Multivariate Analysis Project Report: Analysis of Mortality Rate

Enes Dilber, Selcuk Meric Kostekci, Nilsu Uyar

Abstract

In our analysis, we performed Multiple Linear Regression (section 2), Multivariate Multiple Linear Regression (section 3), Principal Component Analysis (section 4 and 5), Canonical Correlation (section 6), Clustering Methods (section 7), Classification Methods (section 8). We used linear regression ($Y = \beta X$) in all continuous response models and estimated coefficients by ordinary least squares method. We gave brief introduction about the concepts that we used, such as Ordinary Least Squares method, Principal Component Analysis, Stepwise Regression, Mallow's C_p, Wilk's Λ, k-means, Decision Trees and Multi-layer Perceptron. In univariate linear models, we seek model for B: Death Rate (Mortality Rate) variable. In multivariate models, we seek model for A13: relative pollution potential of nitric oxides and A14: relative pollution of sulfur dioxides. In Cluster Analysis, we applied k-Means and Aggloromative Hierarchical Clustering. For classification, we compared: Softmax Regression, Decision Tree, Multilayer Perceptron and k-Nearest Neighbours. We used R programming Language. Project progress and R codes are reachable at Github page of the project[1]. In addition to R, we used nstats[2] for linear regression. It is a R Shiny application programmed by our team.

1 Introduction

The morality rate (Death Rate) of US cities differ dramatically. The question is "what is the cause of it". McDonald et al. [3] introduce several variables that contain information about pollution indices, economic factors and social life conditions. By the context of their research, they constructed a univariate regression model that took *Death Rate* as response and investigate the effect of other factors. In their analysis, they mostly focus on transformation of independent variables and variable selection via step-wise regression.

1.1 Data Information

There are 16 variables and 60 observations. Here are the descriptions about dataset.

- A1 average annual precipitation in inches
- A2 average January temperature in degrees Fahrenheit
- A3 average July temperature in degrees Fahrenheit
- A4 percent of population 65 years old or older
- A5 average household size
- A6 the number of years of school that persons 22 years or older have completed
- A7 the rate of households fully equipped with household appliances
- A8 population per square mile in urbanized areas

- A9 percent nonwhite population
- A10 percent office workers
- A11 poor families (annual income under $3000)
- A12 relative pollution potential of hydrocarbons
- A13 relative pollution potential of nitric oxides
- A14 relative pollution of sulfur dioxides
- A15 percent relative humidity, annual average at 1pm
- B The death rate in some regions of United States(deaths per 100,000)

1.2 Ordinary Least Square Estimation

Unless any other models mentioned in content, it is the fact that we used OLS estimations of parameters. Here is the linear regression model:

$$Y_{n\times k} = X_{n\times p+1}\ \beta_{p+1\times k}\ + \epsilon_{n\times k}$$

$$\begin{bmatrix} y_{11} & \dots & y_{1k} \\ y_{21} & \dots & y_{2k} \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ y_{n1} & \dots & y_{nk} \end{bmatrix}_{n\times k} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{n\times p+1} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}_{p+1\times k} + \begin{bmatrix} \epsilon_1 \\ \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}_{n\times k}$$

Where n = number of observations, p = number of predictors (+1 comes from intercept term), k = number of responses. When $k = 1$ the model is univariate linear regression. Any other integer $k > 1$, the model is multivariate linear regression. However derivation of coefficients by OLS estimation is the same.

To find the OLS estimates of β we need to minimize ϵ^2

$$(\epsilon^T\epsilon)_{k\times k} = (Y - X\beta)^T_{k\times n}(Y - X\beta)_{n\times k} = Y^TY - 2\beta^TX^TY + \beta^TX^T\beta X$$

We want to minimize sum of square errors(this is the main idea of L2-loss).

$$\min_{\beta}(\epsilon^T\epsilon)$$

This is a pretty straightforward optimization problem. Because we know sum of square errors have a strictly convex shape.

$$\frac{\partial\epsilon^T\epsilon}{\partial\beta} = -2X^TY + 2X^TX\beta$$

We estimate β by equating above equation zero. Because of strict convexity, it is going to give us the values of β which gave minimum $\epsilon^T\epsilon$.

$$-2X^TY + 2X^TX\hat{\beta} = 0$$
$$\hat{\beta} = (X^TX)^{-1}X^TY$$

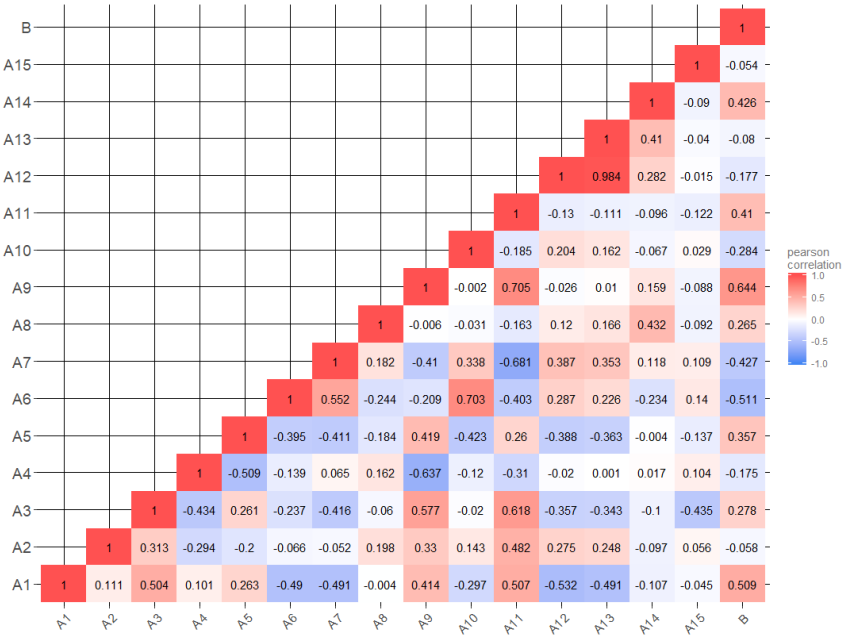
2 Multiple Linear Regression

In this section, we are going to investigate, "What are the variables behind the variability of *Death Rate*. To do that, we seek linear models such that their coefficients are estimated by ordinary least square method.

2.1 Relationship between Variables

Firstly, we plot the pairwise correlation matrix of the data at Figure 1, to see the correlation between variables. In this matrix, the variables which are more correlated with each other are shown in darker red colour if they have positive correlation and they are shown in darker blue if they have negative correlation. As one can see, some independent variables have strong correlation that can ruin our analysis by multi-collinearity. A4 and A9, A6 and A10, A3 and A11, A7 and A11; have a pearson correlation coefficient that is greater than 0.6. More importantly, A12 and A13 are perfectly correlated. To get more sense from the relationship between predictors and response we obtained scatter plot at Figure 2. We could see A12 and A13 have accumulated values around zero with some outliers. That probably will effect our analysis. A1, A6, A7 and A9 have a linear relationship with our response.

Figure 1: Correlation Matrix



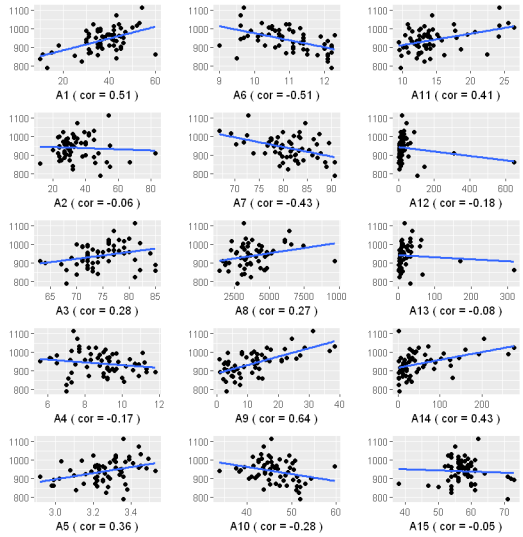
To obtain linearity condition in "Linear Regression" we performed several transformation on independent variables at Table 1. The table numbers indicate power transformation(x^k), m is the mean of the variable. Some variables strongly need the transformation. We performed suggested transformations to $A2, A13, A15$ that are indicated bold. In fact, those transformations made the model worse. In our analysis we concluded that Non-Linear transformation is not necessary. In addition, we did not obtain any significant interaction term.

2.2 Initial Linear Model for Death Rate

We construct the model with and without standardized predictors.
Here are the regression models:

- Without Standardization ($Y \sim X$):
  $Death\ Rate = 1860 + 2.07(A1) - 2.18(A2) - 2.83(A3) - 14(A4) - 115(A5) - 24.2(A6) - 1.15(A7) + 0.01(A8) + 3.53(A9) + 0.523(A10) + 0.267(A11) - 0.889(A12) + 1.87x(A13) - 0.0345(A14) + 0.533(A15)$

