the residuals almost obeys the normality.

****

(Graph 2)

### (2). Test whether the mean of the residuals equal to 0

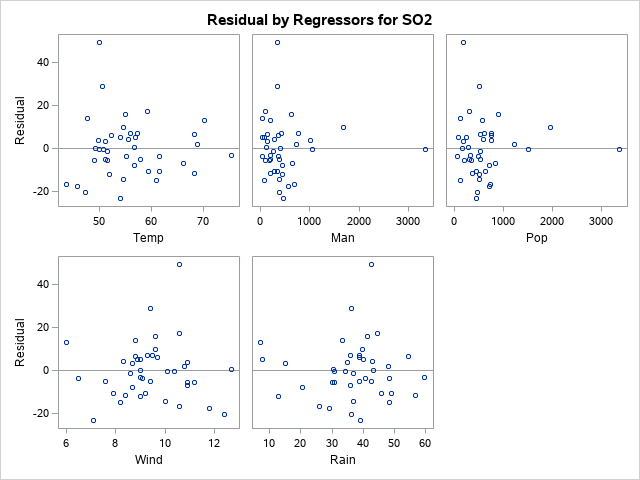
This assumption need not be satisfied, because even though the means are not equal to 0, the part that is more or less than 0 is a constant, which can be incorporated into the intercept term without affecting other assumptions of the model.

### (3). Test whether the residuals have autocorrelation

The DW-value is 1.519, showing that the autocorrelation doesn’t exist.

### (4). Test whether the residuals hold constant variance

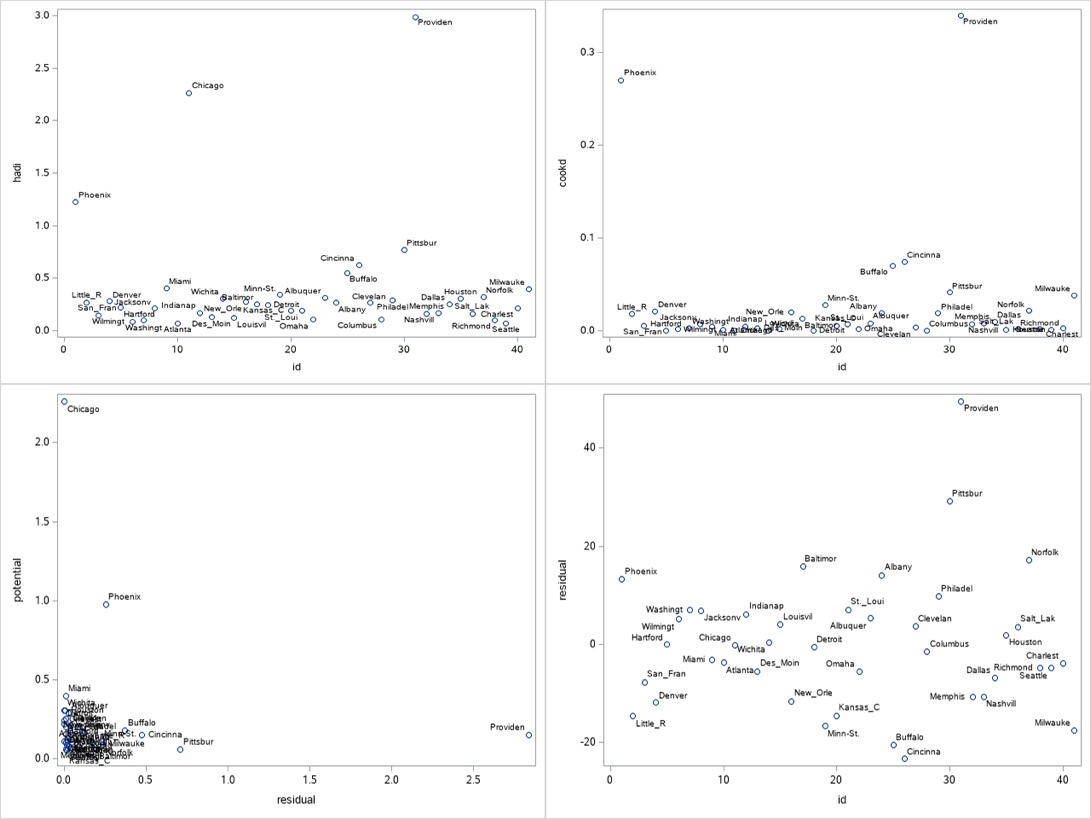
We made the residuals-predictors plots to see whether there exists heteroscedasticity. It is clear that there do have heteroscedasticity, nearly every variable increase or decrease as the residual increases or decreases. Why there will have heteroscedasticity? 1.It may come from the measure error, but we assume that there has no measure error, because we cannot re-collect data and we do not know if the data collection process is reasonable. 2.It may cause of non-linear relationship between X’s and Y, and we may use transformation or introducing dummy variables to solve the problem. 3.It may cause of some influence points (outliers, high-leverage points). We can try those methods later.

