STA 207 Final Project

Winter 2016, University of California, Davis

Project II: Mortality Analysis

By
Zhen Zhang (ezzhang@ucdavis.edu) 913435403
Miao Wang (mgwang@ucdavis.edu) 913495016
Juanjuan Hu (nahu@ucdavis.edu) 913407952

Approved:

Zhen Zhang _____Date_____

Miao Wang _____Date_____

Juanjuan Hu _____Date_____

1 Introduction

Interests in identifying factors that have associations with human mortality arise from their importance in better understanding causes to mortality and in helping people improve their life quality. Many researchers work on studying relationships between mortality and various factors including weather, smoking, social relationship, etc. In our project, we sought to explore associations between mortality with some measures of climate and demography as well as pollution measures. We also discussed in details how the pollution alone affects mortality.

2 Methods and Results

2.1 Data Description and Exploration

2.1.1 Data Description

This dataset contains seven numeric variable: mean annual precipitation (PRECIP), median number of school years completed by persons of age 25 or over (EDUC), percentage of population in 1960 that is nonwhite (NONWHITE), percentage of households with annual income under $3000 in 1960 (POOR), relative pollution potential of oxides of nitrogen (NOX), relative pollution potential of sulphur dioxide (SO2) and total age-adjusted mortality from all causes (MORT). Among these variables, mortality is the response variable and the other six are independent variables. The dataset also contains a column CITY, which is a factor variable. Since the number of its levels is the same as the data size (60), it would contribute nothing to explain the mortality in our models. Therefore, we dropped this column.

2.1.2 Initial data analysis

In order to explore the data, we firstly calculated the summary statistics for all the variables including mean, median and some other important quantiles (Table I.1.2.1). We also constructed each distribution plot (Figure I.1.2.1) and performed simple linear regression for each independent variable (Table I.1.2.2).

2.1.3 Transformations and Standardization

Based on the distribution plots, we found all the independent variables showed some skewness. Therefore, we did three types of transformation (log, square root, square) for each of them and plot the new distribution plots (Figure I.1.3.1). It turned out that square root transformation is preferred for NONWHITE and log transformation is preferred for POOR and SO2. We also standardized response variables and independent variables so that they are all equally scaled. From now on, all the variables we mention are the transformed and standardized ones.

2.1.4 Correlations and Multicollinearity

From the correlation matrix (Table I.1.3.1) and the matrix plot (Figure I.1.3.2), we observed correlation present between variables (for example, SO2 and NOX has correlations of 0.7328). Then we calculated the VIF value for each independent variable (Table I.1.3.2) and found that all the VIF are between 2 and 5, indicating moderate multicollinearity among independent variables.

2.2 Model Building

Here are the results for all the models:

	PRECIP	EDUC	NONWHITE	POOR	NOX	SO2
Stepwise	0.2897	-0.2024	0.4696	0.0000	0.0000	0.3593
PLS	0.3738	-0.2192	0.5058	-0.1251	0.1577	0.2266
Ridge	0.2914	-0.2122	0.3662	-0.0039	0.1454	0.2079
Lasso	0.2904	-0.1972	0.4160	0.0000	0.0931	0.2535

2.2.1 Stepwise Regression

To deal with multicollinearity among the X variables, we can use stepwise method and AIC criterion to find the best subset of X variables sequentially. In each step, one variable is added or deleted from the model according to AIC. We first considered models with only first order variables. The stepwise procedure ended up with best model (Table II.2.1.1).

There was a pattern in the residual vs. fitted value plot (Figure II.2.1.1). Some information may be not covered by current model. We considered adding second order variables (including quadratic terms and interaction terms) into the model. The best model we obtained (Table II.2.1.2) is in the appendix and the residual plot (Figure II.2.1.2) did not improve much. We did cross validation on the above two models. The rates are 0.384, 0.394. Considering the difference was not big and linear model was easier to interpret, we decided the first order model as the final stepwise model.

2.2.2 Ridge Regression

In order to address the issue of multicollinearity, we considered ridge regression model in this part. The optimal penalty coefficient was chosen to be 0.2275 (Figure II.2.2.1). The observed value against the fitted values plot, residuals against fitted value plot and histogram of residuals (Figure II.2.2.2) show appropriateness of our fitted ridge regression model.

2.2.3 Lasso Regression

Lasso is similar method as Ridge Regression except that the penalty is different. We conducted Lasso regression on data. The penalty coefficient was chosen as 0.04165 ((Figure II.2.3.1) according to cross-validation. And it dropped variable POOR.

After checking observed value against the fitted values plot, residuals against fitted value plot and histogram of residuals (Figure II.2.3.2). it was appropriate to use this lasso model.

2.2.4 Partial Least Squares

Partial least squares is another popular method in dealing with data with multicollinearity. Cross validation results (Table II.2.4.1) suggest that 4 components are suitable for this dataset. The loadings are summarized (Table II.2.4.2). They give out the lowest mean squared error rate 0.6440 and can explain 88% information of the predictors as well as 70% of the response variable. The residual plot (Figure II.2.4.1) indicates that the model fits the data pretty well.

2.3 Model Selection

Now we have four models in hand: linear model, partial least square (PLS) model, ridge model and lasso model.

Stepwise	PLS	Ridge	Lasso
0.384	0.644	0.400	0.415

Linear model is very general, and it seems to apply to a dataset that has no particular feature. PLS is suitable for data with high correlation between the independent variables. From the VIF and the correlation table, we know the multi-correlation seems not to be a problem, so PLS is not a good solution here. The result confirms our assumption, which PLS has the largest cross-validation error. Also because PLS is not easy to interpret, we will not choose this model.

Now we discuss the other three regressions. Each model selects coefficients in different methods: stepwise select model by AIC, ridge and lasso select models in cross-validation, while ridge uses L2 penalty while lasso uses L1 penalty, so the coefficient in lasso and stepwise can be zero while in ridge it cannot. All of them are derived from the initial linear regression model, so their coefficients are not significant different from each other, and their cross-validation error are almost identical. By the Occam's Razor principle, we select the simplest model, which is stepwise model.

2.4 Model Validation

We had chosen the stepwise model from our potential models. The fitted model is summarized in the following table.

	Estimate	Std. Error	t value	P value
NONWHITE	0.39628	0.07272	5.450	1.22e-06
EDUC	-0.11816	0.08349	-1.415	0.162605
SO2	0.47188	0.07647	6.171	8.51e-08
PRECIP	0.30388	0.08365	3.633	0.0006

It is important to evaluate goodness of fit, validation of assumptions and prediction ability of our final model.

From the residual plots (Figure II.4.1), we could find the residuals appear to be homogeneity and normality. Therefore, the residuals appear to behave randomly, suggesting that the model fits the data well. And the assumptions of the model were satisfied. Also we have done cross validation previously, the error rate is 0.384, which is the lowest among all the models. So we can deem this model is good in terms of prediction.

In our linear regression model building process in 2.1, we found that there is one potential outlier which has a cooks distance greater than 0.5. If we delete it, the residual plots (Figure II.4.2) shows that the model fits the data better and the cross validation gives lower error rate 0.315.

