Assignment 3

Last name: Du
First name: Min
Student ID: 1002602230
Course section: STA302H1F-Summer 2017

Introduction

Observational studies have suggested that precipitation,monthly average temperature, family status and other environmental fators might affect The death rate. Study subjects were variety of places in 1960. The aim of my data analysis is to find the variables that are important in predicting the response variables(The Death Rate) and construct a regression fits the data.

The data set consists of 60 observations on 9 variables(8 predictor variable and 1 response variable.The variables are:
AVP: the average annual precipitation (inch)
MPH: the number of members per household in 1960
YSP22: the number of years of schooling for persons over 22 in 1960
HFK: the number of households with fully equipped kitchens
NFL3: the number of families with an income less than us dollar 3000
SDP: the relative pollution of Sulfur Dioxide
HP: relative pollution potential of hydrocarbons
DA: percent relative humidity, annual average at 1pm.
Response variable: The death rate

I expect The Death rate to be strongly related to HFK(the number of households with fully equipped kitchens), NFL3(poor families),SDP(the relative pollution of Sulfur Dioxide) and DA(percent relative humidity, annual average at 1pm). Slightly related to AVP(the average annual precipitation),HP(relative pollution potential of hydrocarbons).Do not have linear regression with MPH(the number of members per household), YSP22(the number of years of schooling for persons over 22).

Analysis

Regression between Predict Variables and Response Variable(8 Simple linear regressions, Y_i=b_0+b_1x_1,Death Rate is Y_i)(code given in appendix)

list of table	p-value for H_0:β_1=0
AVP	3.22e-05
MPH	0.00507
YSP22	3.02e-05
HFK	0.000672
NFL3	0.001124
SDP	0.000692
HP	0.1755
DA	0.6819

According to above r-code. MPH,HFK,NFL3,SDP has moderate evidence of relationship with response variable. AVP and YSP22 has strong evidence of relationship with response variable.There is no evidence that the coefficients of HP and DA are different from 0.

First Mutiple Regression(Full Model)(code given in appendix)

list of table	estmate	std.error	t-value	Pr(>
(Intercept)	648.79956	267.29044	2.427	0.01878
AVP	2.50851	0.81884	3.063	0.00349
MPH	83.85424	49.00100	1.711	0.09311
YSP22	-1.88113	9.31822	-0.202	0.84082
HFK	-1.83788	1.77717	-1.034	0.30594
NFL3	1.59976	2.06984	0.773	0.44316
SDP	0.47417	0.10173	4.661	2.3e-05
HP	0.03531	0.08540	0.413	0.68099
DA	0.75991	1.05024	0.724	0.47264

According to the above table, there is strong evidence that the coefficient of SDP is non-zero for a model including all of the other predictor variables.

There is moderate evidence that the coefficients of AVP is non-zero for a model including all of the other predictor variables.

There is weak evidence that the coefficients of MPH is non-zero for a model including all of the other predictor variables.

There is no evidence that the coefficients of YSP22, HFK, NFL3, HP and DA are different from 0.

From this model, It seems that important predictors are SDP(the relative pollution of Sulfur Dioxide),AVP(the average annual precipitation) and MPH(Household size in1960). We will try another mutiple regression by removing YSP22, HFK, NFL3, HP and DA.

Second Mutiple Regression After Removing YSP22, HFK, NFL3, HP and DA(Reduced Model)(code given in appendix)

list of table	estmate	std.error	t-value	Pr(>
(Intercept)	457.25325	136.88389	3.340	0.00149
AVP	3.12122	0.58821	5.306	1.98e-06
MPH	104.53931	43.17468	2.421	0.01873
SDP	0.47141	0.08938	5.274	2.22e-06

(1)According to the above table, we can see all of their p-values are smaller than $\alpha = 0.05$, which means we have evidence to show all of them have linear regression with the Death Rate in this model.

(2)Taking ANOVA of both full and reduced moedels.H_0: There is no difference between full model and reduced model. We get Pr(>F)=0.382, which means full model and redueced model are different. Reduced model is better.(code given in appendix)

(3)According to correlation plots between 9 variables (code given in appendix), we can see the first 6 varaibles have strong correlation between each other.

