R code, STA 108 Project

**Problem 1:**

>deathrate = read.table("~/Desktop/STA 108/deathrate\_dataset.txt", header=TRUE)

>reg=lm(death\_rate~1, data=deathrate)

>step(reg,scope=death\_rate~precip+jan\_temp+jul\_temp+age\_65\_plus+household+school+kitchen+pop\_density+nonwhite+office+inc\_30k+HC\_index+NOx\_index+SOx\_index+atmos, direction="both")

>best=lm(formula = death\_rate ~ nonwhite + school + jan\_temp + >pop\_density + precip + SOx\_index + jul\_temp, data = deathrate)

>summary(best)

Residuals:

Min 1Q Median 3Q Max

-84.679 -18.451 0.424 17.739 93.278

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.167e+03 1.174e+02 9.943 1.27e-13 \*\*\*

nonwhite 4.528e+00 6.448e-01 7.023 4.55e-09 \*\*\*

school -1.572e+01 6.173e+00 -2.547 0.013849 \*

jan\_temp -1.481e+00 4.064e-01 -3.644 0.000619 \*\*\*

pop\_density 8.076e-03 3.471e-03 2.327 0.023905 \*

precip 1.681e+00 5.774e-01 2.912 0.005276 \*\*

SOx\_index 1.723e-01 8.374e-02 2.057 0.044683 \*

jul\_temp -2.143e+00 1.198e+00 -1.790 0.079311 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

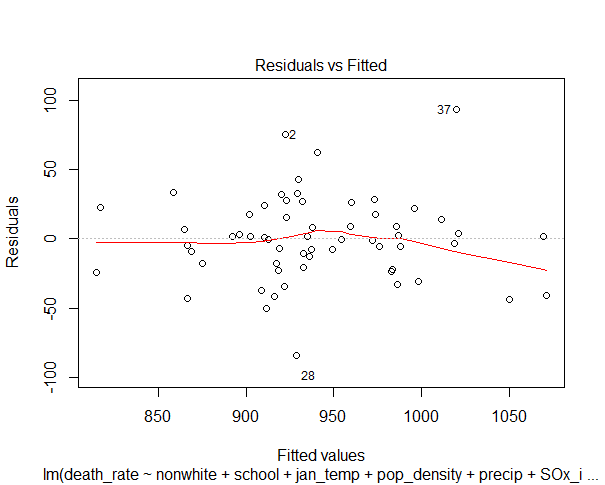
Residual standard error: 32.43 on 52 degrees of freedom

Multiple R-squared: 0.7604, Adjusted R-squared: 0.7281

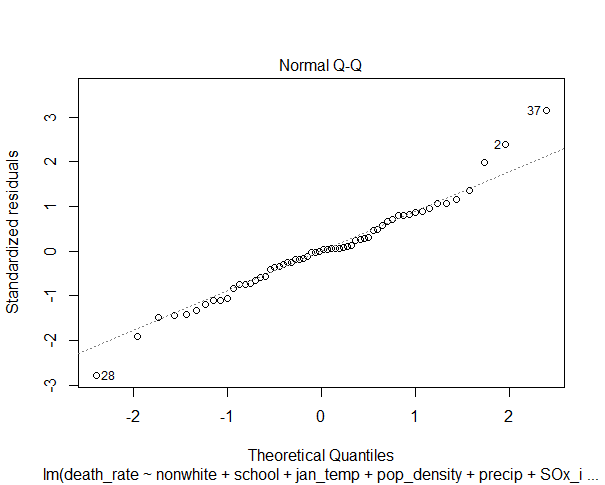
F-statistic: 23.58 on 7 and 52 DF, p-value: 4.659e-14

From the summary statistics, the variables school, jan\_temp, and jul\_temp are all negative. It is notable that the p-value for jul\_temp is greater than 0.05, while all other variables are less, but with our cp comparison, we determined that jul\_temp is still a significant variable in our model. This implies that on average, an increase in each respective variable variable, while holding all other variables constant, would lead to a decrease deathrate. On the other hand, all other variables listed above excluding jan\_temp, school, and jul\_temp are associated with an average increase in death rate if each corresponding variable increases while holding all other variables constant. Also, all the p-values for the variables are small, meaning they are statistically significant in explaining variability in deathdate. Seeing that the r-squared is 0.7604 is a good indicator that these variables approximately estimate 76.04% of the variability in death rate.