3 Conclusions and Discussion

3.1 Interpretation of the final model

Our final model includes four predictors, among which annual precipitation, percentage of nonwhite and SO2 pollution are positively associated with mortality while education is negatively associated with mortality. It looks reasonable since people with high education levels might have higher awareness of health status and thus a lower mortality. On the other hand, a higher level of SO2 stands for a more severe air pollution, which is regarded as a cause to respiratory diseases and higher mortality. However, it should be noticed that association do not mean causation. Precipitation and nonwhite cannot be reasons for mortality. Possibly, they are associated with other unknown factors that lead to mortality. Since some variables have been transformed, increase in independent variables may cause changes in response variable in different scales. For example, unit increase in precipitation is associated with 0.2897 unit increase in mortality, while we expect mortality to increase by 0.00359 units along with SO2 increasing by one percent.

3.2 The relationship between pollution and mortality

First let's focus on the correlation table. We found that the correlation coefficient between NOX and SO2 is 0.7328, which is very high. So we think that one of them will not be significant in affecting mortality, which is validated in the regression of mortality to these two variables:

(Table III.2.1) refer toAPPENDIEX

Here we can see SO2 is significant while NOX is not significant at all, and pollution has a positive relationship with mortality. Considering the correlation between them, we conclude that the information lies in NOX is explained by SO2. This can also be seen as the coefficient table listed in the model section part: although the coefficients of SO2 and NOX are different from each other, the sum of them falls in the interval of (0.34, 0.36). This means that pollution indeed influences mortality, and we can only use SO2 to account for this.

In common sense, we know pollution gas are always discharged into the air altogether. In particular, NOX and SO2 are the emission of fossil oil combustion. C. Arden Pop et al. have shown that particular air pollution can be a predictor of mortality. So when NOX is polluted, SO2 will be polluted as well. This phenomenon, from another perspective, explains why SO2 and NOX are correlated with each other.

3.3 Nonwhite and Poor

For the same reason that is discussed above, Nonwhite and Poor are modestly correlated with each other, with a coefficient of 0.6427. This collinearity explains the coefficient difference in the three models (stepwise, lasso and ridge). And we see that when seeing them as a group, the coefficients do not vary that much. In the cities of this dataset, the white people might have higher annual income in average than nonwhite people. However, this correlation may differ in other cities.

3.4 City and Region

In this project we dropped the city variable, and used the other six variables to do the regression. However, we can categorize these cities into different groups by their geographic regions, and see whether our new variable, region, has an effect on the mortality. Since different regions may have different social environments and cultures, it is possible that region will influence mortality.

Reference

[1] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li, (2004). "Applied Linear Statistical Model" (5th edition).

[2] C. Arden Pope, III, Michael J. Thun, Mohan M. Namboodiri, Douglas W. Dockery, John S. Evans,Frank E. Speizer, and Clark W. Heath, Jr. "Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults", American Journal of Respiratory and Critical Care Medicine, Vol. 151, No. 3_pt_1 (1995), pp. 669-674.

Appendix

Table I.1.2.1 Summary statistics

	PRECIP	EDUC	NONWHITE	POOR	NOX	SO2	MORTALITY
Min.	10.00	9.00	0.80	9.40	1.00	1.00	790.7
1st Qu.	32.75	10.40	4.95	12.00	4.00	11.00	898.4
Median	38.00	11.05	10.40	13.20	9.00	30.00	943.7
Mean	37.37	10.97	11.87	14.37	22.65	53.77	940.4
3rd Qu.	43.25	11.50	15.65	15.15	23.75	69.00	983.2
Max.	60.00	12.30	38.50	26.40	319.00	278.00	1113.0

Figure I.1.2.1 Histogram of each variable

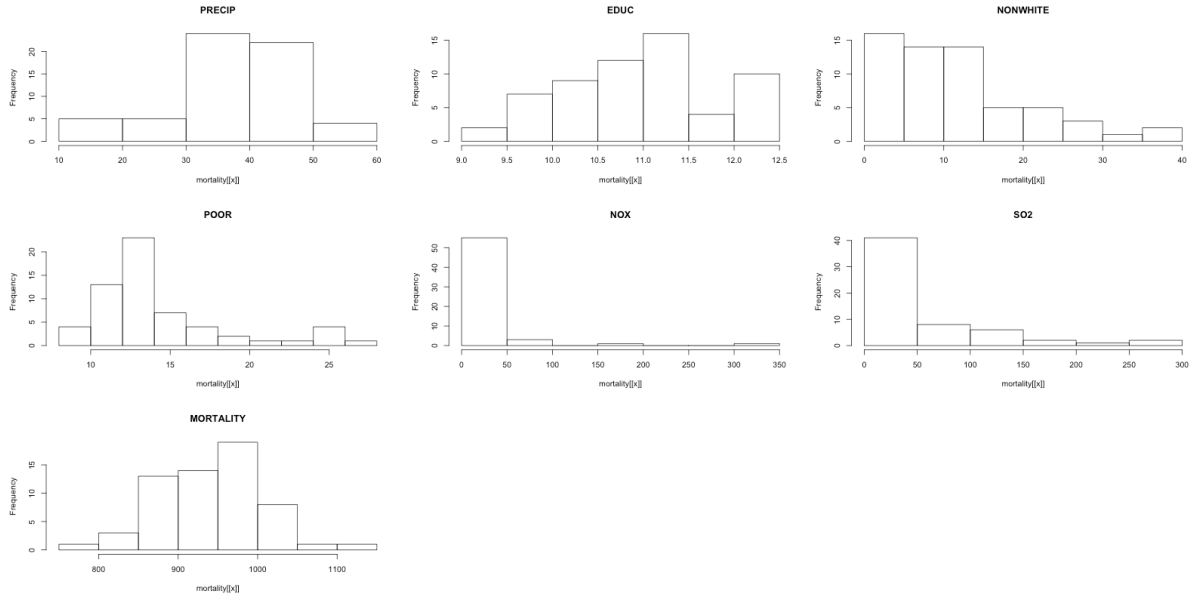


Table I.1.2.2 Single linear regression results summary

	Estimate	Std. Error	t value	Pr(>|t|)	R square
PRECIP	**3.174**	**0.7039**	**4.509**	**3.22e-05**	**0.2596**
EDUC	**-37.601**	**8.306**	**-4.527**	**3.02e-05**	**0.2611**
NONWHITE	**4.4884**	**0.7006**	**6.406**	**2.89e-08**	**0.4144**
POOR	**6.137**	**1.79**	**3.428**	**0.00112**	**0.1685**
NOX	**-0.1039**	**0.1757**	**-0.591**	**0.557**	**0.005988**
SO2	**0.418**	**0.1166**	**3.585**	**0.000691**	**0.1814**