(4)According to correlation plots between 9 variables, COR(AVP,MPH)=0.26,COR(AVP,SDP)=-0.1069,COR(MPH,SDP)=-0.004.There is a weak postive correlation between AVP and MPH, no correlation between AVP & SDP, MPH & SDP.

(5)Taking linear regression between AVP and SDP, we get mutiple R squared=0.1143, VIF =1.00, which means they are linearly independent. Same as AVP and MPH, R squared=0.0694, VIF =1.00; MPH and SDP,R squared=1.668e-05, VIF =1.00.(code all given in appendix) All of them are linearly independent to each other.

(6)Double check this mutiple linear regression for Gauss Markov Assumption. (plot given in appendix).

Regression is linear.

Error terms are independent.(Because error terms are relatively uncorrelated.)

Error terms relatively have a constant variance.

Error terms are normally distributed.(Because the normal QQ plot is a straight line.)

No outliers.(No points are far away from the line in the plot.)

No any important predictor variables is ommitted from the model.

The above 6 points illustrate that reduced model is a good linear regression model, which is better than the full model.

Conclusion

For the full mutiple regression model, we find that most of the variables are greater than 0.05, then we remove the variables that are apparently greater than 0.05. Thus, we get our reduced model. According to the table of the reduced model, all the p-values are smaller than 0.05, also according to ANOVA of both full and reduced model, we get result that there is difference between reduced model and full model, which illustrates again the reduced model is better than the full model. Moreover, we find that AVP,SDP,MPH almost have no correlation between eath other, and they mutually linearly independent. Finally, I double check the reduced model for GM assumption.

AVP(the average annual precipitation (inch)), SDP(the relative pollution of Sulfur Dioxide) and MPH(the number of members per household in 1960) construct a good mutiple regression model for the Death Rate. The environmental factors and household size affect the Death rate most.

The Death Rate = 457.25325 + 3.12122 * AVP + 0.47141*SDP +104.5393*MPH
457.25325 means when AVP, SDP, MPH equals to 0, the mean value of Death Rate equals to 457.25325.

3.12122 means the change in the mean value of the Death Rate when AVP change by one unit and all other predictor variables are held constant.

0.47141 means the change in the mean value of the Death Rate when SDP change by one unit and all other predictor variables are held constant.

104.539 means the change in the mean value of the Death Rate when MPH change by one unit and all other predictor variables are held constant.

Those numbers illustrate that MPH affects the response variable(the Death Rate) most. Then AVP ,SDP.

Reference

Richard Gunst, Robert Mason, Regression Analysis and Its Applications: a data-oriented approach, Dekker, 1980, pages 370-371. ISBN: 0824769937.

Gary McDonald, Richard Schwing, Instabilities of regression estimates relating air pollution to mortality, Technometrics, Volume 15, Number 3, pages 463-482, 1973.

Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991, ISBN 0-12-656460-4.