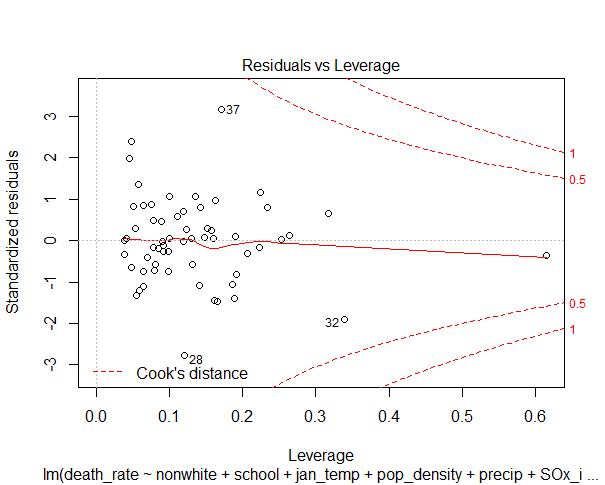
>plot(best)



Looking at the residual v.s fitted plot, we can see that the assumption of equal variance and linearity are mostly is satisfied. The three notable outliers might cause the distortion of data.



Looking at the qq plot we can see the assumption of normality is in question due to the skewed tail ends of the plot.



Looking at the residual v.s leverage plot, once again, we see that variance is is relatively equal. We can see that plots 28,38,37 are outliers in the data set but not extreme enough to have affected the data.

Update on March 12 14:29 by Peggy

#Use other criterion to check

library(leaps)

subsets=regsubsets(death\_rate~precip+jan\_temp+age\_65\_plus+household+school+kitchen+pop\_density+nonwhite+office+inc\_30k+HC\_index+NOx\_index+SOx\_index+atmos+jul\_temp, data=deathrate, nbest=1)

c1=summary(subsets)$which # Get the subsets used

c2=summary(subsets)$cp # Get Cp for Best Subsets

c3=summary(subsets)$bic # Get SBC for Best Subsets

p = rowSums(summary(subsets)$which) # Retrieve p = # betas

summary(subsets)$bic + 2\*log(nrow(mtcars))\*p - 2\*p # Get AIC for Best Subsets, by SBC w/different penalty

cbind("p"=p, "cp"=c2,"sbc"=c3,"adjRsp"=summary(subsets)$adjr2,c1)

#!!!!!!analysis

#best1 is considered a better model obtained from other criterion

best1=lm(formula = death\_rate ~ nonwhite + school + jan\_temp + pop\_density + precip +HC\_index+NOx\_index, data = deathrate)

plot(best1)

summary(best1)

#boxcox on best1

library(MASS)

boxcox(best1y)

#Transform Y, since best1 is not normal distributed

best1y=lm(formula = sqrt(death\_rate)~ nonwhite + school + jan\_temp + pop\_density + precip +HC\_index+NOx\_index, data = deathrate)

#Correlation between variables

pairs(death\_rate~nonwhite + school + jan\_temp + pop\_density + precip + NOx\_index + HC\_index, data = deathrate)

summary(best1y)

cor(cbind(deathrate$nonwhite,deathrate$school ,deathrate$jan\_temp, deathrate$pop\_density ,deathrate$precip , deathrate$NOx\_index,deathrate$HC\_index), y=NULL)

plot(best1y)

#Interaction between HC and So

best1i=lm(formula = sqrt(death\_rate)~ nonwhite + school + jan\_temp + pop\_density + precip +HC\_index+NOx\_index+HC\_index:NOx\_index, data = deathrate)

plot(best1i)

summary(best1i)

#The p-value is 0.408673, so we don't need to do the interaction

#Focus back on our model obtained from Stepwise function

plot(best)

#Found non-normalbilty, do Y transformation

library(MASS)

boxcox(best)

#lamda is 0.5, consider to use square root

besty=lm(sqrt(death\_rate) ~ nonwhite + school + jan\_temp + pop\_density + precip + SOx\_index + jul\_temp, data = deathrate)

plot(besty)

boxcox(besty)

#However, the model is still not normal distributed, consider transformation on X