Figure I.1.3.1 Histogram of possible transformations

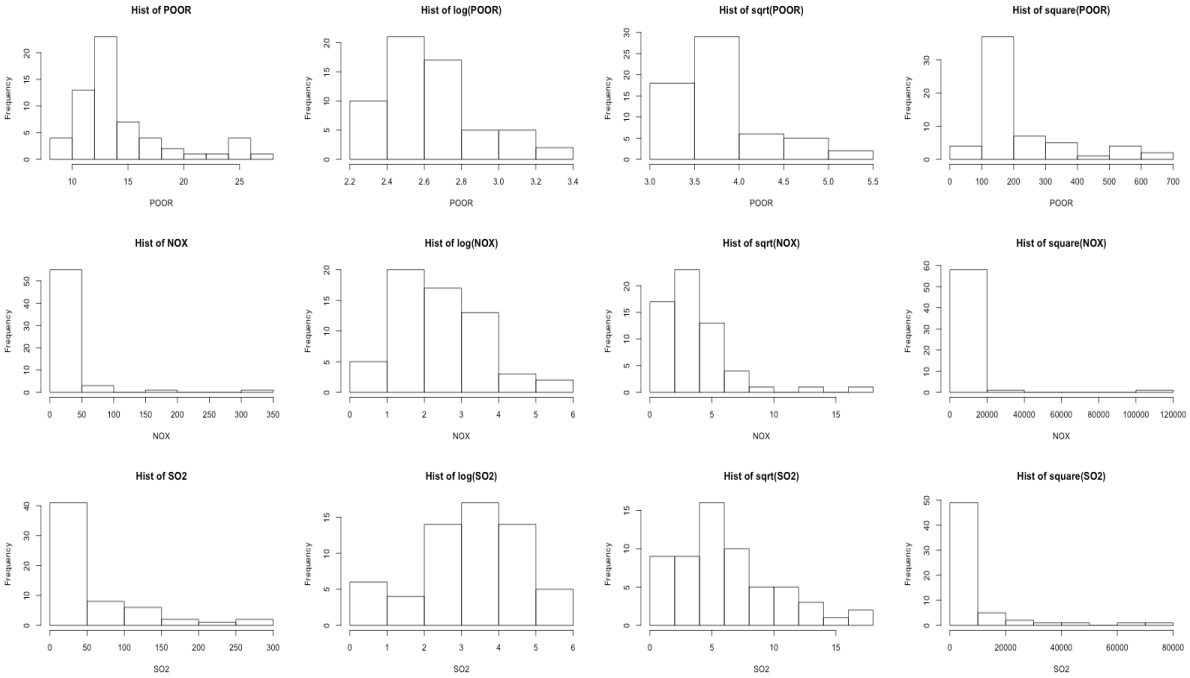


Table I.1.3.1 Correlation matrix

	PRECIP	EDUC	NONWHITE	POOR	NOX
PRECIP	1.0000000	-0.49042518	0.34931594	0.50658543	-0.36830267
EDUC	-0.4904252	1.00000000	-0.15861781	-0.40333845	0.01798472
NONWHITE	0.3493159	-0.15861781	1.00000000	0.64267418	0.19583176
POOR	0.5065854	-0.40333845	0.64267418	1.00000000	-0.09027348
NOX	-0.3683027	0.01798472	0.19583176	-0.09027348	1.00000000
SO2	-0.1211723	-0.25616219	0.05780639	-0.19336723	0.73280742
MORTALITY	0.5094924	-0.51098130	0.62366084	0.41045399	0.29199967

	SO2	MORTALITY
PRECIP	-0.12117231	0.5094924
EDUC	-0.25616219	-0.5109813
NONWHITE	0.05780639	0.6236608
POOR	-0.19336723	0.4104540
NOX	0.73280742	0.2919997
SO2	1.00000000	0.4031300
MORTALITY	0.40313003	1.0000000

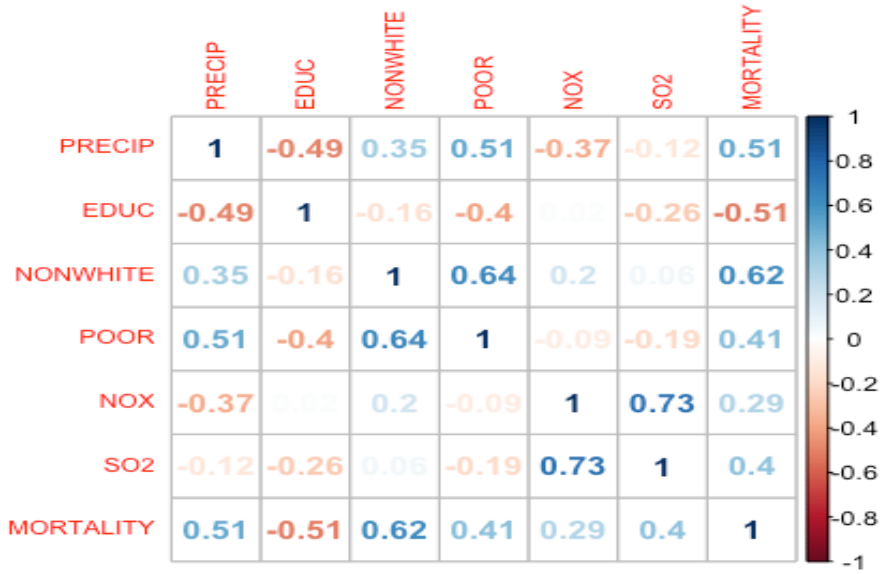
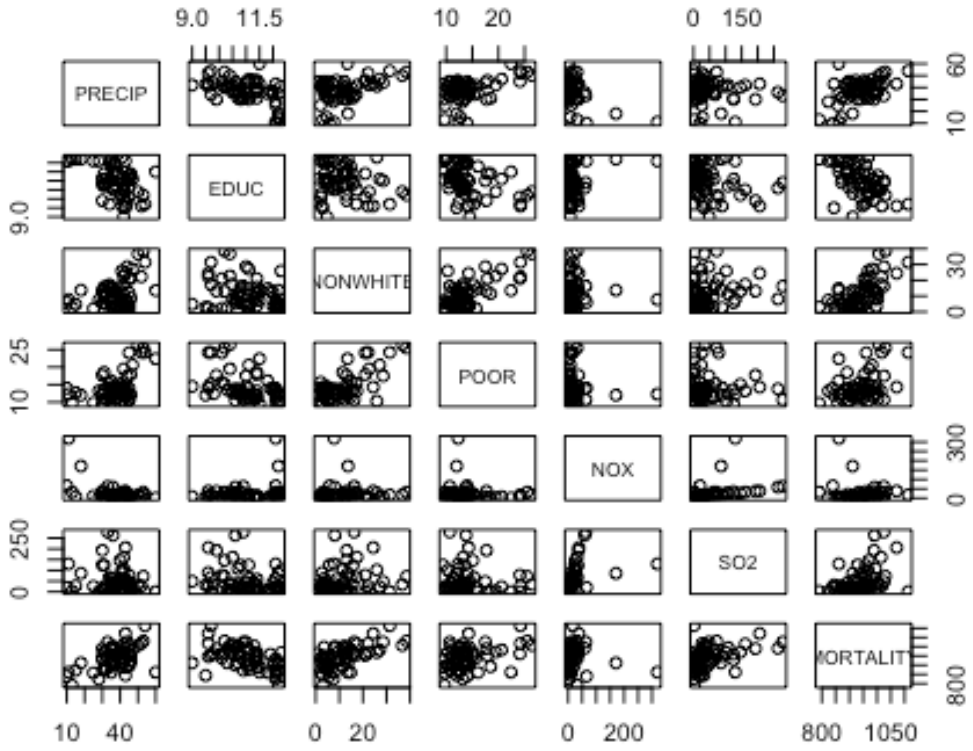