Figure 2: Scatter Plot of Death Rate vs Rest of the Variables



f(x)	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
-2	-0.46	-0.14	-0.31	0.14	-0.35	0.47	0.43	-0.28	-0.09	0.24	-0.39	-0.23	-0.33	-0.15	-0.06
-1	-0.5	-0.08	-0.3	0.15	-0.35	0.49	0.43	-0.29	-0.28	0.26	-0.4	-0.3	-0.4	-0.21	-0.03
-0.5	-0.51	-0.04	-0.29	0.16	-0.35	0.49	0.43	-0.29	-0.43	0.27	-0.41	-0.3	**-0.39**	-0.29	-0.01
0.5	0.52	-0.03	0.28	-0.17	0.36	-0.51	-0.43	0.28	0.62	-0.28	0.41	-0.07	0.08	0.45	-0.03
1	0.51	-0.06	0.28	-0.17	0.36	-0.51	-0.43	0.27	0.64	-0.28	0.41	-0.18	-0.08	0.43	-0.05
2	0.47	-0.1	0.27	-0.18	0.36	-0.52	-0.42	0.22	0.6	-0.29	0.41	-0.19	-0.17	0.37	-0.09
3	0.41	-0.12	0.25	-0.19	0.36	-0.53	-0.42	0.15	0.53	-0.29	0.4	-0.17	-0.17	0.33	-0.12
4	0.35	-0.12	0.24	-0.19	0.36	-0.54	-0.42	0.1	0.47	-0.29	0.4	-0.17	-0.17	0.3	-0.15
log	0.52	0.01	0.29	-0.16	0.36	-0.5	-0.43	0.29	0.55	-0.27	0.41	0.15	0.28	0.4	-0.01
exp	-0.17	-0.06	-0.18	-0.17	0.36	-0.55	-0.37	NaN	0.21	0.06	0.26	-0.17	-0.17	0.18	-0.17
$(x-m)^2$	-0.35	**-0.18**	-0.24	-0.1	-0.06	-0.14	0.14	0.03	0.4	-0.11	0.32	-0.19	-0.18	0.3	**-0.35**
$(x-m)^3$	0.44	-0.09	0.07	-0.12	0.28	-0.31	-0.43	0.05	0.45	-0.06	0.39	-0.17	-0.17	0.31	-0.02

Table 1: Correlation between Death Rate and Transformed Predictors

- With Standardization ($Y \sim standardization(X)$):
  $Death\ Rate = 940 + 20.7(A1) - 26.1(A2) - 13.5(A3) - 20.6(A4) - 15.6(A5) - 20.5(A6) - 5.89(A7) + 14.6(A8) + 31.5(A9) + 2.42(A10) + 1.11(A11) - 81.8x(A12) + 86.5(A13) - 2.19(A14) + 2.91(A15)$

To compare, we will refer standardized coefficients. A13 has a huge impact on response. However we know from Figure 2 that, they are not linearly related to each other. So it is probably an insignificant variable that effects our model. Very dangerous to comment on a variable which is not significant but has a higher effect on response. We will make more comment after significance test. In later findings and comments, we are going to always mention standardized model as our model.

2.2.1 R^2 values:

- $R^2 = 0.7985$, $R^2_{adj} = 0.7298$, $R^2_{prediction} = 0.5215$

We know R^2 is an indicator for "How good we model variation in response with variation in predictors. However there is a huge gap between R^2 and R^2_{adj}. That is probably cause by some insignificant variables. In addition $R^2_{prediction}$ value is much lower than R^2_{adj}, this is probably because of some insignificant variables

and outlier observations. $R^2_{prediction} = PRESS/SST$. PRESS has basically same intuition between Leave-one-out Cross Validation. So low press score indicates we have some influention and outlier observation.

2.3 Assumption and Anomaly Check

In linear models, there may some problems like missing important predictors; having nonlinear relation, multicollinearity, outliers, influential observations etc. Furthermore, there are some assumptions. There must be a linear relationship between response and predictors. Also, error must have zero mean, constant variance, uncorrelated and normally distributed. In this section, we will focus on these issues.

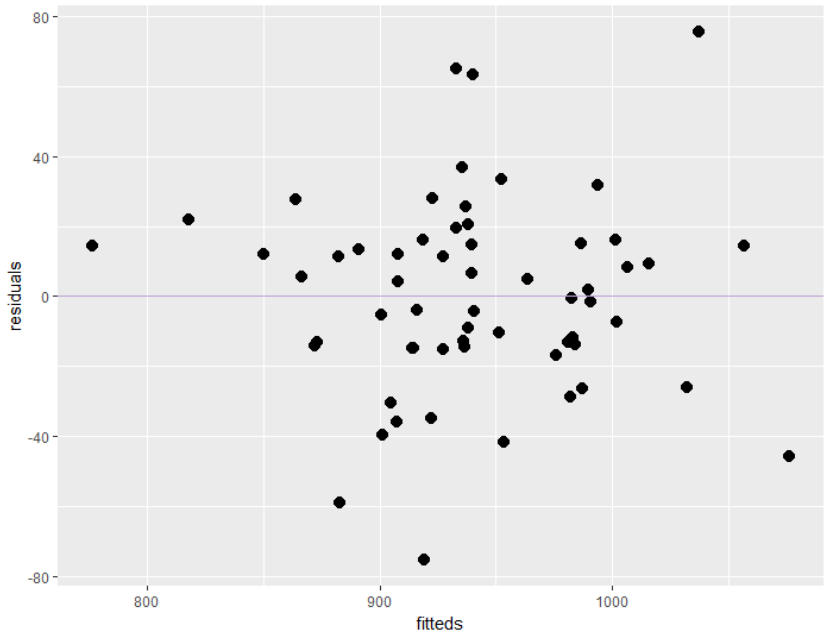
2.3.1 Linear relation between X and Y

We gave sufficient explanation about relationship between predictors and response. It is quite week for majority of variables as seen on Figure 1 and Figure 2.

2.3.2 Zero Mean and Heteroscedasticity of Variance of error terms

The error term of our model has a mean zero value. Here is Heteroscedasticity analysis. We performed *Breusch-Pagan test* [4]. The p *value* $= 0.41$ of H_0: *Error variance do not change with the level of the response.* We confidently conclude that, we do not have a non constant variance problem. We also obtain *Fitted vs Residual* plot at Figure 3. There is no association.

Figure 3: Fitted vs Residuals Plot



2.3.3 Autocorrelation

There is no autocorrelation problem either. According to Durbin-Watson test [5] test score is 2.32 and p *value* $= 0.28$ with H_0: *Errors are serially uncorrelated.*

2.3.4 Multicollinearity

To check multicollinearity, we obtained VIF values. $A12$ and $A13$ have very large values, which are far larger than 10. This is not a surprise because they are strongly correlated. Their correlation is 0.98.

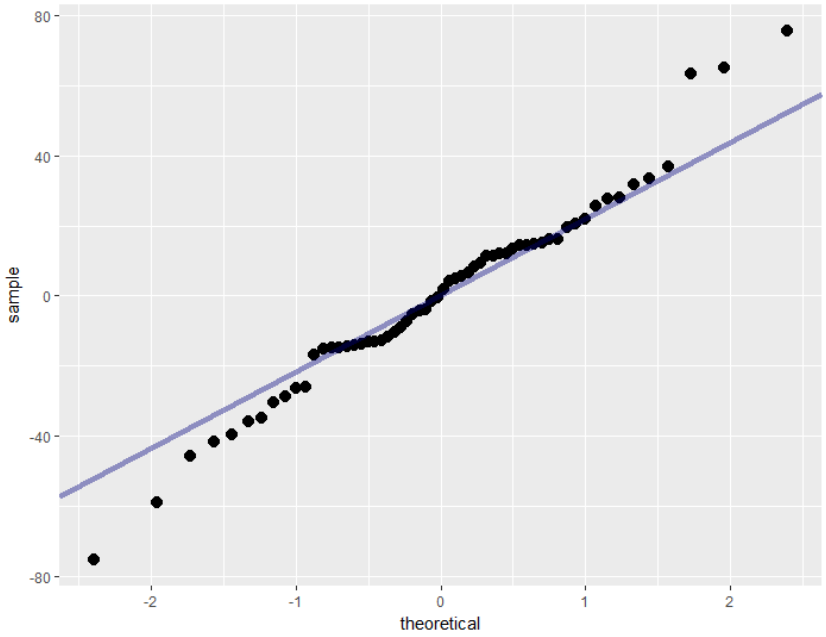
Variables	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
VIF	3.99	3.69	4.02	7.26	3.97	5.07	3.21	2.03	7.38	2.91	6.43	**97.70**	**106.00**	4.59	1.86

Table 2: Variance Inflation Factors

2.3.5 Normality

To check normality, we obtained QQ-plot(Figure 4) and Shapiro-Wilk test [6]. $p\ value = 0.28$ with H_0: *Residuals are normally distributed.* We do not enough evidence to reject H_0. No problem with normality.

Figure 4: QQ-plot for Normality of Error terms



2.3.6 Leverage, Outlier and Influential Point Detection

Observations	cooks.d	leverage	res	del.res	stu.res
29	0.871	0.905	12.0	127.0	1.22
48	0.395	0.601	-41.8	-105.0	-2.13
32	0.357	0.602	-39.7	-99.6	-2.01
37	0.185	0.281	75.6	105.0	3.00
28	0.138	0.237	-75.3	-98.7	-2.88

Table 3: Unusual Observations

If we down-weight observation $32, 37, 28$ we could increase R^2 from 0.8 to 0.83 and increase R^2_{pred} from 0.52 to 0.69. The increase in R^2_{pred} quite significant.

2.4 Subset Selection

2.4.1 Without $A12$

At multicollinearity section we found that $A12$ and $A13$ have a perfect linear relation. Therefore we deleted $A12$ to solve this problem. When $A13$ is in the model $A12$ does not introduce a new information. Here is the model without $A12$:
$Death\ Rate = 940 + 22.8(A1) - 24.7(A2) - 13(A3) - 16.2(A4) - 13.5(A5) - 15.3(A6) - 8.88(A7) + 14.1(A8) + 32.6(A9) - 0.579(A10) + 0.971(A11) + 2.78(A13) + 10.7(A14) + 1.77(A15)$
R^2 values:

- $R^2 = 0.7808$, $R^2_{adj} = 0.7126$, $R^2_{prediction} = 0.6001$

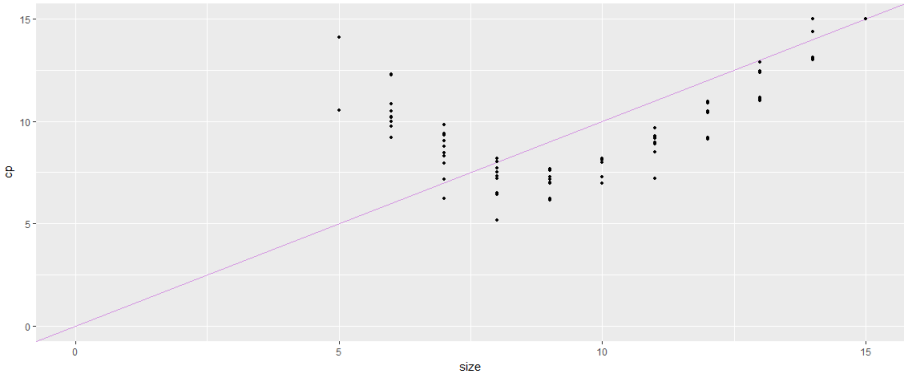
2.4.2 Mallow's C_p

We run Mallow's C_p algorithm to find best subsets of our model. Here is the formula of Mallow's C_p[7]:

$$C_p = \frac{SSE_p}{MSE} - N + 2P$$

SSE is the Sum of Square Errors of subset model. N is number of observations and P is number of variables in subset. To selection on C_p is: Take the model which has few variables, closer and under the diagonal line. At Figure 5, one can see the subset variation.