#use pairs to the linearity between death\_rate and variables

pairs(death\_rate~nonwhite + school + jan\_temp + pop\_density + precip + sqrt(SOx\_index) + jul\_temp, data = deathrate)

#It is obvious that Y~Sox\_index

bestx=lm(sqrt(death\_rate)~ nonwhite + school + jan\_temp + pop\_density + precip + sqrt(SOx\_index) + jul\_temp, data = deathrate)

#Check out the correlation of the variables

cor(cbind(deathrate$nonwhite,deathrate$school ,deathrate$jan\_temp, sqrt(deathrate$pop\_density) ,deathrate$precip , sqrt(deathrate$SOx\_index) ,deathrate$jul\_temp), y=NULL)

lm(sqrt(death\_rate)~ nonwhite + school + jan\_temp + pop\_density + precip + log(SOx\_index) + jul\_temp, data = deathrate)

#Use sqrt(SOx\_index) rathen than log(SOx\_index), because it will cause #37 observation out of cook's distance

#Conside doing transformation on pop\_density

plot(death\_rate~(pop\_density), data=deathrate)

plot(death\_rate~sqrt(pop\_density), data=deathrate)

summary(lm(sqrt(death\_rate)~ nonwhite + school + jan\_temp + sqrt(pop\_density) + precip + sqrt(SOx\_index) + jul\_temp, data = deathrate))

#It seems useless to do transformation on poo\_density.

#Finally, we decide to use bestx as our final model

model=bestx

summary(model)

plot(model)

**Final model:**

**lm(sqrt(death\_rate)~ nonwhite + school + jan\_temp + pop\_density + precip + sqrt(SOx\_index) + jul\_temp, data = deathrate)**

**Problem 2:**

> summary(model2)

Call:

lm(formula = death\_rate ~ school, data = deathrate\_dataset)

Residuals:

Min 1Q Median 3Q Max

-151.71 -36.69 2.41 43.84 124.91

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1352.893 91.408 14.801 < 2e-16 \*\*\*

school -37.598 8.306 -4.527 3.03e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.93 on 58 degrees of freedom

Multiple R-squared: 0.2611, Adjusted R-squared: 0.2483

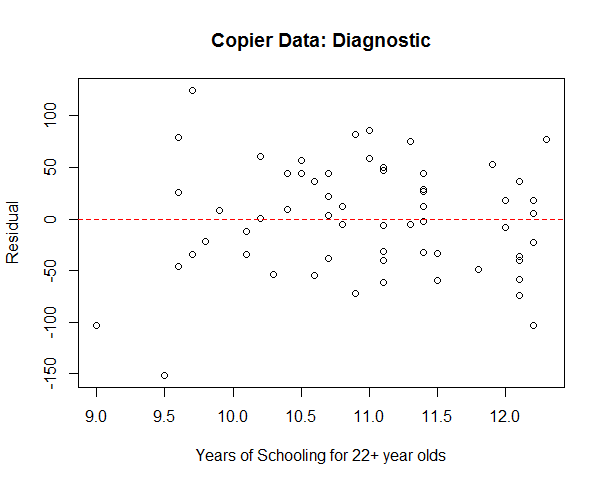
F-statistic: 20.49 on 1 and 58 DF, p-value: 3.027e-05

Using p-value test, we can see that 3.03e-05 is less than 0.05 which means the number of schooling over 22 is significant in explaining variation in death rate.

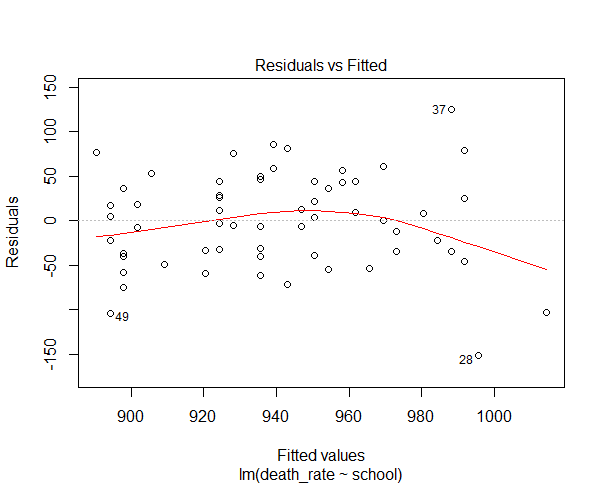
> plot(deathrate\_dataset$school, model2$residuals,, main = "Copier Data: Diagnostic",