a3 = read.csv("/Users/Joy/Desktop/STA302/A3/A3 JOY DATA zhiqian.csv",header=TRUE)

```
#Regression between Death Rate and the average annual precipitation
a3 = read.csv("/Users/Joy/Desktop/STA302/A3/A3 JOY DATA zhiqian.csv",header=TRUE)
lm_fit1 = lm(a3$DEATH.RATE ~ a3$AVP)
summary(lm_fit1)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$AVP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.764  -38.199    3.787   34.342  119.999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 821.7466    27.2108  30.199  < 2e-16 ***
## a3$AVP        3.1743     0.7039   4.509 3.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.99 on 58 degrees of freedom
## Multiple R-squared:  0.2596, Adjusted R-squared:  0.2468
## F-statistic: 20.33 on 1 and 58 DF,  p-value: 3.215e-05
```

```
#Regression between Death Rate and the number of members per household
lm_fit3 = lm(a3$DEATH.RATE ~ a3$MPH)
summary(lm_fit3)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$MPH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -165.539  -35.614   -0.189   36.657  156.884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    404.1      184.2   2.194  0.03228 *
## a3$MPH         164.3       56.4   2.914  0.00507 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.6 on 58 degrees of freedom
## Multiple R-squared:  0.1277, Adjusted R-squared:  0.1126
## F-statistic: 8.489 on 1 and 58 DF,  p-value: 0.005068
```

```
#Regression between Death Rate and the number of years of schooling for persons over 2
lm_fit4 = lm(a3$DEATH.RATE ~ a3$YSP22)
summary(lm_fit4)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$YSP22)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -151.708  -36.691    2.417   43.811  124.916
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1352.996     91.412  14.801  < 2e-16 ***
## a3$YSP22     -37.604      8.306  -4.527 3.02e-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:  0.2484
## F-statistic:  20.5 on 1 and 58 DF,  p-value: 3.022e-05
```

```
#Regression between Death Rate and the number of households with fully equipped kitche
lm_fit5 = lm(a3$DEATH.RATE ~ a3$HFK)
summary(lm_fit5)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$HFK)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -116.147  -37.063   -2.069   26.847  149.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1358.207    116.480  11.660  < 2e-16 ***
## a3$HFK        -5.164      1.437  -3.594 0.000672 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.74 on 58 degrees of freedom
## Multiple R-squared:  0.1822, Adjusted R-squared:  0.1681
## F-statistic: 12.92 on 1 and 58 DF,  p-value: 0.0006718
```

```
#Regression between Death Rate and the number of families with an income less than $30
lm_fit6 = lm(a3$DEATH.RATE ~ a3$NFL3)
summary(lm_fit6)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$NFL3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.188  -31.752    1.504   31.331  131.381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  852.134     26.773  31.828  < 2e-16 ***
## a3$NFL3        6.138      1.790   3.428  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.21 on 58 degrees of freedom
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1542
## F-statistic: 11.75 on 1 and 58 DF,  p-value: 0.001124
```

```
#Regression between Death Rate and the sulfur dioxide pollution index
lm_fit7 = lm(a3$DEATH.RATE ~ a3$SDP)
summary(lm_fit7)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$SDP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -128.408  -35.079   -8.669   34.338  194.851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 917.8874     9.6435  95.182  < 2e-16 ***
## a3$SDP        0.4179     0.1166   3.585 0.000692 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.77 on 58 degrees of freedom
## Multiple R-squared:  0.1814, Adjusted R-squared:  0.1673
## F-statistic: 12.85 on 1 and 58 DF,  p-value: 0.0006922
```

```
#Mutiple Regression
lm_fit8 = lm(a3$DEATH.RATE ~ a3$AVP + a3$MPH + a3$YSP22 + a3$HFK + a3$NFL3 + a3$SDP + a3
summary(lm_fit8)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$AVP + a3$MPH + a3$YSP22 + a3$HFK +
##     a3$NFL3 + a3$SDP + a3$HP + a3$DA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -108.40  -22.84    0.95   23.79  112.18
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 648.79956  267.29044   2.427  0.01878 *
## a3$AVP        2.50851    0.81884   3.063  0.00349 **
## a3$MPH       83.85424   49.00100   1.711  0.09311 .
## a3$YSP22     -1.88113    9.31822  -0.202  0.84082
## a3$HFK       -1.83788    1.77717  -1.034  0.30594
## a3$NFL3       1.59976    2.06984   0.773  0.44316
## a3$SDP        0.47417    0.10173   4.661  2.3e-05 ***
## a3$HP         0.03531    0.08540   0.413  0.68099
## a3$DA         0.75991    1.05024   0.724  0.47264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.1 on 51 degrees of freedom
## Multiple R-squared:  0.585,  Adjusted R-squared:  0.5199
## F-statistic: 8.988 on 8 and 51 DF,  p-value: 1.365e-07
```

```
#Reduced Model
lm_fit9 = lm(a3$DEATH.RATE ~   a3$MPH+ a3$SDP+a3$AVP  )
```

```r
summary(lm_fit9)
```

```
##
## Call:
## lm(formula = a3$DEATH.RATE ~ a3$MPH + a3$SDP + a3$AVP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.841  -23.922   -0.169   21.125  135.633
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 457.25325  136.88389   3.340  0.00149 **
## a3$MPH      104.53931   43.17468   2.421  0.01873 *
## a3$SDP        0.47141    0.08938   5.274 2.22e-06 ***
## a3$AVP        3.12122    0.58821   5.306 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.26 on 56 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5165
## F-statistic: 22.01 on 3 and 56 DF,  p-value: 1.527e-09
```