Figure 5: Mallow's C_p



Mallow's C_p suggested Model with 6 variables:
$Death\ Rate = 940 + 13.6(A1) - 20.5(A2) - 11.7(A3) - 15.9(A6) + 16.3(A8) + 44.7(A9)$ with $R^2 = 0.7808$, $R^2_{adj} = 0.7126$, $R^2_{prediction} = 0.6001$

Mallow's C_p suggested Model with 7 variables:
$Death\ Rate = 940 + 12.6(A1) - 18.9(A2) + 2.89(A4) - 12.8(A6) + 11.9(A8) + 38.4(A9) + 11.7(A14)$ with $R^2 = 0.7466$, $R^2_{adj} = 0.7124$, $R^2_{prediction} = 0.6525$

2.4.3 Step-wise Regression

Step-wise regression is another popular tool for subset selection. We performed a Forward Selection. Which starts the model with the independent variable which has strongest correlation with dependent, and add other variables with order of importance.

Forward Selection with alpha-to-enter value = 0.25
$Death\ Rate = 940 + 16.8(A1) - 17.7(A2) - 10.2(A3) - 13.3(A6) + 11.7(A8) + 40.4(A9) + 10.9(A14)$ with $R^2 = 0.7605$, $R^2_{adj} = 0.7283$, $R^2_{prediction} = 0.6774$

2.5 Final Linear Model for Death Rate

We obtain our "best" subset by step-wise regression.

$$Death\ Rate = 940 + 16.8(A1) - 17.7(A2) - 10.2(A3) - 13.3(A6) + 11.7(A8) + 40.4(A9) + 10.9(A14)$$

The most important variables in initial model are not in final model. This indicates how bad to give conclusion without checking assumptions. For this model we could now interpret the effect of independent variables. In our model, A1, A8, A9, A14 have positive effect on *Death Rate*. Which is quite interesting because this indicates; precipitation, population per square mile in urbanized areas, percent nonwhite population, relative pollution of sulfur dioxides have a significant effect on the response. Especially in the change on A9:percent nonwhite population is effecting that rate significantly. On the other hand A2, A3 and A6 effect *Death Rate* negatively.

2.5.1 Assumption Check

All linear model assumptions are satisfied. To write them in detail here is waste of space. Therefore we give a brief summary of assumption check.

- All VIF scores are less than 2
- Durbin-Watson score is 2.23
- Shapiro-Wilk p-value is 0.19
- Non-constant variance p-value = 0.45

2.5.2 ANOVA

H_0 : all coefficients are equal to zero
$p\ value = 2.3 \times 10^{-13}$, at 0.01 significance level reject H_0

2.5.3 Significance of Variables

All variables are significant at 0.1 significance level. Only A3 is not significance at 0.05 level. However when we discard it model performs worse. We rather decided to keep it.

coefficients	estimate	standard error	t value	p value
(Intercept)	940.40	4.19	224.60	0.00
A1	16.79	5.76	2.91	0.01
A2	-17.74	4.87	-3.65	0.00
A3	-10.22	5.71	-1.79	0.08
A6	-13.29	5.22	-2.55	0.01
A8	11.75	5.05	2.33	0.02
A9	40.39	5.75	7.02	0.00
A14	10.93	5.31	2.06	0.04

Table 4: t - test

2.6 Discussion

In multiple linear regression, we firstly look relationship between variables. Variable A12:*relative pollution potential of hydrocarbons* and A13:*relative pollution potential of nitric oxides* are strongly correlated. Later, we tried to find transformation to get better linearity condition but in fact transformation did it worse. Later, we construct the model with standardized predictors to make comparable interpretation. After that, we check linear regression assumptions. There was no problem in terms of errors and having a linear relation

between response and predictors. However, as we mentioned before, there is a perfect relation between A12 and A13 that cause a multicollinearity problem. To overcome this problem, we drop A12. We are going to more technical information about "Why we drop A12" at *Multivariate Multiple Linear Regression* part. Lastly, to find best model we run Mallow's Cp and stepweise regression. Final model shows that Death Rate can be explained by average annual precipitation, average January temperature, average July temperature, the number of years of school that persons 22 years or older have completed, population per square mile in urbanized areas, percent nonwhite population and relative pollution of sulfur dioxides. Death Rate is increasing in regions where there are many nonwhite people. This may happen due to unfavorable working and living conditions of the non-white people in the research area.

3 Multivariate Multiple Linear Regression

In our multivariate response analysis, our research purpose is to understand the reasons behind air pollution. To achieve this goal, pollution potential of *Nitric Oxide*, pollution potential of *Sulfur Dioxides* and pollution potential of *Hydrocarbons* are selected as dependent variables, since these substances are chemical compounds, any increase of their presence in the air may lead to critical air pollution. Rest of the variables are selected for independent variables except *Death Rate*. We believe, weather conditions and human factors have crucial effects for the density of the toxic gases such as Nitric Oxide and Sulfur Dioxides. Meanwhile *Death Rate* may increase by the toxic gases in atmosphere(reverse causality). Initial analysis revealed that pollution potential of *Hydrocarbons* are highly correlated with pollution potential of *Oxides of Nitrogen* with a *pearson correlation coefficient* of 0.98. This is not a surprising fact though. Car exhaust systems and House heaters have catalytic converters. *Carbon Monoxide* goes into reaction with *Nitric oxide* and we get *Carbon Dioxide* as product[8]. Here is the equation this chemical equation proof of the correlation.

$$2CO + 2NO \rightarrow N_2 + 2CO_2$$

Because of that, pollution potential of *Hydrocarbons* is removed from the research. Initial analysis also showed that the responses are positively skewed and heteroscedasticity exists. Therefore, logarithmic and square root power of 4 transformations are applied to *Nitric Oxide* and *Sulfur Dioxides*, respectively. Those transformations are suggested by Box-Cox method[9], in our initial analysis. Also it is difficult to compare regression coefficients due to varying magnitudes, so predictors are standardized. All models are based on these transformed variables.

1. Standardized every independent variable
2. Applied *log* transformation to NO
3. Applied $\sqrt[4]{x}$ transformation to SO_2

3.1 Linear Model for *Relative Pollution Potential of* NO*'s* and *Relative Pollution Potential of* SO_2

As we introduce at the beginning, our responses are $A13$ and $A14$. Independent variables are from $A1$ to $A15$ (beside $A12$, $A13$ and $A14$). Here is our regression model:

$$Y = \beta_0 + \beta_1 A1 + \beta_2 A2 + ... + \beta_{15} A15 + \epsilon$$

Here is estimated coefficients and estimated responses in matrix form:

$$\underbrace{\begin{bmatrix} 1 & -0.14 & -0.65 & \dots & 0.27 \\ 1 & -0.24 & -0.99 & \dots & -0.10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0.06 & -0.57 & \dots & 0.09 \end{bmatrix}}_{X_{60\times 13}} \underbrace{\begin{bmatrix} 2.31 & 2.37 \\ -0.61 & -0.21 \\ 0.09 & -0.22 \\ -0.58 & -0.31 \\ 0.32 & 0.16 \\ -0.05 & -0.11 \\ -0.31 & -0.36 \\ 0.2 & 0.07 \\ 0.18 & 0.24 \\ 1.06 & 0.6 \\ 0.08 & 0.08 \\ -0.11 & -0.13 \\ -0.07 & -0.13 \end{bmatrix}}_{\hat{\beta}_{13\times 2}} = \underbrace{\begin{bmatrix} 2 & 2.13 \\ 2.29 & 2.71 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 2.35 & 2.38 \end{bmatrix}}_{\hat{Y}_{60\times 2}}$$

First column of β matrix corresponding for NO and the second for SO_2. First rows are intercept terms and rest continue from $A1$ to $A15$.

3.1.1 R^2 values:

- For NO $R^2 = 0.6221$, $R^2_{adj} = 0.5256$, $R^2_{prediction} = 0.2285$
- For SO_2 $R^2 = 0.6103$, $R^2_{adj} = 0.5108$, $R^2_{prediction} = 0.4215$

R^2 values are quite similar for both models. However there are significant gaps between R^2 and R^2_{adj} values, this suggests the initial model consist some insignificant variables. $R^2_{prediction}$ value for SO_2 model quite acceptable but NO model probably suffers from outlier observations.

3.1.2 Wilks Λ:

After assumption check (we did not include it, we give detailed assumption check at final model). We applied Wilk's Lambda[10]. Wilk's lambda is a useful test statistic to see the individual contribution of the variables to the model. It is a likelihood ratio test statistic.

$$\Lambda^* = \frac{|W|}{|B+W|}$$

W is the variation that is explained by the variable and B is unexplained variation. Since greater ratio leads to more significant variable. In Table 5, we test them at 0.05 p *value*. We conclude that specific variables explain the group of responses well. As we can see in the model, variables: the size of the population older than 65, the number of households with fully equipped kitchens, the number of office workers, the number of families with an income less than $3000, (coded as A4,A7,A10,A11,A15) have p-values greater that 0.05. This shows that we should expect very few contribution of these variables to the model. A valid remedy to this problem would be to remove these variables from the model and check the model performance again.