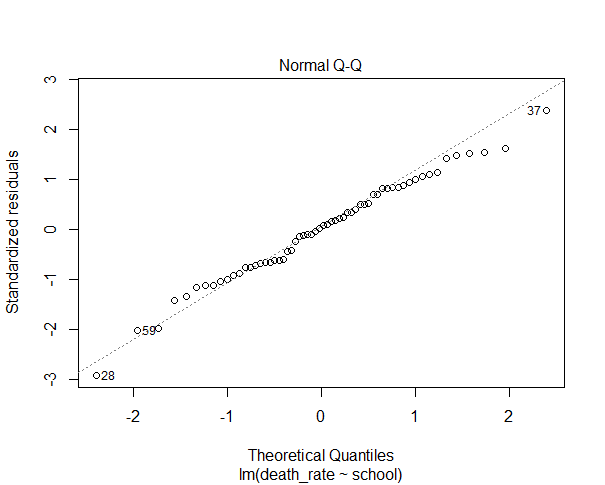
+ xlab = "Years of Schooling for 22+ year olds", ylab = "Residual")

> abline(h = 0, col = "red", lty=2)



> plot(model2, which = c(1:2))





> anova(model2)

Analysis of Variance Table

Response: death\_rate

Df Sum Sq Mean Sq F value Pr(>F)

school 1 59595 59595 20.492 3.027e-05 \*\*\*

Residuals 58 168680 2908

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(deathrate\_dataset$school, deathrate\_dataset$death\_rate, main = "Death Rate Data", xlab = "school (years)", ylab = "death rate")

> abline(model2, col = "red")



> -37.598 + c(-1,1)\*qt(0.95, 58)\*8.306

95% C.I for change in death rate when the average years of school for 22+ year olds increases by one (-51.48192 -23.71408)

**Problem 3:**

> #model with three pollution index variables

> model3=lm(formula = death\_rate ~ HC\_index + NOx\_index + SOx\_index, data = deathrate\_dataset)

> summary(model3)

Call:

lm(formula = death\_rate ~ HC\_index + NOx\_index + SOx\_index, data = deathrate\_dataset)

Residuals:

Min 1Q Median 3Q Max

-100.292 -33.352 -5.458 37.506 172.586

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 924.3782 9.0897 101.695 <2e-16 \*\*\*

HC\_index -1.5064 0.6255 -2.408 0.0193 \*

NOx\_index 2.7082 1.3053 2.075 0.0426 \*

SOx\_index 0.2226 0.1771 1.257 0.2140

---

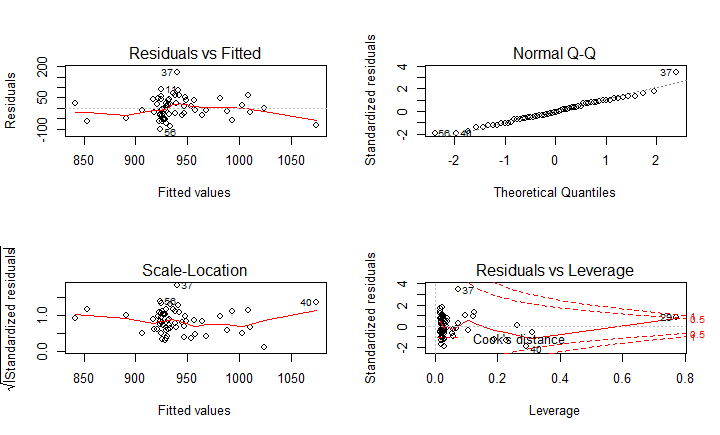
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.3 on 56 degrees of freedom

Multiple R-squared: 0.329, Adjusted R-squared: 0.293

F-statistic: 9.152 on 3 and 56 DF, p-value: 5.05e-05

> plot(model3)



> #standard error of parameter of the hydrocarbon pollution index

> sehydrocarbon=0.6255

> #standard error of parameter of the nitric oxide pollution index

> senitric=1.3053

> #value of parameter of the hydrocarbon pollution index

> bHC\_index=-1.5064

> #value of parameter of the nitric oxide pollution index

> bNOx\_index=2.7082

**Interpretation:**

F-test:

we reject the null hypothesis and claim that the hydrocarbon pollution index,the nitric oxide pollution index, and the sulfur dioxide pollution index are jointly significant at 0.05 level of significance.