Figure I.1.3.2 Matrix plot



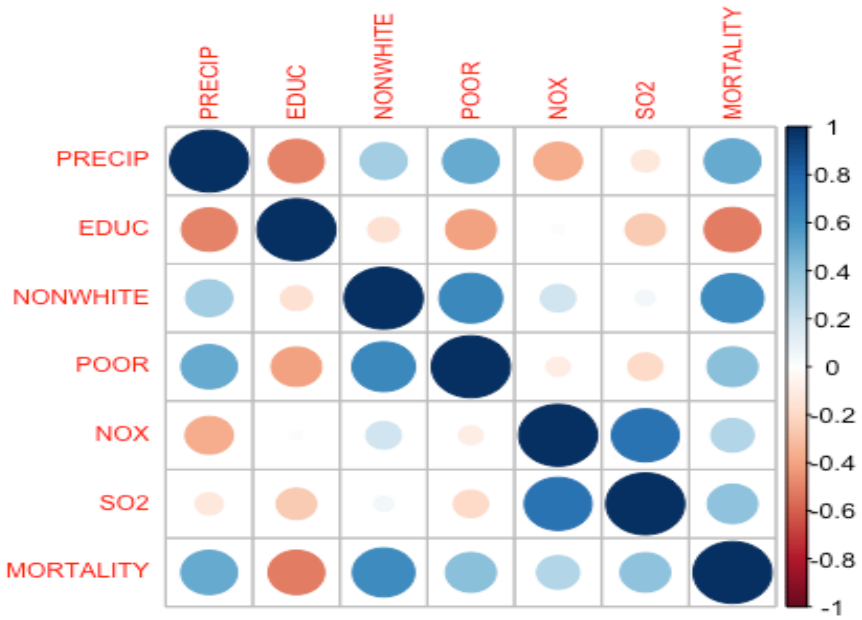


Table I.1.3.2 VIF

	PRECIP	EDUC	NONWHITE	POOR	NOX	SO2
VIF	2.114044	1.871773	2.127563	2.680043	3.381715	3.272394

Table II.2.1.1 Stepwise regression (first order)

```
## Step:  AIC=-61.03
## MORTALITY ~ NONWHITE + EDUC + SO2 + PRECIP
##
##             Df Sum of Sq    RSS     AIC
## <none>                   18.366 -61.031
## + NOX        1    0.4283 17.938 -60.447
## + POOR       1    0.3862 17.980 -60.306
## - EDUC       1    1.5853 19.951 -58.064
## - PRECIP     1    3.0516 21.417 -53.809
## - SO2        1    6.4080 24.774 -45.074
## - NONWHITE   1   11.2504 29.616 -34.361

##
## Call:
## lm(formula = MORTALITY ~ NONWHITE + EDUC + SO2 + PRECIP, data =
mortality_std)
##
## Coefficients:
## (Intercept)     NONWHITE         EDUC          SO2       PRECIP
##   3.322e-16    4.696e-01   -2.024e-01    3.593e-01    2.897e-01
```

Figure II.2.1.1 Residual plots for first order regression models

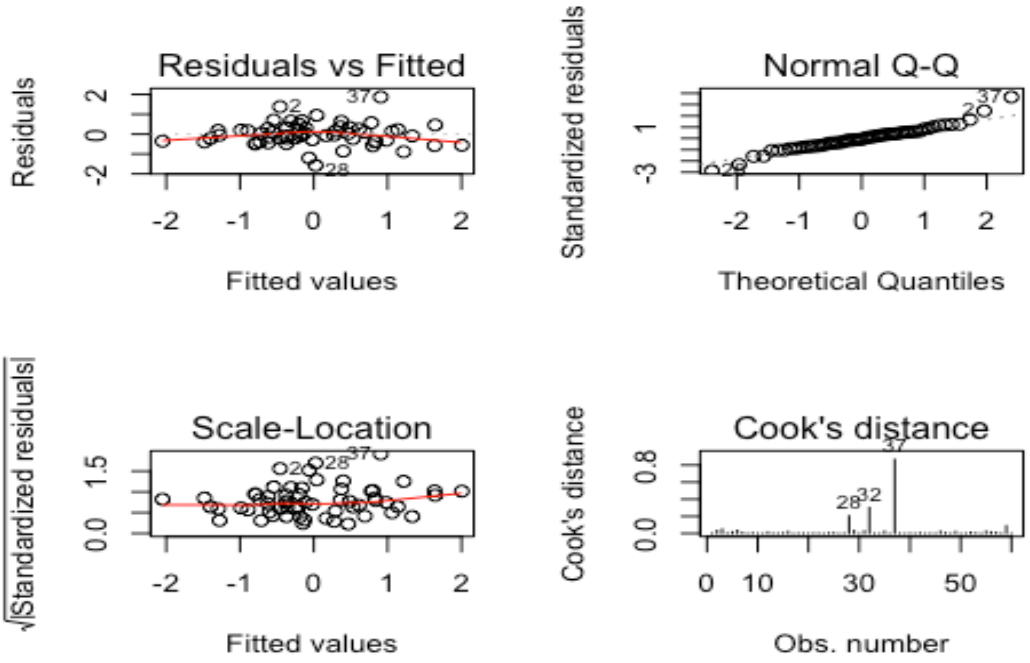


Table II.2.1.2 Stepwise regression (second order)

```
## Step:  AIC=-68.03
## MORTALITY ~ NONWHITE + EDUC + SO2 + PRECIP + NONWHITE:SO2 + EDUC:SO2
+
##     SO2:PRECIP
##
##                    Df Sum of Sq     RSS     AIC
## <none>                          14.788 -68.030
## + EDUC:NONWHITE     1   0.44327 14.345 -67.856
## + POOR              1   0.41002 14.378 -67.717
## - SO2:PRECIP        1   0.66771 15.456 -67.381
## + PRECIP:EDUC       1   0.32327 14.465 -67.356
## + PRECIP:NONWHITE   1   0.12835 14.660 -66.553
## + NOX               1   0.03115 14.757 -66.157
## - EDUC:SO2          1   1.07945 15.868 -65.803
## - NONWHITE:SO2      1   1.46619 16.255 -64.358

##
## Call:
## lm(formula = MORTALITY ~ NONWHITE + EDUC + SO2 + PRECIP +
NONWHITE:SO2 +
##     EDUC:SO2 + SO2:PRECIP, data = mortality_std)
##
## Coefficients:
##  (Intercept)      NONWHITE          EDUC           SO2        PRECIP
##      0.07058       0.47763      -0.19947       0.41399       0.28651
## NONWHITE:SO2      EDUC:SO2    SO2:PRECIP
##     -0.23873       0.16954       0.12004
```