Variable	$Wilks\ \Lambda$	$p\ value$
A1	0.797	0.005
A2	0.814	0.009
A3	0.846	0.021
A4	0.971	0.513
A5	0.981	0.64
A6	0.875	0.047
A7	0.973	0.531
A8	0.868	0.038
A9	0.709	0
A10	0.99	0.786
A11	0.989	0.781
A15	0.948	0.293

Table 5: Wilks Λ

3.2 Final Linear Model for *Relative Pollution Potential of* NO*'s* and *Relative Pollution Potential of* SO_2

We deleted some insignificant variables according to *Wilk's* test. In addition we performed "subset selection" methods that we mentioned in previous chapter.

Here is our regression model:

$$Y = \beta_0 + \beta_1 A1 + \beta_2 A2 + \beta_3 A3 + \beta_5 A5 + \beta_6 A6 + \beta_8 A8 + \beta_9 A9 + \beta_{15} A15 + \epsilon$$

Here is estimated coefficients and estimated responses in matrix form:

$$\underbrace{\begin{bmatrix} 1 & -0.14 & -0.65 & \dots & 0.27 \\ 1 & -0.24 & -0.99 & \dots & -0.10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0.06 & -0.57 & \dots & 0.09 \end{bmatrix}}_{X_{60\times 9}} \underbrace{\begin{bmatrix} 2.31 & 2.37 \\ -0.49 & -0.16 \\ -0.01 & -0.29 \\ -0.71 & -0.38 \\ -0.23 & -0.2 \\ -0.21 & -0.28 \\ 0.28 & 0.31 \\ 0.85 & 0.48 \\ -0.1 & -0.14 \end{bmatrix}}_{\hat{\beta}_{9\times 2}} = \underbrace{\begin{bmatrix} 2.24 & 2.27 \\ 2.32 & 2.72 \\ . & . \\ . & . \\ . & . \\ 2.27 & 2.38 \end{bmatrix}}_{\hat{Y}_{60\times 2}}$$

A1, A2, A3, A5, A6 and A15 have negative effect on both indication of pollution. On the contrary, A8 and A9 have positive effect.

- A1, A2, A3 and A15 are weather related variables: The biggest negative coefficient is A3 (Average July temperature) has a negative effect on pollution. As we mentioned before these pollutants are mostly coming from catalyst reactions, in warm climates the average pollution caused by to heat houses is lower compared to cold climates.
- The biggest positive coefficient is A9 (Percent nonwhite population). This is rather an indirect effect of factory regions. For example, imagine Detroit region, there are a lot of factories. The pollutants in air is at a life-threatening level. And factory workers mostly non-white such as African-Americans and Hispanics.

3.2.1 R^2 values:

- For NO $R^2 = 0.5966$, $R^2_{adj} = 0.5333$, $R^2_{prediction} = 0.353$
- For SO_2 $R^2 = 0.5904$, $R^2_{adj} = 0.5261$, $R^2_{prediction} = 0.4614$

R^2 values decreased, naturally. But adjusted ones slightly increased. Moreover, prediction R^2's significantly increased. Unnecessary variables dramatically effect PRESS score.

3.2.2 Wilks Λ

Variable	$Wilks\ \Lambda$	$p\ value$
A1	0.787	0.003
A2	0.684	0
A3	0.745	0.001
A5	0.921	0.128
A6	0.842	0.013
A8	0.761	0.001
A9	0.61	0
A15	0.946	0.248

Table 6: Wilks Λ for Enhanced Model

After removing the variables A4,A7,A10,A11 the p-values of the remaining variables decreased, Table 6. Their significance become higher, enhanced model performs better in that regard. $A15$ is only insignificant variable at the model, but when we discard it, model performs worse.

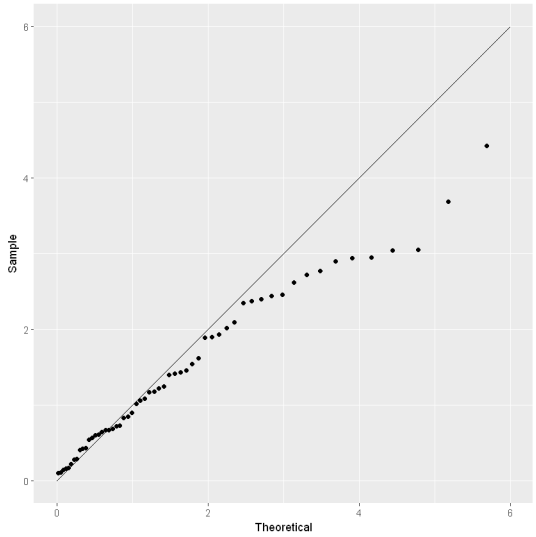
3.2.3 Linear relation between X and Y

We seek for any transformation for independent variables to increase linear relationship, however untransformed variables gave the best performance.

3.2.4 Multivariate Normality Check

We plotted Multivariate QQ-plot by standardized distance formula. Although the fit at right tail far from normality, any transformation made no difference. We also check univariate normality for each error term by Shapiro-Wilk test. P-values for error terms are 0.16 for NO and 0.87 for SO_2.

Figure 6: Multivariate QQ-plot



3.2.5 Zero Mean and Heteroscedasticity of Variance of error terms

Mean of the error terms are equal to zero for both error vectors. Non-constant Variance Test p values are 0.76 and 0.45, respectively. Therefore, we could conclude there is no heteroscedasticity.

3.2.6 Autocorrelation

We got Durbin-Watson p-values are greater than 0.05 for several runs(Since the test consist some bootstrap methods we tried several runs). In addition scores are near to 2.

3.2.7 Multicollinearity

To check multicollinearity, we obtained VIF values. There is no indication for multicollinearity problem.

Variables	A1	A2	A3	A5	A6	A8	A9	A15
VIF	1.88	1.52	2.59	1.82	1.77	1.28	2.05	1.50

Table 7: Variance Inflation Factors for Enhanced Model

3.3 Discussion

In Multivariate Multiple Regression Model, first of all we identified dependent and independent variables. The variable "Death Rate" is removed because it does not hold sufficient meaning to explain dependent variables. After that, A12: Relative pollution potential of Hydrocarbons is removed, because it consist similar information with: A13 Relative pollution potential of Nitric Oxides. Since we would like to compare the predictors, we have standardized them before setting up the models. We applied appropriate transformations to the responses and removed the unnecessary variables. In the end, enhanced model had all properties that is necessary for Multivariate Multiple Regression Model, such as homoscedasticity, linearity, normality, no multicollinearity. It also brings meaningful explanations to what can effect the relative pollution potential of toxic gases such as oxides of Nitrogen and Sulfur Dioxide. Precipitation(A1), Temperature(A2,A3), size of the household(A5) and years of schooling for persons over age 22(A6) have negative relation with the relative pollution potential of NO and SO_2 whereas population density in urbanized areas(A8) and degree of atmospheric moisture in our model(A15), increases such pollution potential.

4 Principal Component Analysis

4.1 Introduction to Principal Component Analysis

Let X contains p continuous variables. Our aim is to find p different linear combination of X's. That satisfies two conditions.

1. Y's are independent
2. $var(Y_i) \geq var(Y_k) \; \forall \; i > k$

Here is the model for PCA:

$$Y_1 = e_{11}X_1 + e_{12}X_2 + ... + e_{1p}X_p$$
$$Y_2 = e_{21}X_1 + e_{22}X_2 + ... + e_{2p}X_p$$
$$.$$
$$.$$
$$.$$
$$Y_p = e_{p1}X_1 + e_{p2}X_2 + ... + e_{pp}X_p$$

Therefore PCA could be viewed as a *variance maximization* problem. Where;

$$var(Y_i) = e_i^T var(X) e_i$$

And there are two **constraints** of this optimization problem:

1. $e_i^T e_i = 1$
2. $cov(e_i, e_j) = 0$

To obtain this such e_i's, we should calculate **normalized** eigenvectors.

4.2 Preparing Data for Principal Component Analysis

Applying PCA to our dataset, we are going to work on the variables that affect $DeathRate(B)$. One can see our first six observations at Table 8.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
36	27	71	8.1	3.34	11.4	81.5	3243	8.8	42.6	11.7	21	15	59	59
35	23	72	11.1	3.14	11	78.8	4281	3.6	50.7	14.4	8	10	39	57
44	29	74	10.4	3.21	9.8	81.6	4260	0.8	39.4	12.4	6	6	33	54
47	45	79	6.5	3.41	11.1	77.5	3125	27.1	50.2	20.6	18	8	24	56
43	35	77	7.6	3.44	9.6	84.6	6441	24.4	43.7	14.3	43	38	206	55
53	45	80	7.7	3.45	10.2	66.8	3325	38.5	43.1	25.5	30	32	72	54

Table 8: First six observations of the variables that we work in PCA

First of all, most important assumption in PCA is *All variables should be numeric.* Because, we perform standardization. And then we find covariance matrix. We need numeric varaibles both of the cases. It is satisfied, we do not have any categorical or ordinal data. As we mentioned, we standardize the Variables. Because, if we do not, since PCA is a variance maximization procedure, the variables with higher variances would dominate Principal Components. At Table 9, we provide standardized values.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
-0.14	-0.65	-0.76	-0.48	0.57	0.5	0.11	-0.44	-0.34	-0.75	-0.64	-0.18	-0.16	0.08	0.27
-0.24	-0.99	-0.55	1.57	-0.91	0.03	-0.41	0.28	-0.93	1	0.01	-0.32	-0.27	-0.23	-0.1
0.66	-0.49	-0.13	1.09	-0.39	-1.39	0.13	0.26	-1.24	-1.44	-0.47	-0.35	-0.36	-0.33	-0.65
0.96	0.85	0.92	-1.57	1.09	0.15	-0.66	-0.52	1.71	0.89	1.5	-0.22	-0.31	-0.47	-0.28
0.56	0.02	0.5	-0.82	1.31	-1.62	0.72	1.76	1.4	-0.51	-0.02	0.06	0.33	2.4	-0.46
1.57	0.85	1.13	-0.75	1.38	-0.91	-2.75	-0.38	2.99	-0.64	2.67	-0.09	0.2	0.29	-0.65

Table 9: Standardized values of first six observations of the Variables that we work in PCA

4.3 Principal Component Model

Writing raw model here would be loss of space (There are 15 equations each with 15 variables). Instead, we will provide summary information such as tables and plots. One could get more detailed information at the Github page of the project[1].