****

(Graph 3)

### (5). Influential point

In a fitting model, we want to ensure that the fit is not overly determined by a few observations. Hence, we are interested in detecting points whose deletion might cause large changes. In this case, we consider hadi’s measure (plus function), cook’s distance (product function), L-R plot (comprehensive), and residuals-index plot to detect outliers, high-leverage points, and high-influential points. The four plots are shown in graph 4.

****

(Graph 4)

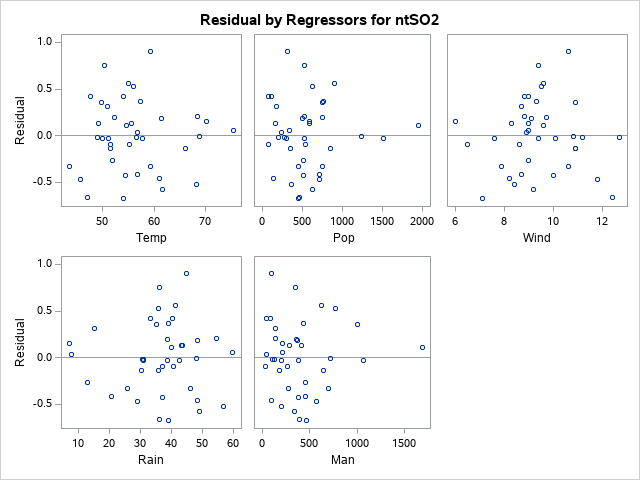
From the graph above: we can detect that Chicago has high leverage, and high influential. This city has high population and many manufacturers, which makes Chicago an influential observation. Providence is an outlier, which is a small city with heavy industry. Therefore, their characteristics are different with most other cities, then we delete those two points.

## Step 5: Improving model

#### (1). Stabilizing Variance

Using Box-Cox method, we obtained the best lambda value equaling -0.25, which can minimize the variance of residuals and make it become constant. So that this model become:

Model 2:

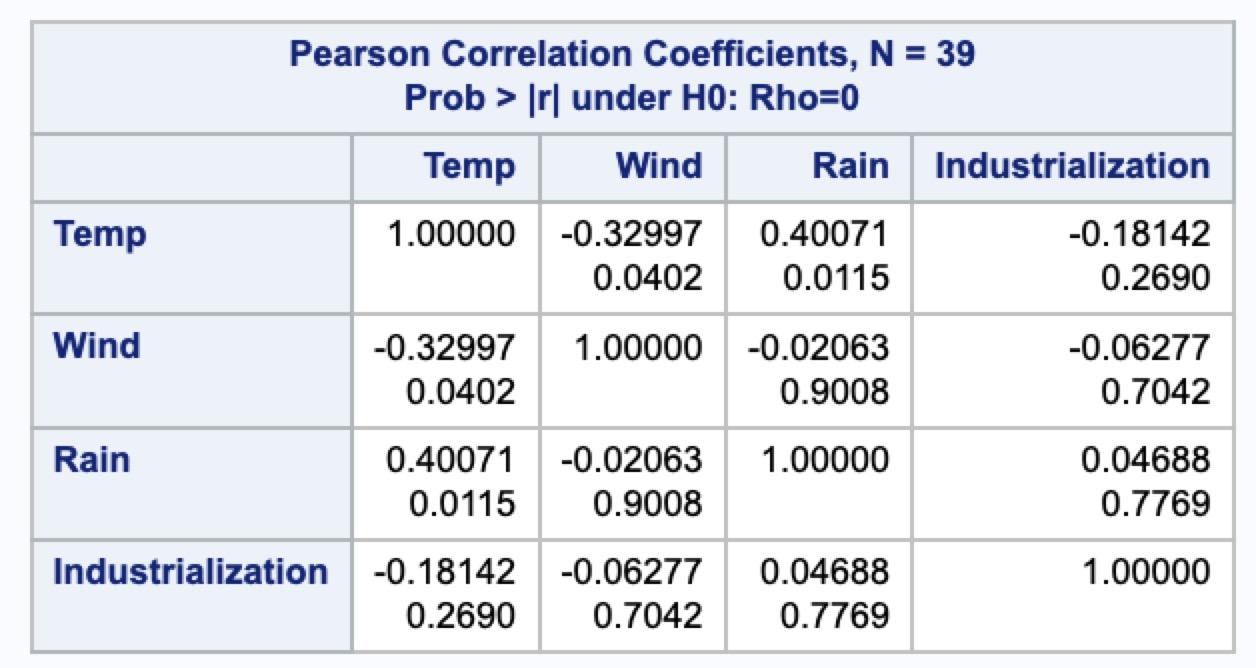
****

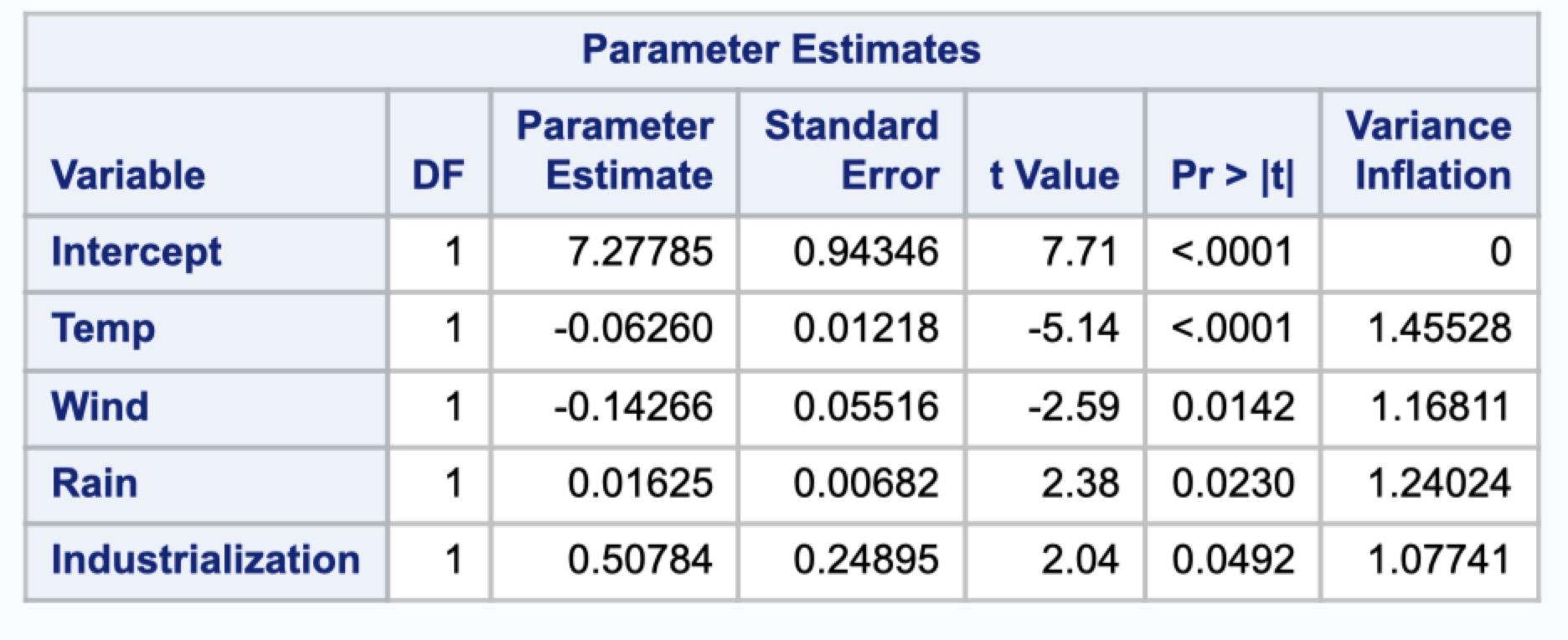
(Graph 5)

The graph 5 shows the plot between residual and predictors, after the transformation. The problem of heteroscedastic have been relieved.

#### (2). Eliminate collinearity

There are four reasons for the dummy variable here: there is collinearity between man and pop; we want to keep the model as simple as possible but use as many variables as well; pop do not follow strict linear relationship with ntso2; the parameter coefficients of man and pop are relatively small. So that we decide to introduce dummy variable industrialization. Z = man/pop: when z > 1, industrialization=1. When z <1, industrialization = 0. After we did this, we found that the collinearity has been eliminated, and the coefficient become larger.

****

****

(Table 5)

Therefore, our final model is:

# Conclusion

In this article, we used linear regression to determine how the climatic factors influence sulfur dioxide contents in the air. We find that three variables contribute the most. They are Temp, Wind, and Rain. Since they have the exponential relationship, the sulfur dioxide content will change faster as the variables changed constantly.

Our final model will be:

The R-square for this model is 0.5375. From this model, it will be enough for us to deduce the list of best cities to develop the polluting industry, based on the assumptions that the ratio of local people’s occupation will not be changed. Cities like Hartford, Charleston and Albany will be the best places to develop polluting industry as their natural environment can minimize human influence, while cities as Houston, Phoenix, and Dallas.

# References

World Health Organization. (2014). “*WHO | 7 Million Premature Deaths Annually Linked to Air Pollution.*” Retrieved from https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/