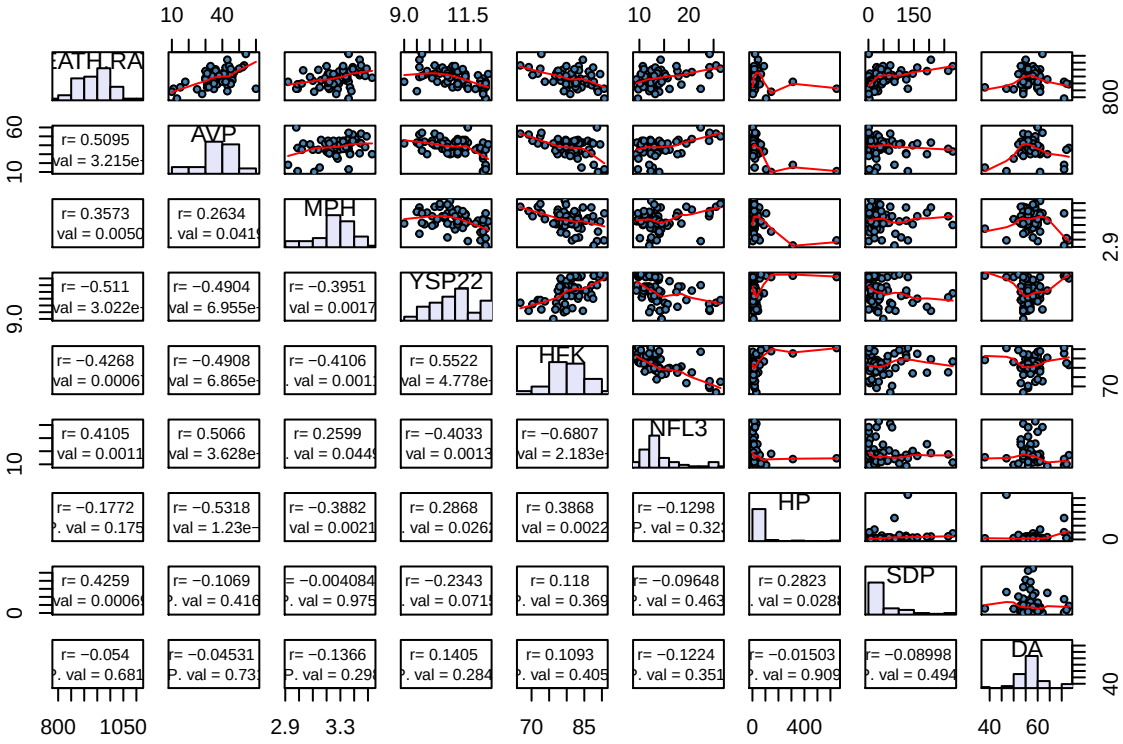
```r
anova(lm_fit9,lm_fit8)
```

```
## Analysis of Variance Table
##
## Model 1: a3$DEATH.RATE ~ a3$MPH + a3$SDP + a3$AVP
## Model 2: a3$DEATH.RATE ~ a3$AVP + a3$MPH + a3$YSP22 + a3$HFK + a3$NFL3 +
##     a3$SDP + a3$HP + a3$DA
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     56 104779
## 2     51  94739  5     10040 1.081  0.382
```

```r
#correlation plots between 9 variables
mycor <- function ( data ){
# ------------ put histograms on the diagonal -----------------
panel.hist <- function (x , ...) {
  usr <- par ("usr") ; on.exit ( par ( usr ))
  par ( usr = c( usr [1:2] , 0, 1.5) )
  h <- hist (x , plot = FALSE )
  breaks <- h$ breaks ; nB <- length ( breaks )
  y <- h$ counts ; y <- y/ max (y)
  rect ( breaks [ - nB ] , 0, breaks [ -1] , y , col ="lavender", ...)}
panel.cor <- function (x , y , digits =4 , prefix ="", cex.cor , ...) {
    usr <- par ("usr") ;
```

```
    on.exit ( par ( usr ))
    par ( usr = c(0 , 1 , 0 , 1) )
    txt1 <- format ( cor (x ,y) , digits = digits )
    txt2 <- format (cor.test (x ,y)$p.value , digits = digits )
    text (0.5 ,0.5 , paste ("r=",txt1 , "\n P. val =", txt2 ) , cex =0.8)}
pairs (data , lower.panel = panel.cor , cex =0.7 , pch = 21 , bg =" steelblue ", diag.pa
   }

mycor(a3[,c(9,1:8)])
```




```
#Find Multiple R-squared of SDP,AVP
summary(lm(a3$SDP ~ a3$AVP))
```

```
##
## Call:
## lm(formula = a3$SDP ~ a3$AVP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.31 -40.10 -18.80  13.81 221.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.1325    32.0404   2.470   0.0165 *
## a3$AVP       -0.6788     0.8289  -0.819   0.4161
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.57 on 58 degrees of freedom
## Multiple R-squared:  0.01143,    Adjusted R-squared:  -0.005612
## F-statistic: 0.6708 on 1 and 58 DF,  p-value: 0.4161
```

```
#Find Multiple R-squared of MPH,AVP
summary(lm(a3$MPH ~ a3$AVP))
```