4.3.1 Marginal and Cumulative Proportional Variance of Principal Components

We could compute marginal contribution of each principal component to explain the variation in X. Moreover, we could analyze their combined contribution on explaining the variance. As we see in the Table 10, contributions after 11^{th} Principal Component are less then 1%.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Marginal Proportional Variance	0.302	0.171	0.133	0.094	0.08	0.066	0.049	0.031	0.024
Cumulative Proportional Variance	0.302	0.473	0.606	0.7	0.781	0.847	0.896	0.928	0.951

	PC10	PC11	PC12	PC13	PC14	PC15
Marginal Variance	0.017	0.011	0.009	0.008	0.004	0
Cumulative Proportional Variance	0.968	0.979	0.988	0.996	1	1

Table 10: Variance Contribution of Principal Components

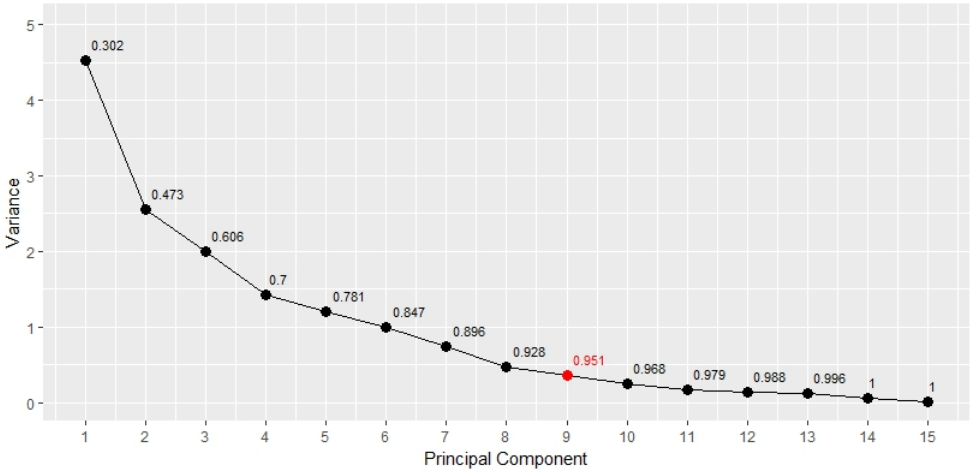
4.3.2 Scree Plot

Scree plot is a $IndexValue\ of\ Principal\ Component$ vs $Its\ Variance$ plot. It is a decreasing plot due to the second condition that we mentioned at *Introduction to Principal Component Analysis* part. We also provide cumulative proportional variance at the top of each point. For further analysis we are going to cut last 6 principal components. Because we could explain 95% of the variation in X with first 9 principal components.

4.3.3 Correlation Between Principal Components and Variables

With this correlation matrix (Table 11), we could get an intuition about "which variables effect given principal component". In our correlation matrix, we see that the first Principal Component has a strong positive linear relation with $A6, A7$ and negative with $A1, A3, A11$. We could make similar comments on other Principal Components too. However one may notice that magnitude of the correlation between Principal Components and Variables decrease through the Principal Components. This is an expected result, when the index of Principal Component increases, its variance decreases(condition 2 at *Introduction to Principal Component Analysis* part). In other words, it become more non-informative about Variables.

Figure 7: Scree Plot



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
A1	-0.737	-0.163	-0.114	-0.362	-0.083	0.103	0.205	0.389	0.199
A2	-0.154	0.665	0.044	-0.524	0.1	0.126	-0.393	-0.026	-0.114
A3	-0.729	0.313	0.202	-0.097	-0.409	-0.114	-0.019	0.116	-0.209
A4	0.34	-0.561	-0.408	-0.505	-0.006	-0.239	0.222	0.027	0.073
A5	-0.632	-0.093	-0.001	0.658	0.137	0.117	-0.144	-0.013	0.182
A6	0.615	0.141	0.683	0.044	-0.108	0.037	0.136	0.058	0.087
A7	0.77	0.052	0.101	0.099	-0.234	0.225	-0.213	0.431	-0.069
A8	0.143	0.168	-0.614	-0.186	-0.451	0.412	-0.222	-0.189	0.237
A9	-0.636	0.611	0.063	0.137	0.047	0.223	0.238	0.089	0.163
A10	0.42	0.312	0.567	-0.178	-0.345	0.139	0.328	-0.221	0.098
A11	-0.758	0.43	0.022	-0.258	0.218	-0.149	0.114	-0.078	-0.038
A12	0.608	0.645	-0.217	0.038	0.259	-0.249	-0.002	0.08	0.125
A13	0.577	0.654	-0.314	0.067	0.23	-0.228	0.073	0.087	0.09
A14	0.147	0.294	-0.663	0.304	-0.186	0.238	0.402	0	-0.292
A15	0.209	-0.183	0.125	-0.229	0.64	0.633	0.109	0.003	-0.085

Table 11: Correlation between Standardized Variables and Principal Components

4.3.4 Distribution of Principal Components

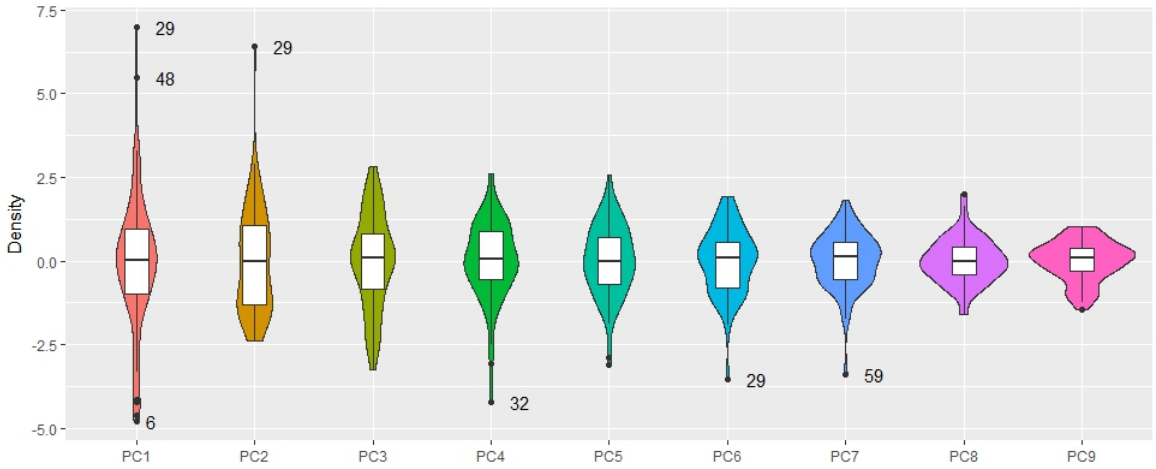
Violin plot provides kernel density estimations of Principal Components. We also provide box plots and indicate outliers at Figure 8. We learned three facts from this plot:

1. 29^{th} observation is extreme outlier for several Principal Component
2. We proved 2^{nd} condition of Principal Component Analysis graphically$(var(Y_i) \geq var(Y_k) \; \forall \; i > k)$
3. Kernel density estimations and box-plots indicate that distribution of Principal Components are symmetric

4.4 Discussion

At the start of this section, we gave a brief introduction to Principal Component Analysis. Later we applied it to our Variables. We observed that, dimension could be reduced from 15 to 9 by losing less than 5% of the total variation in X. We are going to use these 9 principal components.

Figure 8: Kernel Density Estimation of Principal Components



5 Regression with Principal Components

We repeated the steps at *Multiple Linear Regression* with taking Principal Components as response.

5.1 Regression with Ordinary Least Squares Setup

We performed a Linear Regression where first 9 Principal Components are our independent variables and *DeathRate* is our response. So far we expect similar results with section 2. Here is the model that we want to fit our data:

$$[Death\ Rate]_{60\times1} = [PCA]_{60\times10}[\beta]_{10\times1} + [\epsilon]_{60\times1}$$

We are going to estimate β by OLS estimation.

5.2 Initial Linear Model for Death Rate by first 9 PC's

We estimated regression coefficient with first 9 principal components. Here is the model formula:

$$Death\ Rate = 940 - 15.5(PC1) + 6.52(PC2) - 18.1(PC3) + 7.05(PC4) - 1.73(PC5) + 16.6(PC6) + 25.7x(PC7) + 5(PC8) + 18.4(P$$

R^2 values:

- $R^2 = 0.7297$, $R^2_{adj} = 0.681$, $R^2_{prediction} = 0.5904$

Since R^2 values are quite low compared to section 2, we run stepwise regression to eliminate some principal components. According to stepwise regression, we dropped $PC5$ and $PC8$. Beside the stepwise regression, those variables are obviously insignificant.

5.3 Linear Model for Death Rate by 7 PC's

Again, we estimated regression coefficient, this time with first 7 principal components. Here is the model formula:

$$Death\ Rate = 940 - 15.5(PC1) + 6.52(PC2) - 18.1(PC3) + 7.05(PC4) + 16.6(PC6) + 25.7(PC7) + 18.4(PC9)$$

R^2 values:

- $R^2 = 0.7257$, $R^2_{adj} = 0.6888$, $R^2_{prediction} = 0.635$

5.3.1 Significance of Variables

As we see in t-tests (Table 12), all variables are significant with 0.1 significance level. $PC7$ is the variable has the most effect on response estimation. $PC1$, $PC3$ have negative effect and $PC2$, $PC4$, $PC6$, $PC7$, $PC9$ have positive effect.