T-test:

We reject the hypothesis and claim that the hydrocarbon pollution index and the nitric oxide pollution index are separately and statistically significant at 0.1 level of significance. However, we reject the hypothesis that the nitric oxide pollution index is not statistically significant in this model.

Confidence interval of parameters:

>cbind(lower=-1.5064-qt(1-0.05/2,56)\*sehydrocarbon,upper=-1.5064+qt(1-0.05/2,56)\*sehydrocarbon)

lower upper

[1,] -2.759427 -0.2533729

>cbind(lower=2.7082-qt(1-0.05/2,56)\*senitric,upper=2.7082+qt(1-0.05/2,56)\*senitric)

lower upper

[1,] 0.09336989 5.32303

-interpretation: we are 95% confident that the true value of parameter corresponding to hydrocarbon pollution index falls in [-2.759427,-0.2533729], and the true value of the parameter corresponding to the nitric oxide pollution index falls in [0.09336989,5.32303]

**# Now i am going to construct a model without SOx\_index**

> model3.1=lm(death\_rate~HC\_index+NOx\_index, data=deathrate\_dataset)

> summary(model3.1)

Call:

lm(formula = death\_rate ~ HC\_index + NOx\_index, data = deathrate\_dataset)

Residuals:

Min 1Q Median 3Q Max

-105.833 -35.312 -1.773 37.367 157.370

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 929.9486 7.9766 116.585 < 2e-16 \*\*\*

HC\_index -2.0936 0.4180 -5.008 5.63e-06 \*\*\*

NOx\_index 3.9796 0.8294 4.798 1.19e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.57 on 57 degrees of freedom

Multiple R-squared: 0.3101, Adjusted R-squared: 0.2859

F-statistic: 12.81 on 2 and 57 DF, p-value: 2.548e-05

> library(leaps)

> ll=regsubsets(death\_rate~HC\_index+NOx\_index+SOx\_index,data=deathrate\_dataset, nbest=2)

> ci=summary(ll)$which

> c1=summary(ll)$which

> c2=summary(ll)$cp

> c2=summary(ll)$bic

> c2=summary(ll)$cp

> c3=summary(ll)$bic

> p = rowSums(summary(ll)$which)

> summary(ll)$bic + 2\*log(nrow(deathrate\_dataset))\*p - 2\*p

1 1 2 2 3

8.562932 18.652065 8.580070 11.353960 17.193609

> c4=summary(ll)$bic + 2\*log(nrow(deathrate\_dataset))\*p - 2\*p

> cbind("p"=p, "cp"=c2,"SBC"=c3,"adjRsq" = summary(ll)$adjr2,"AIC"=c4,c1)

p cp SBC adjRsq AIC (Intercept) HC\_index NOx\_index SOx\_index

1 2 12.324941 -3.814446 0.16719672 8.562932 1 0 0 1

1 2 24.836378 6.274687 0.01469654 18.652065 1 1 0 0

2 3 3.580076 -9.985998 0.28585149 8.580070 1 1 1 0

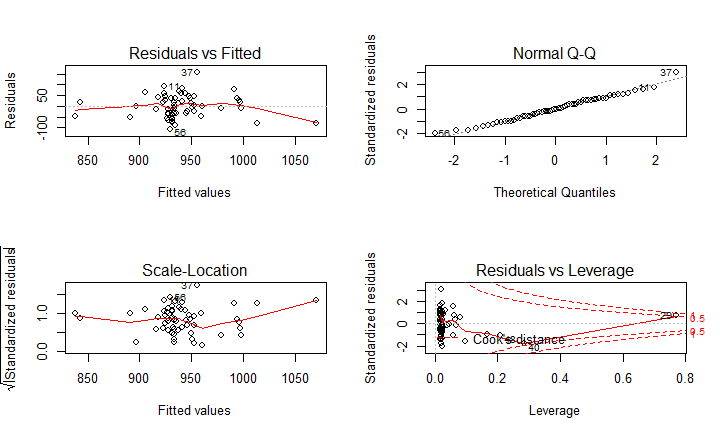
2 3 6.304583 -7.212108 0.25206024 11.353960 1 1 0 1