Figure II.2.1.2 Residual plots for second order regression models

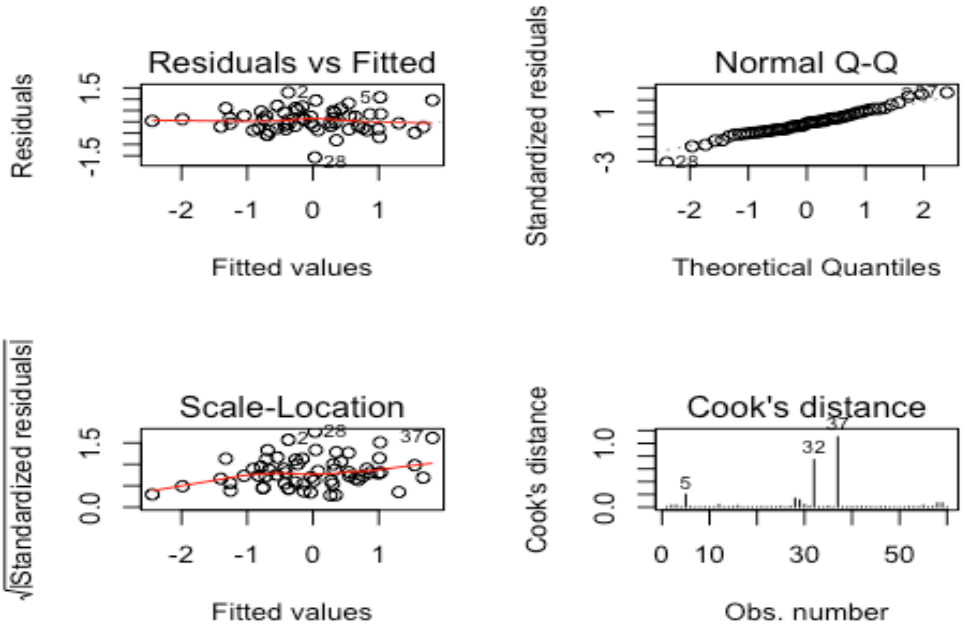


Figure II.2.2.1 Ridge regression lambda

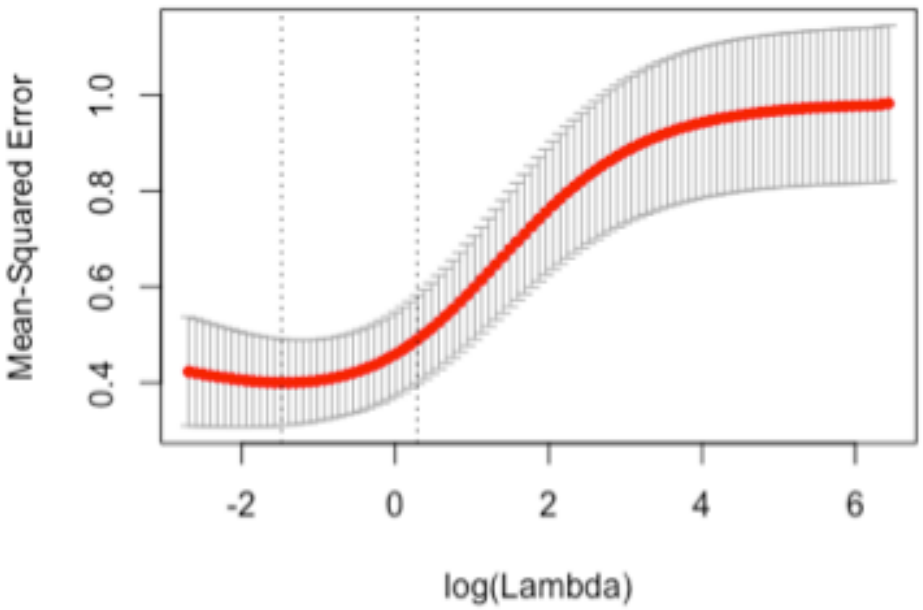


Figure II.2.2.2 Residual plot of ridge regression

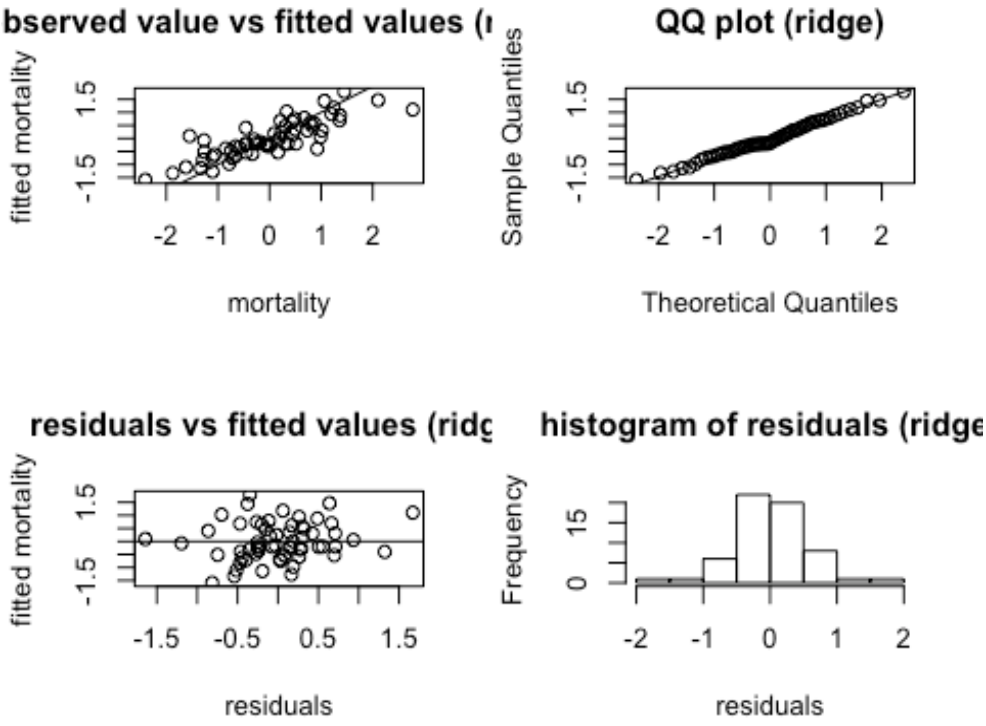


Figure II.2.3.1 Ridge regression lambda

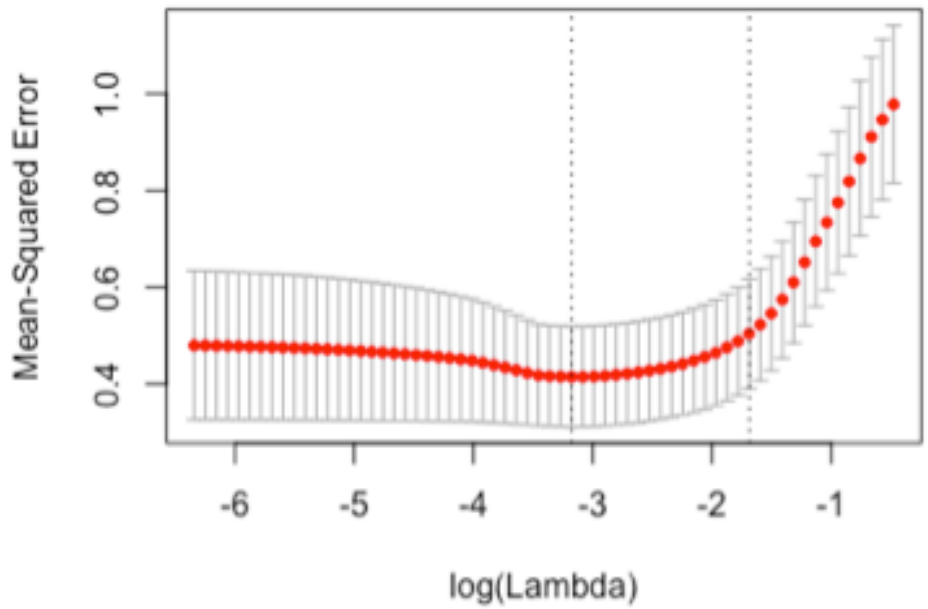


Figure II.2.3.2 Residual plot of lasso regression

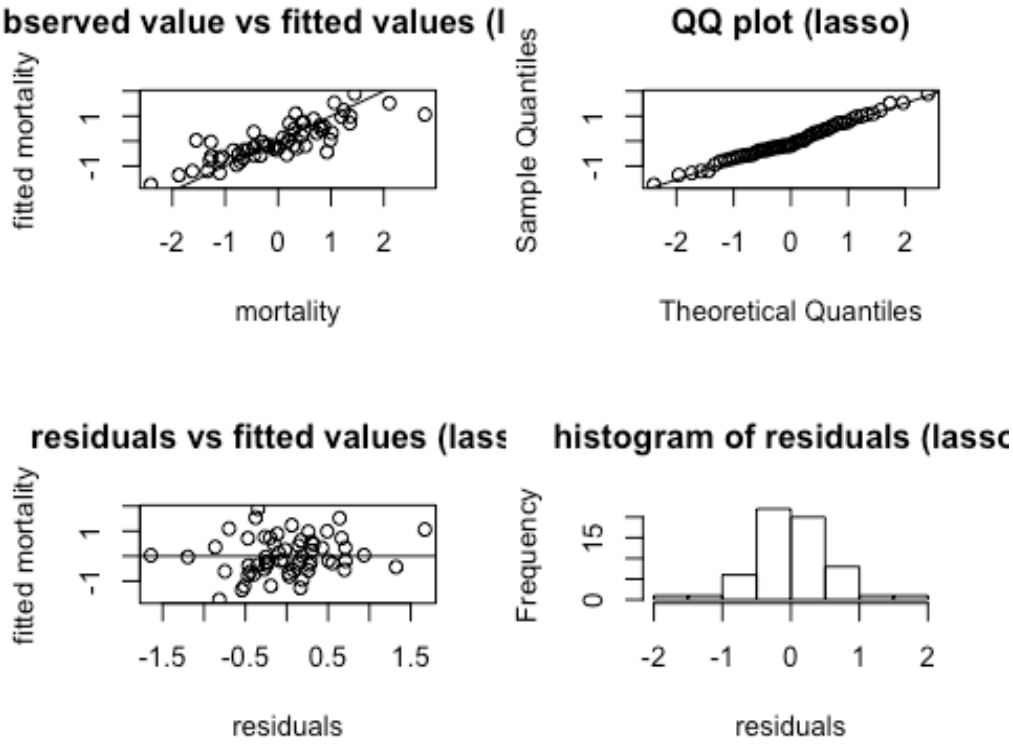


Table II.2.4.1 PLS

```
## Data:    X dimension: 60 6
## Y dimension: 60 1
## Fit method: kernelpls
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           1.008   0.6402   0.6428   0.6538   0.6440   0.6699
## adjCV        1.008   0.6375   0.6405   0.6476   0.6394   0.6630
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps
## X             35.48    64.84    76.44    88.02    97.12
## MORTALITY     64.10    67.32    70.10    70.40    70.54
```

Figure II.2.4.1 Residual plot of PLS

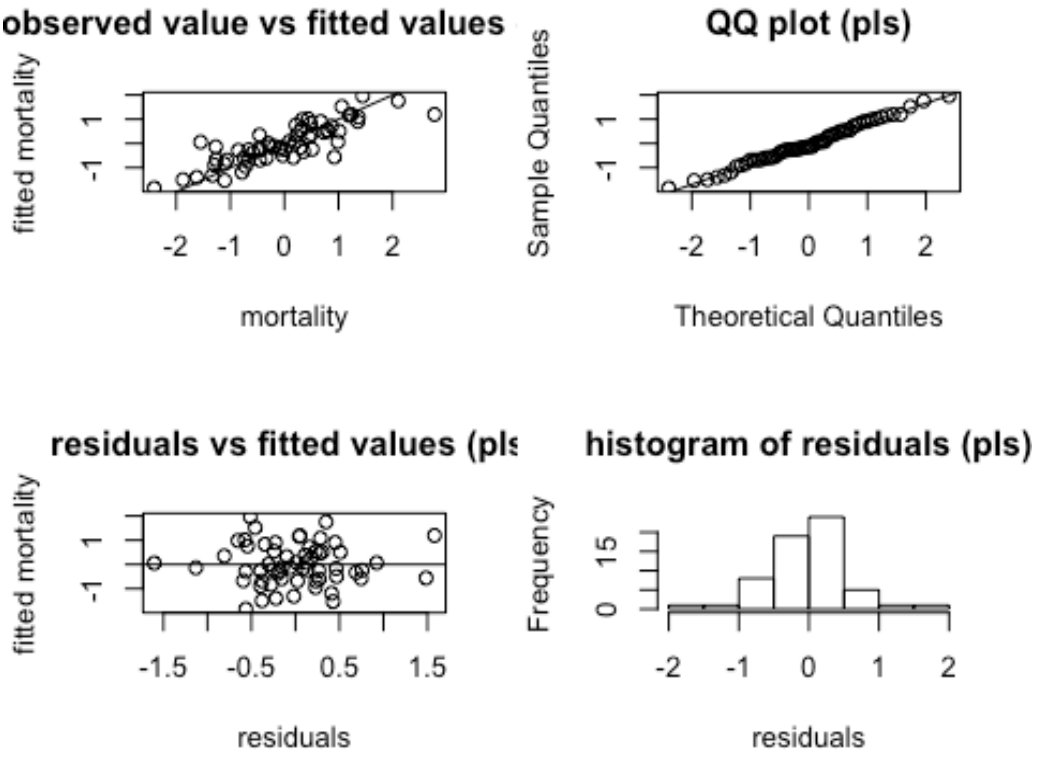


Table II.2.4.2 Loadings

```
## Loadings:
##          Comp 1 Comp 2 Comp 3 Comp 4
## PRECIP    0.432 -0.322  0.703 -0.545
## EDUC     -0.472  0.123  0.104  0.484
## NONWHITE  0.515         0.176  0.894
## POOR      0.492 -0.560 -0.301  0.111
## NOX       0.200  0.676 -0.603  0.292
## SO2       0.270  0.756 -0.286 -0.403
##
##                Comp 1 Comp 2 Comp 3 Comp 4
## SS loadings     1.029  1.465  1.073  1.590
## Proportion Var  0.172  0.244  0.179  0.265
## Cumulative Var  0.172  0.416  0.595  0.860
```