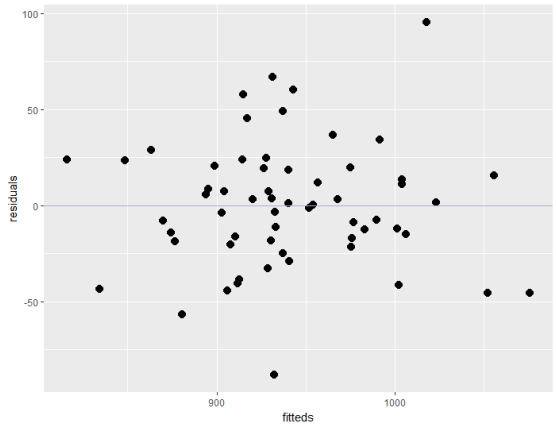
coefficients	estimate	standard error	t value	p value
(Intercept)	940.40	4.48	209.90	0.00
PC1	-15.51	2.12	-7.31	0.00
PC2	6.52	2.82	2.31	0.02
PC3	-18.13	3.19	-5.68	0.00
PC4	7.05	3.80	1.86	0.07
PC6	16.58	4.53	3.66	0.00
PC7	25.71	5.26	4.89	0.00
PC9	18.38	7.57	2.43	0.02

Table 12: t - test

5.3.2 Zero Mean and Heteroscedasticity of Variance of error terms

The error term of our model has a mean zero value. Here is Heteroscedasticity analysis. We performed *Breusch-Pagan test* [4]. The $p\ value = 0.6$ of H_0: *Error variance do not change with the level of the response.* We confidently conclude that, we do not have a non constant variance problem.

Figure 9: Fitted vs Residuals Plot



5.3.3 Autocorrelation

There is no autocorrelation problem either. According to Durbin-Watson test [5] test score is 2.04 and $p\ value = 0.88$ with H_0: *Errors are serially uncorrelated.*

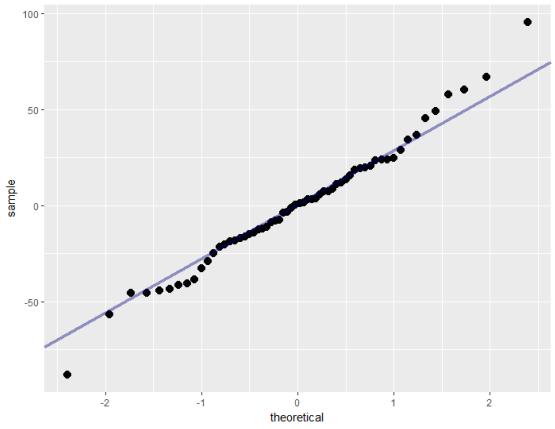
5.3.4 Multicollinearity

Multicollinearity is not an issue with a PCA model. Since all Principal Components are linearly independent (first condition that we mentioned at *Introduction to Principal Component Analysis* part).

5.3.5 Normality

To check normality, we obtained qqplots and shapiro-wilk test [6]. $p\ value = 0.28$ with H_0: *Residuals are normally distributed.*

Figure 10: QQ-plot for Normality of Error terms



5.4 Discussion

We try to construct a linear model for Death Rate with the PCA's that we found in previous section. Keeping track of which original variable is contributing model much is a hard work to do when the predictors are Principal Components. And in our PC's, we can not give a clear information such: This variable contribute these specific PC's. Therefore, if reader wants to get a neat interpretation about the variables, we could refer section 2 and 3.

6 Canonical Correlation Analysis

Canonical analysis provide us to identify the linear relation magnitude among sets of variables. This analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. In this analysis, firstly we determine the pair of linear combinations which have the largest correlation. Next, we find the pair of linear combinations having the largest correlation among all pairs uncorrelated with the first selected pair. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations which measure the strength of association between the two set of variables.

6.1 Grouped Variables

We grouped our variables according to information they carry. These five groups of variables at below are measured and the relationship among them are to be considered.

- Weather Related Variables
  - A1:average annual precipitation in inches
  - A2 average January temperature in degrees Fahrenheit
  - A3 average July temperature in degrees Fahrenheit
  - A15 percent relative humidity, annual average at 1pm
- Demographic-Social Variables
  - A4 percent of population 65 years old or older
  - A6 number of years of school, persons 22 years or older have completed
  - A8 population per square mile in urbanized areas
  - A9 percent nonwhite population
- Demographic-Economical Variables
  - A5 average household size
  - A7 the rate of households fully equipped with household appliances
  - A10 percent office workers
  - A11 poor families (annual income under $3000)
- Pollutant Index Variables
  - A12 relative pollution potential of hydrocarbons
  - A13 relative pollution potential of nitric oxides
  - A14 relative pollution of sulfur dioxides
- Death Rate
  - B The death rate in cities of US (deaths per 100,000)

6.2 Canonical Correlations

In this section we focus on the necessary details for obtaining the canonical variables and their correlations. The best squared values of the canonical variate pairs are shown in Table 13. As one can see, in weather variables, %49 of the variation is explained by the variation in social situation variables; %56 of the variation is explained by economic situation variables; %51 of the variation is explained by pollution and only %28 of the variation can be explained by death rate. If we focus on variables related to social situations, %78 of the variation is explained by the variation in economical situation variables; %62 of the variation is explained by death rate and only %25 of the variation can be explained by the pollution variables. Moreover, in variables related to economic situation, besides weather and social variables, pollution and death rate variables also explain the variability with %27 and %26 respectively. Lastly, in pollution variables, apart from weather variables, social and economic situation variables, %33 of the variation can be explained by death rate.

Table 13: Highest Canonical Correlation Between Groups

	weather	**social**	**economic**	**poll**	**DR**
weather	1	0.49	0.56	0.51	0.28
social	0.49	1	0.78	0.25	0.62
economic	0.56	0.78	1	0.27	0.26
poll	0.51	0.25	0.27	1	0.33
DR	0.28	0.62	0.26	0.33	1

6.3 Discussion

These squared values of the canonical variate pairs shows that social life situation highly correlated with economical variables and death rate. Therefore variation in Death Rate, highly understandable with social life related variables.

7 Cluster Analysis

Before starting Cluster Analysis, we updated our knowledge about data. We found another source of it [11], which contains city abbreviation for each observation. Therefore, we added city names into our dataset as row names.

Data clustering is a method such that it grous a set of variables in a way that elements in the same group are more similar to each other than those in other groups (clusters). We want to achieve high within-cluster similarity and low inter-cluster similarity. We used 2 different techniques:

- K-means
- Agglomerative Hierarchical Clustering

We also provide a spatial visualization of data clusters.

7.1 K-means

Given k, the k-means algorithm is implemented in four steps[12]:

1. Partition objects into k nonempty subsets
2. Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster)
3. Assign each object to the cluster with the nearest seed point
4. Go back to Step 2, stop when the assignment does not change

To be more specific, we used Lloyd's algorithm. It differs from classical k-means, because it treats input as continuous geometric region rather than a set of points [13]. K-means algorithm could only discover linear-decision boundaries. It is not able to detect complex structured data.

7.1.1 Elbow Method

We used elbow method to decide the number of clusters. It is the most common method that is used in k-means. It basically plot *Total Within Sum of Squares* vs *Number of Clusters*. We hope to see an elbow shape, and we select the number of cluster where elbow seen. The stopping rule is pretty similar to scree plot and pretty straightforward, *After this point additional clusters do not significantly lower the inter-cluster variety*. According to Figure 11, we stopped at 6 clusters. After 6 clusters we could not be able to lower inter-cluster variability, significantly.

Figure 11: Elbow Plot

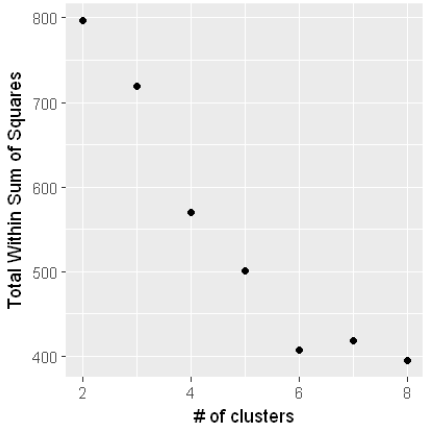
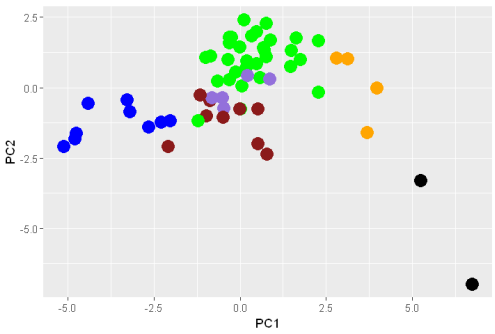
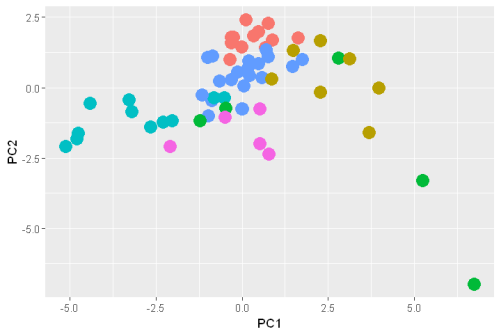


Figure 12: Principal Component 1 vs 2 Conditioned on Cluster Number

(a) Hierarchical Clustering



(b) K-Means



7.1.2 Illustration

Illustration of a High-Dimensional clustering is not easy. One method is: draw the highest contributed 2 variables conditioned on cluster numbers. Beside that we could use principal components. Therefore we obtained first two principal component for our input variables. They explain a total 47% variability. To be clear, we did not run the cluster algorithm by principal components, we only obtained them for illustrative purposes.