3 4 4.000000 -7.561148 0.29304600 17.193609 1 1 1 1

#pick p=3, explanatory variables include HC\_index and NOx\_index

By comparing the cp value of each model, we found that the model with variables HC\_index and NOx\_index SOx\_ has the closest cp value to p value. It also has the lowest SBC value. This model has largest adjusted value compared to other models. Although its AIC value is slightly higher than the model with one explanatory variable SOx\_index.

plot(model3.1)



* Observation 29,Los Angeles, Long Beach, CA has cook distance larger than 0.5. Compared to other cities, this city has relatively lower average annual precipitation and extremely high hydrocarbon pollution index.

geography

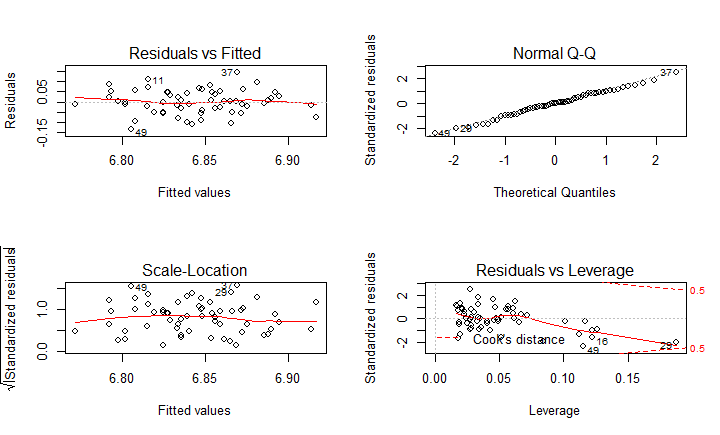
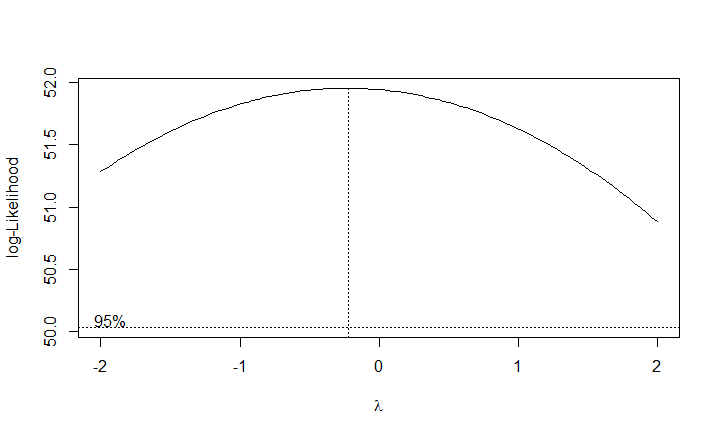
discuss model with and without the observation

interpret it with special geographic features.

Transformation:

> library(MASS)

> boxcox(model3.1)



> model3.3=lm(log(death\_rate)~log(HC\_index) + log(NOx\_index), data=deathrate\_dataset)

> plot(model3.3)

> summary(model3.3)

Call:

lm(formula = death\_rate ~ log(HC\_index) + log(NOx\_index), data = deathrate\_dataset)

Residuals:

Min 1Q Median 3Q Max

-114.832 -34.193 0.556 42.230 149.158

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 942.08 19.46 48.407 < 2e-16 \*\*\*

log(HC\_index) -63.97 19.71 -3.245 0.001966 \*\*

log(NOx\_index) 75.35 19.61 3.843 0.000308 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.75 on 57 degrees of freedom

Multiple R-squared: 0.2238, Adjusted R-squared: 0.1966

F-statistic: 8.219 on 2 and 57 DF, p-value: 0.0007304

#interpretation:

The original model fails the assumptions of normality, independent variance and it has a outlier with Cook’s distance larger than 0.5, which means it has strong influence on the slopes of the regression model. After the command “boxcox” is applied, the result indicates that a logarithm transformation should be applied to the model. After the transformation, the model performs fairly well.