Figure II.4.1 Residual plots for final model

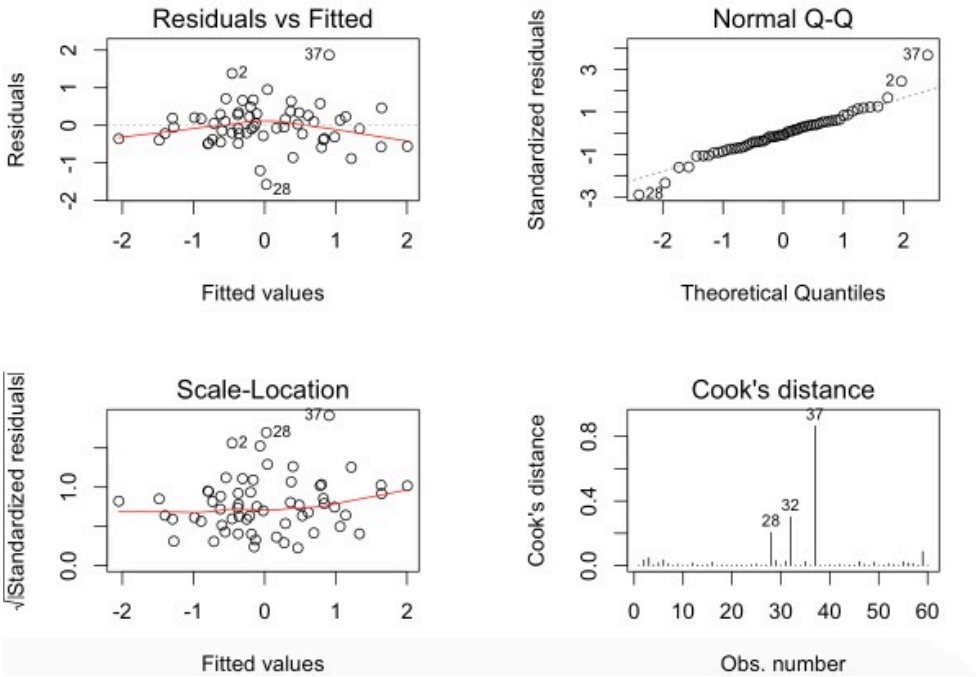


Figure II.4.2 Residual plots for final model after deleting outlier

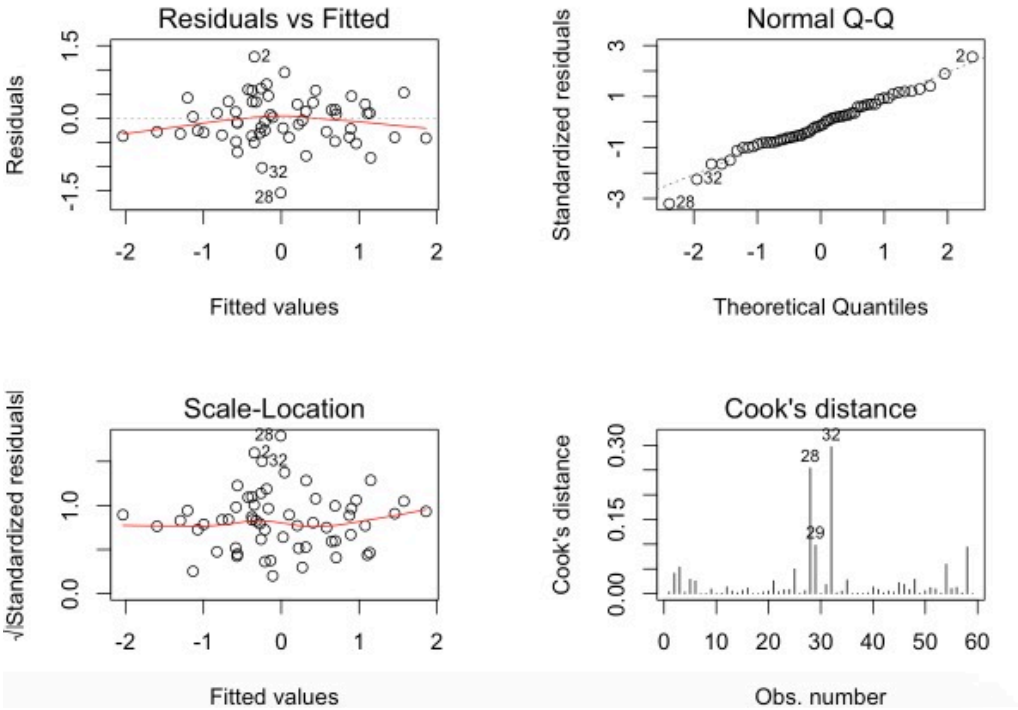


Table III.2.1 regression model of mortality to NOX and SO2

```
Call:
lm(formula = MORTALITY ~ NOX + SO2, data = mortality_std)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8259 -0.5532  0.0329  0.4833  3.6518

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.488e-16  1.202e-01   0.000   1.0000
NOX         -7.380e-03  1.781e-01  -0.041   0.9671
SO2          4.085e-01  1.781e-01   2.293   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.931 on 57 degrees of freedom
Multiple R-squared:  0.1625,    Adjusted R-squared:  0.1332
F-statistic: 5.531 on 2 and 57 DF,  p-value: 0.006375
```