We colored the clusters on $PC1$ and $PC2$ (Figure 12b). The plot seems very complicated. At the center red and blue ones seems very close to each other, we could not say it is a successful clustering by looking $PC1 \times PC2$ subset.

7.2 Hierarchical Clustering

We applied *Ward's minimum variance method* [14]. It is a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function, which is error sum of squares. Since we selected 6 clusters for k-means, to make a clear comparison let us cut the hierarchical tree where there are 6 branches decides the cluster of leaf nodes.

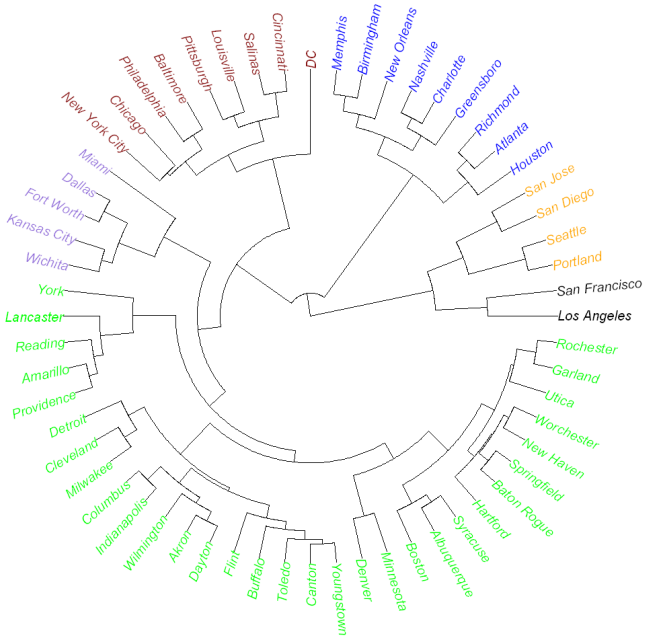
7.2.1 Illustration

We used dendogram (Figure 13) for visualizing our implementation. Which is the most common visualization method of hierarchical clustering algorithms. We colored them according to branches that cuts the tree in to 6 pieces. The brown group contains biggest cities in US. The two biggest western cities, San Francisco and Los Angeles are in the same cluster. We also provide PCA plot for hierarchical clustering in Figure 12a. It looks well separated than k-means. Their clustering -somehow- close to each other.

7.3 Spatial Visualization of Clustering Methods

Since we have city information for each observation, we obtained latitude and longitude data by sending queries to open map servers. After obtaining it, we visualized cluster assignments of each cities on US map in Figure 14. We observed that, there are some location patterns of clusters. For example in Figure 14a, small western cities are in orange color and big western cities are in black color. In addition, southern cities are in blue and northeast in red. We could comment on similar things for Figure 14b too. Clustering can give us idea about cities based on similarities and dissimilarities. Although it should be noted that it does not give answer for the question "what are those similarities or differences". Clustering is only a grouping techniques that consider similarities as distances based on the property of the variables. On the other hand, we observed that we could extract a geographic information from our data. Observations -somehow- separable spatially with given variables.

Figure 13: Fan type Dendogram of Hierarchical Clustering Algorithm

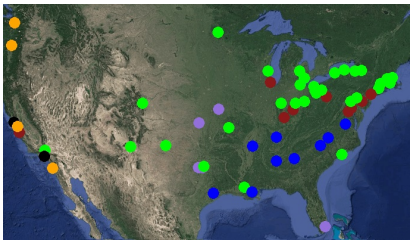


7.4 Discussion

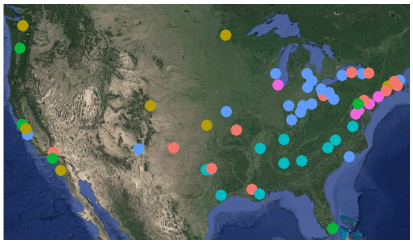
We applied two clustering methods one is a Partitional Clustering: K-means and the other one is Hierarchical Clustering: Agglomerative. We do not use any evaluation algorithm about the goodness of them. Since they are pretty different algorithms, a comparison approach would be insufficient. Instead we gave comparison by graphical intuition. We provide several plots. One of them was spatial one. Although, we did not provide any geographic information while clustering, clustering algorithm discovered interesting spatial patterns. This brings us to our next chapter.

Figure 14: Spatial Visualization of Cluster Algorithms

(a) Hierarchical Clustering



(b) K-Means

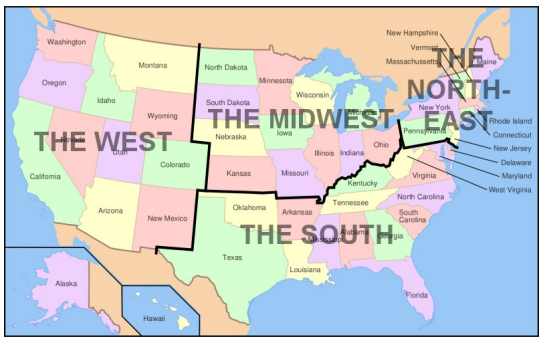


8 Comparison of Classification Methods

In clustering part, we observed that our observation with given variable structure could explain a meaningful information, spatially. Therefore for classification task we obtained regions of each cities. US has 4 regions according to US Census Bureau [15]. Here we show them in Figure 15. We labeled our data according to "which region their cities are belong to". We have total 60 observations. Here are total occurrences of regions:

$Midwest$	$Northeast$	$South$	$West$
17	14	19	10

Figure 15: Census Regions of the United States



In this section we try to find a function such that:

$$f(y = Midwest|X) = p_1$$
$$f(y = Northeast|X) = p_2$$
$$f(y = South|X) = p_3$$
$$f(y = West|X) = p_4$$
$$where \sum_{i=1}^{4} p_i = 1 \; and \; X = \{A1, A2, A3 \ldots, A15, B\}$$

To find such functions we applied 4 different classification methods. Namely, Softmax Regression, Multi-Layer Perceptron, Decision Tree and K-Nearest Neighbors. At the end, we compare their performance via K-folds Cross Validation. Since all functions return a probability distribution, for simplicity we state $y = k$ where k is the Region with highest probability for given input.

8.1 Descriptive Analysis of Data by Softmax Regression

8.1.1 Softmax Function

Since we do not have binary output, we can not apply logistic regression to our data. Instead we could define a softmax regression which guarantees that output will always fall between zero and one. It is one-versus-all classifier. That means, It calculates the signals coming from variables to interested output and divides to whole signals. To be clear, here is an example for calculating $Northeast$ probability of a given observation. Suppose that $Midwest = a, Northeast = b, South = c, West = d$

$$P(y = b|X) = \frac{e^{X^T w_b}}{\sum_{i=a}^{d} e^{X^T w_i}}$$

Table 14: Coefficients of Softmax Model

	Coefficient			
Variable	*Midwest*	*Northeast*	*South*	*West*
Intercept	0.4	-0.32	0.43	-0.54
A1	-0.3	**0.51**	**0.74**	**-0.94**
A2	**-0.85**	-0.14	**0.52**	0.47
A3	0.26	-0.26	0.24	-0.24
A4	**-0.65**	**1.05**	**-1.1**	**0.7**
A5	-0.02	0.06	0.05	-0.08
A6	**0.57**	-0.05	-0.33	-0.19
A8	-0.17	**0.66**	-0.48	-0.01
A10	**-0.72**	0.14	-0.13	**0.7**
A11	-0.03	-0.05	0.07	0.01
A15	0.03	-0.03	-0.03	0.02
B	0.02	0	0.02	-0.03

8.1.2 Model Building and Coefficient Summary

As objective function we used *Cross Entropy Loss*. It was the only option in our library. Since it makes easier to take derivative with respect to weights in a softmax function.

We have very limited data. For example, we have 10 observations for $West$ class. If we apply softmax directly, we immediately overfit because of the $number\ of\ weights > number\ of\ observations$ problem. We have 10 observation for $West$ class with 15 predictors. So that, we applied lasso penalty to regularize the model. This has the effect of shrinking the coefficient values, and eventually, coefficients with minor contribution getting closer to zero. Also, in this section we constructed our model with all observation, so we did not split the data as *train* and *validation* sets.

In Table 14 we gave estimated values of coefficient in each region. Since variables are standardized, they are comparable. We also indicated high contributions with bold font. As one can see, Variables $A1, A2, A4, A6, A8, A10$ are quite influential for all region models. For example increasing in $A1$ negatively effects probability of $West$ and $Midwest$ but positively effects $Northeast$ and $South$ probabilities. This makes sense because A1 represents *Average annual precipitation in inches*. The climate of $West$ quite dry. But it rains in $South$ and $Northeast$ regions. A2 represents *Average January temperature in degrees Fahrenheit*. In winters $Midwest$ temperature is dramatically low, so that helps the model to classify regions. These results are quite obvious but the effect of $A4$ is quite interesting. It represents *Percent of population 65 years old or older*. The senior ratio is a quite high in $Northeast$ and $West$ regions compared to other regions.

Also one could notice that some variables are dropped due to the lasso penalty.

8.1.3 Accuracy and Sensitivity

We used Accuracy and Sensitivity measures to interpret our classification scheme.

$$\begin{aligned} Overall\ Accuracy: &\ 0.78 \\ Sensitivity\ Midwest: &\ 0.88 \\ Sensitivity\ Northeast: &\ 0.79 \\ Sensitivity\ South: &\ 0.79 \\ Sensitivity\ West: &\ 0.6 \end{aligned}$$

In general our model correctly label 47 out of 60 observations. It performs very well at classifying $Midwest$ observations. Its performance at $West$ quite poor, compared to other ones.