> anova(lm(death\_rate~HC\_index,data=deathrate\_dataset))

Analysis of Variance Table

Response: death\_rate

Df Sum Sq Mean Sq F value Pr(>F)

HC\_index 1 7167 7167.1 1.88 0.1756

Residuals 58 221109 3812.2

> anova(lm(death\_rate~HC\_index+NOx\_index,data=deathrate\_dataset))

Analysis of Variance Table

Response: death\_rate

Df Sum Sq Mean Sq F value Pr(>F)

HC\_index 1 7167 7167 2.5939 0.1128

NOx\_index 1 63612 63612 23.0220 1.194e-05 \*\*\*

Residuals 57 157496 2763

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm(death\_rate~NOx\_index,data=deathrate\_dataset))

Analysis of Variance Table

Response: death\_rate

Df Sum Sq Mean Sq F value Pr(>F)

NOx\_index 1 1476 1475.9 0.3774 0.5414

Residuals 58 226800 3910.3

> anova(lm(death\_rate~NOx\_index+HC\_index,data=deathrate\_dataset))

Analysis of Variance Table

Response: death\_rate

Df Sum Sq Mean Sq F value Pr(>F)

NOx\_index 1 1476 1476 0.5341 0.4679

HC\_index 1 69303 69303 25.0817 5.631e-06 \*\*\*

Residuals 57 157496 2763

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

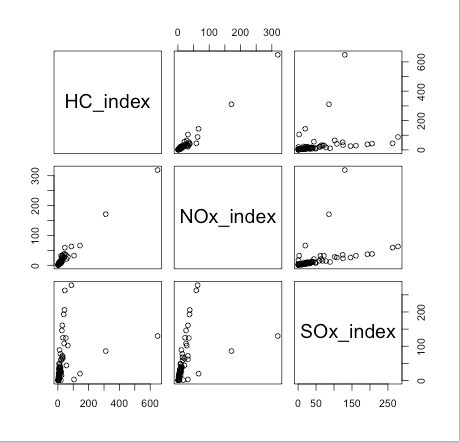
HC\_index=deathrate$HC\_index

NOx\_index=deathrate$NOx\_index

SOx\_index=deathrate$SOx\_index

#Make a **scatter plot** of the data.

pairs(cbind(HC\_index,NOx\_index,SOx\_index))



Interpretation: From the plot, there is a correlation between the hydrocarbon pollution index and the nitric oxide pollution index.

>cor(cbind(HC\_index, NOx\_index, SOx\_index), y=NULL)

HC\_index NOx\_index SOx\_index

HC\_index 1.0000000 0.9840337 0.2822963

NOx\_index 0.9840337 1.0000000 0.4100839

SOx\_index 0.2822963 0.4100839 1.0000000

Interpretation:

correlation…….

> #Check the **Variance Inflation Factor** for the test of multicollinearity

> vif(model3)

HC\_index NOx\_index SOx\_index

71.393408 78.987159 2.718799

> vif(lm(formula = death\_rate ~ NOx\_index + HC\_index, data = deathrate\_dataset))

NOx\_index HC\_index

31.56791 31.56791

Interpretation: Since vif greater than 10, it indicates multicollinearity

> vif(lm(formula = death\_rate ~ NOx\_index + SOx\_index, data = deathrate\_dataset))

NOx\_index SOx\_index

1.202167 1.202167

> vif(lm(formula = death\_rate ~ HC\_index + SOx\_index, data = deathrate\_dataset))

HC\_index SOx\_index

1.086592 1.086592

Interpretation: If we drop either NOx\_index or HC\_index, vif is longer than 10, which indicates that the predictors included in the models are not highly related, and it will not cause problem if we include both of them in the model.

**Problem 4:**

Obtained the information about San Francisco from the data set.

Use predict function to obtain the confidence (prediction?) interval with 95% confidence from the final model.

with 95% confidence, the death rate of San Francisco is from 877.6386 to 942.3976 per 100,000.

The death rate of San Francisco in our data set is 911 which is in the confidence interval.

So we would say the model can predict the death rate of San Francisco very well.

R code:

> dataSF=data.frame("nonwhite"=13.7,"school"=12.2, "jan\_temp"= 48, "pop\_density"=4253, "precip"=18 , "SOx\_index"=86, "jul\_temp"= 63)

> a=predict(model, dataSF, interval="confidence")

> #Transfer the result to its original unit

> a^2

fit lwr upr

1 909.7299 877.6386 942.3976