CODE

```
# Part I data exploration

# read data
library(xlsx)
mortality <- read.xlsx("mortality.xls", sheetIndex = 1)
mortality <- mortality[, -8]
plot(mortality)
par(mfrow = c(3, 3))
sapply(colnames(mortality), function(x) hist(mortality[[x]], main = x))

# summary statistics
apply(mortality, 2, summary)

# single regression
single_regression <- lapply(colnames(mortality)[-7], function(x) {
 fit <- lm(as.formula(sprintf("MORTALITY ~ %s", x)), data = mortality_std)
 summary(fit)
})


# compare initial, log, square, square root
par(mfrow = c(3, 4))
sapply(colnames(mortality)[1:6], function(x) {
 hist(mortality[[x]], main = sprintf("Hist of %s", x), xlab = x)
 hist(log(mortality[[x]]), main = sprintf("Hist of log(%s)", x), xlab = sprintf("%s", x))
 hist(sqrt(mortality[[x]]), main = sprintf("Hist of sqrt(%s)", x), xlab = sprintf("%s", x))
 hist(mortality[[x]]^2, main = sprintf("Hist of square(%s)", x), xlab = sprintf("%s", x))
})

# do the transformation based on our discussion, log for NOX and SO2, sqrt for NONWHITE
mortality_new <- mortality
mortality_new$NONWHITE <- sqrt(mortality_new$NONWHITE)
mortality_new$NOX <- log(mortality_new$NOX)
mortality_new$SO2 <- log(mortality_new$SO2)

# standardize
mortality_std <- as.data.frame(scale(mortality_new))

# see the correlation
plot(mortality_std)
cor(mortality_std)
library(corrplot)
par(mfrow=c(1,1), mar=c(1,1,1,1))
corrplot(cor(mortality_std), method = "circle", pch = 1, tl.cex=0.75)
corrplot(cor(mortality_std), method = "number", pch = 1, tl.cex=0.75)


# vif
```

```
vif_regrssion <- sapply(colnames(mortality_std)[-7], function(x) {
 fit <- lm(as.formula(sprintf("%s ~ . - MORTALITY", x)), data = mortality_std)
 return(1/(1 - summary(fit)$r.squared))
})

# Additional: regress y only on pollution
lm_fit0 <- lm(MORTALITY ~ NOX + SO2, data = mortality_std)
summary(lm_fit0)
# not significant

#-------------------------------------------------------------------------------
# Part II stepwise regression

# first order initial
lm_fit1 <- lm(MORTALITY ~ 0 + ., data = mortality_std)
summary(lm_fit1)
par(mfrow = c(2, 2))
plot(lm_fit1, which = 1:4)

# first order stepwise
library(MASS)
lm_fit2 <- stepAIC(lm(MORTALITY ~ 1, data = mortality_std), scope = list(upper = lm_fit1),
                   direction = "both", k = 2)
lm_fit2 <- lm(formula = MORTALITY ~ 0 + NONWHITE + EDUC + SO2 + PRECIP, data =
mortality_std)
plot(lm_fit2, which = 1:4)

# we can see outliers, but to compare this model to subsequent methods easily, we do not
# drop them, also because this is not a severe proble

# second order initial
lm_fit3 <- stepAIC(lm(MORTALITY ~ 1, data = mortality_std),
                   scope = list(upper = lm(MORTALITY ~ 0 + . ^ 2, data = mortality_std)),
                   direction = "both", k = 2)
lm_fit3 <- lm(formula = MORTALITY ~ 0 + NONWHITE + EDUC + SO2 + PRECIP +
NONWHITE:SO2 +
                EDUC:SO2 + SO2:PRECIP, data = mortality_std)
plot(lm_fit3, which = 1:4)

# cross validation statistics
lm_fit2_cv <- 1 / 60 * sum((residuals(lm_fit2) / (1 - influence(lm_fit2)$hat))^2)
lm_fit3_cv <- 1 / 60 * sum((residuals(lm_fit3) / (1 - influence(lm_fit3)$hat))^2)
# cross validation 0.384, 0.394, so select linear model

#-------------------------------------------------------------------------------
# Part III PLS

library(pls)
par(mfrow = c(1,1))
set.seed(1)
pls_fit1 <- plsr(MORTALITY ~ 0 + ., 5, data = mortality_std, validation = 'CV')
```

```r
summary(pls_fit1)

# after the random seed is set, I see 4 components
set.seed(1)
pls_fit2 <- plsr(MORTALITY ~ 0 + ., 4, data = mortality_std, validation = 'CV')
summary(pls_fit2)
pls_fit2_coef <- pls_fit2$coefficients[,,4]
pls_fit2_fitted <- pls_fit2$fitted.values[,,4]
pls_fit2_residual <- mortality_std$MORTALITY - pls_fit2_fitted
loadings(pls_fit2)

# diagnostic plot
plot_diagnostic <- function(fitted_values, residual_values, string) {
 par(mfrow = c(2, 2))
 # the observed against the fitted values
 plot(mortality_std$MORTALITY, fitted_values,
     main = sprintf("observed value vs fitted values (%s)", string),
     xlab = "mortality", ylab = 'fitted mortality')
 abline(a = 0, b = 1)
 # qqnorm
 qqnorm(fitted_values, main = sprintf("QQ plot (%s)", string))
 qqline(fitted_values)
 # residuals against fitted
 plot(residual_values, fitted_values,
     main = sprintf("residuals vs fitted values (%s)", string),
     xlab = "residuals", ylab = 'fitted mortality')
 abline(h = 0)
 # histogram of residuals
 hist(residual_values, main = sprintf('histogram of residuals (%s)', string),
     xlab = 'residuals')
 par(mfrow = c(1, 1))
}

plot_diagnostic(pls_fit2_fitted, pls_fit2_residual, "pls")

# cross-validation result
summary(pls_fit2)

#-----------------------------------------------------------------------------
# Part IV ridge

set.seed(1)
ridge_fit1 <- cv.glmnet(as.matrix(mortality_std[, -7]), mortality_std$MORTALITY,
                  alpha = 0, intercept = F)
plot(ridge_fit1)
ridge_lambda <- ridge_fit1$lambda.min
ridge_coef <- coef(ridge_fit1, s = ridge_lambda)@x
ridge_fit1_fitted <- as.matrix(mortality_std[, -7]) %*% ridge_coef
ridge_fit1_residual <- mortality_std$MORTALITY - ridge_fit1_fitted

# ridge_vif
```

```r
R <- as.matrix(t(mortality_std[, -7])) %*% as.matrix(mortality_std[, -7])
A <- solve(R + diag(ridge_lambda, 6)) %*% R %*% solve(R + diag(ridge_lambda, 6))
ridge_vif <- diag(solve(A)) * diag(A)
# only a little smaller, since our lambda is very small

# diagnostic plot
plot_diagnostic(ridge_fit1_fitted, ridge_fit1_residual, "ridge")

# cross validation result
ridge_fit1_cv <- ridge_fit1$cvm[which(ridge_fit1$lambda == ridge_lambda)]

#------------------------------------------------------------------------------
# Part V lasso

set.seed(1)
lasso_fit1 <- cv.glmnet(as.matrix(mortality_std[, -7]), mortality_std$MORTALITY,
                        alpha = 1, intercept = F)
plot(lasso_fit1)
lasso_lambda <- lasso_fit1$lambda.min
lasso_coef <- coef(lasso_fit1, s = lasso_lambda)@x
# someone is set to zero, see here
lasso_coef <- c(lasso_coef[1:3], 0, lasso_coef[4:5])
lasso_fit1_fitted <- as.matrix(mortality_std[, -7]) %*% lasso_coef
lasso_fit1_residual <- mortality_std$MORTALITY - ridge_fit1_fitted

# diagnostic plot
plot_diagnostic(lasso_fit1_fitted, lasso_fit1_residual, "lasso")

# cross validation result
lasso_fit1_cv <- lasso_fit1$cvm[which(lasso_fit1$lambda == lasso_lambda)]

#------------------------------------------------------------------------------
# Part VI linear model revisit (validation)

# we know the outliers may be 37.
lm_fit4 <- lm(formula = MORTALITY ~ 0 + NONWHITE + EDUC + SO2 + PRECIP,
          data = mortality_std[-37, ])
par(mfrow = c(2, 2))
plot(lm_fit4, which = 1:4)
lm_fit4_cv <- 1 / 60 * (sum((residuals(lm_fit4) / (1 - influence(lm_fit4)$hat))^2) +
                    sum(mortality_std[37, 7] - predict(lm_fit4, mortality_std[37, ])))
```