8.1.4 Multi-Layer Perceptron

In multi-layer perceptron (or artificial neural network) section we constructed a highly regularized network. Because, according to Universal approximation theorem[16], "a simple neural networks can represent a wide variety of interesting functions when given appropriate parameters". So they have an incredible overfitting tendency. And we have very few observations. To deal with, we constructed the network at Figure 16. We have 2 hidden layers each with size 11, one for bias term. As one can see in each hidden layer we used BN:Batch Normalization and DO:Dropout as regularization. In addition we add $L2$ regulizers to each hidden layer. In total it has 354 trainable parameters.

8.1.5 Accuracy and Sensitivity

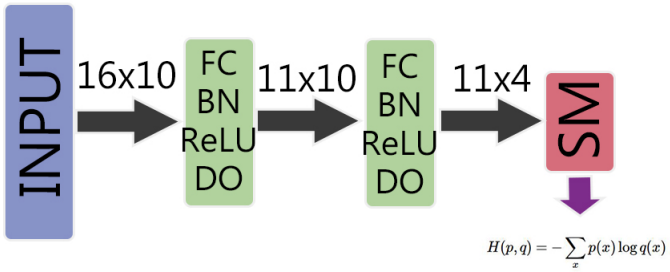
Since Neural Networks are black boxes, we do not give any interpretations about variable contribution.

We used Accuracy and Sensitivity measures to interpret our classification scheme same with Softmax Regression.

$Overall\ Accuracy:$	0.9
$Sensitivity\ Midwest:$	0.94
$Sensitivity\ Northeast:$	1
$Sensitivity\ South:$	0.84
$Sensitivity\ West:$	0.8

If we do not use regularization methods excessively, we overfit the data directly. But now, we believe it does not memorize, it learns the structure. However we will test it at K-folds CV section.

Figure 16: Neural Network Architecture with 2 Hidden Layers

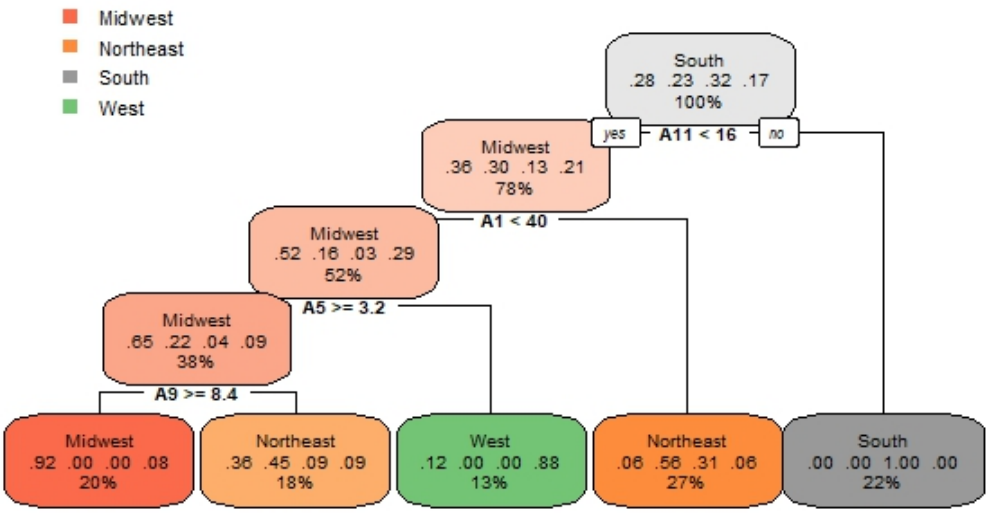


8.1.6 Decision Tree

Decision Tree is great for decision making. In previous section, we fit the model with a Neural Network, but we could not interpret the effect of predictors. In decision trees we could clearly conclude decision patters. In our case at Figure 17, *If the proportion of families with annual income under $3000 (A11) is less than 16% then region is South*. If it is less than 16% and average annual precipitation in inches (A1) is greater than 40, region is *Northeast* with probability 0.56. We could continue interpreting in a similar way.

We used simplest Decision Tree with entropy based. Therefore it only used A11, A1, A5 and A9 variables. One could remember that, in softmax regression section: A1, A2, A4, A6, A8, A10 are affecting the probabilities. Only A1 is the common variable in here. They both correctly classify majority of the variables. Does it make sense? It is indeed. They discover different patterns. Our Softmax classifier is a sophisticated linear classifier but Decision Tree algorithm could discover non linear patterns too. Its mapping function is not even close to linear.

Figure 17: Decision Tree splitted by Information Gain



8.1.7 K-Nearest Neighbors

Lastly we fitted K-Nearest Neighbors. kNN is a non-parametric method which classify items according to their surrounding items. However since it performs lazy evaluation, we could not interpret it without a validation set. We set $k = 7$, by deciding it with trial and error.

8.2 K-folds Cross Validation

We divided our set into 6 folds such that each contains equal amount of those 4 classes. Equal allocation is very important in our case because we have a very small set. If we randomly divide it in some arbitrary folds, we could not be able to observe some classes. Furthermore, we fit our 4 models that we discussed above to dataset. In each run we have 50 observations for training and 10 for validation. We monitored the Validation Accuracy and Validation Sensitivity in each model and each run. Later we took the averages of 6 runs and report both mean and median performances of given model.

One could see in Table 15 and 16 Multi-Layer Perceptron is better than other models. Softmax is quite good too. The performance of our simple Decision Tree is not a match for other models. One could increase the performance of Decision Tree by customizing it with Boosting Algorithms. However we are not going to implement it. Another simple method: k-NN performed quite good due to its simplicity. Its time cost is more than 100 times less than MLP. The main difference between Softmax and MLP arises when predicting $West$ region. $West$ has only 10 observation and in each fold it has 2-3 variables. Detecting it very hard for a shallow learner. However, MLP probably detect some higher interaction and non-linearities in the relation.

8.3 Discussion

We proved that our variables could be used to model region information. They contains spatial information. We also provide a model comparison for our data. We showed that, with given strong regularization, neural networks could be applied for small data.

Table 15: Median of the Measures in 6-fold Cross Validations

	Softmax Regression	Multi-Layer Perceptron	Decision Tree	K-Nearest Neighbors
Overall Accuracy	0.7	**0.75**	0.5	0.65
Sensitivity Midwest	**0.83**	**0.83**	0.5	0.67
Sensitivity Northeast	0.5	**0.83**	0.5	**0.83**
Sensitivity South	**0.71**	**0.71**	0.67	0.67
Sensitivity West	0.5	**0.75**	0.25	**0.75**

Table 16: Mean of the Measures in 6-fold Cross Validations

	Softmax Regression	Multi-Layer Perceptron	Decision Tree	K-Nearest Neighbors
Overall Accuracy	0.7	**0.75**	0.52	0.63
Sensitivity Midwest	**0.81**	0.75	0.56	0.72
Sensitivity Northeast	0.64	**0.78**	0.47	0.61
Sensitivity South	**0.79**	0.74	0.67	0.62
Sensitivity West	0.42	**0.75**	0.33	0.58

9 Conclusion

Since we have investigated different questions in each section, we are going to give a brief summary to all of them.

In second chapter, we seek a linear model to explain the variability in *Death Rate*. As a result of our Linear Model, A1, A8, A9 and A14 have a positive linear effect on Death Rate. On the other hand; A2, A3 and A6 have a negative linear effect.

In third chapter, we investigated the factors that effect pollutant potentials (sulfur dioxide and oxides of nitrogen). A1, A2, A3, A5, A6 and A15 have negative effect on both indication of pollution. On the contrary, A8 and A9 have positive linear effect.

In fourth and fifth chapters, we applied Principal Component Analysis to our data. First 9 PC's explained the 95% variation in the data, therefore they have been used to construct a linear regression model to explain Death Rate. It should be noted that PC does not give clear representation about the variables neither contribution of the variables nor what would be their magnitude to the responses.

In chapter six, we applied Canonical Correlation Analysis. We grouped our variables as: weather, social life, economical indexes, pollutants and Death Rate. Social and Economical variables highly correlated with each other. In addition social factors are also correlated with Death Rate.

In seventh chapter, we conducted a cluster analysis. With observed that, cluster indexes associated between geographical information of observations.

Lastly, in chapter eight, we applied several classification methods to predict the region of observations. Although neural networks performed best in our analysis, it lack of interpretability made us to perform other classification methods, such as: multinomial regression and decision trees.

All in all, we applied a comprehensive multivariate analyze on 1960 Death Rate statistic of United States. The research direction should be a time series analysis. In US government website, we could obtain Death Rate year by year. The comparison between current data and 1960 data, could insight about "the Death Rate changes among years and which factors continue to effect Death Rate, which factors do not". In addition we should analyze "what are the new factors that effect Death Rate?".

References

[1] Enes Dilber. Project for multivariate class, 2017. URL `https://github.com/enesdilber/467/tree/master/project`. [Online; accessed December 31, 2017].

[2] Enes Dilber. nstats: A browser app for linear regression, 2017. URL `https://nstats.shinyapps.io/nstats/`. [Online; accessed January 1, 2018].

[3] Richard Schwing Gary McDonald. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15(3):463–482, 1973.

[4] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, (47):1287–1294, 1979.

[5] J. Fox. *Applied Regression Analysis and Generalized Linear Models*. Sage, 2008.

[6] Patrick Royston. An extension of shapiro and wilk's w test for normality to large samples. *Applied Statistics*, (31):115–124, 1982.

[7] C. L. Mallows. Some comments on c_p. *Technometrics*, 15(4):661–675, 1973.

[8] Albert R. Leeds. The conversion of carbon monoxide to carbon dioxide by active oxygen. *Journal of the American Chemical Society*, (5):78–91, 1883.

[9] G. E. P. Box and D. R. Cox. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, (26):211–252, 1964.

[10] J. Chambers Becker, R. and A. Wilks. *The new S language*. CRC, 1992.

[11] Mortality rates in us cities. `http://people.stat.sc.edu/habing/courses/data/sascity.txt`. Accessed: 2018-01-10.

[12] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[13] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

[14] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[15] Census regions and divisions of the united states. `https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf`. Accessed: 2018-01-10.

[16] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.