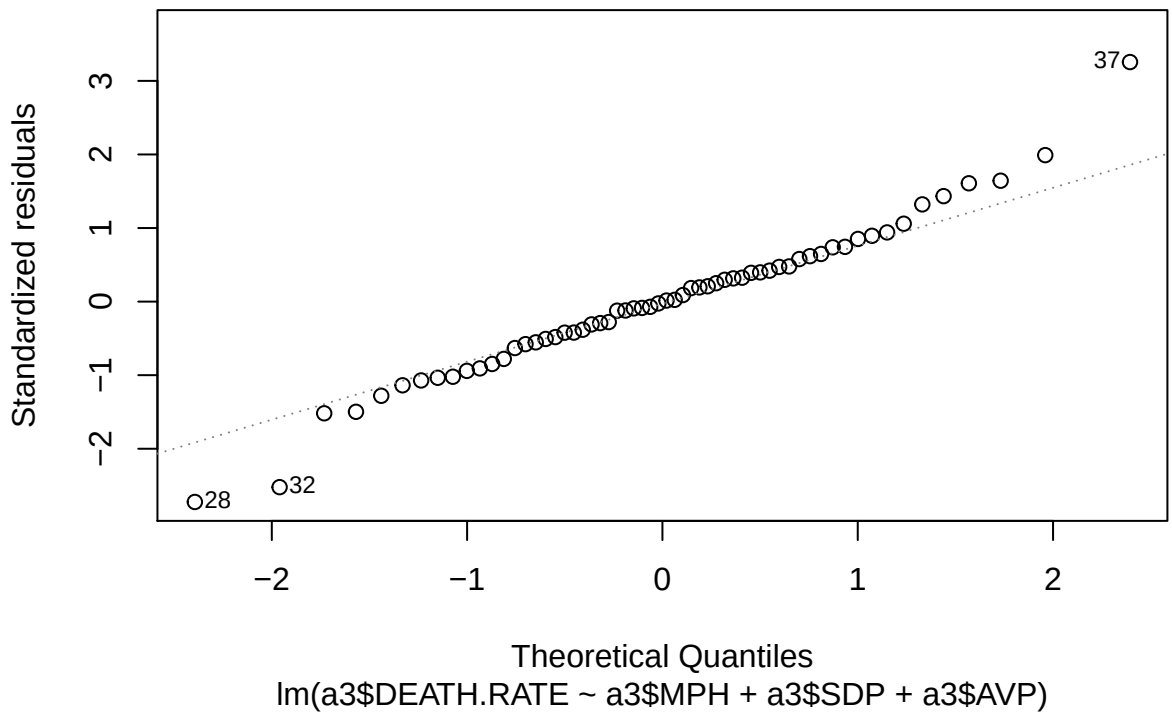
```
##
## Call:
## lm(formula = a3$MPH ~ a3$AVP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36394 -0.05655  0.00351  0.08421  0.29312
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.129820   0.066328   47.19   <2e-16 ***
## a3$AVP      0.003569   0.001716    2.08    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1316 on 58 degrees of freedom
## Multiple R-squared:  0.0694, Adjusted R-squared:  0.05336
## F-statistic: 4.326 on 1 and 58 DF,  p-value: 0.04197
```

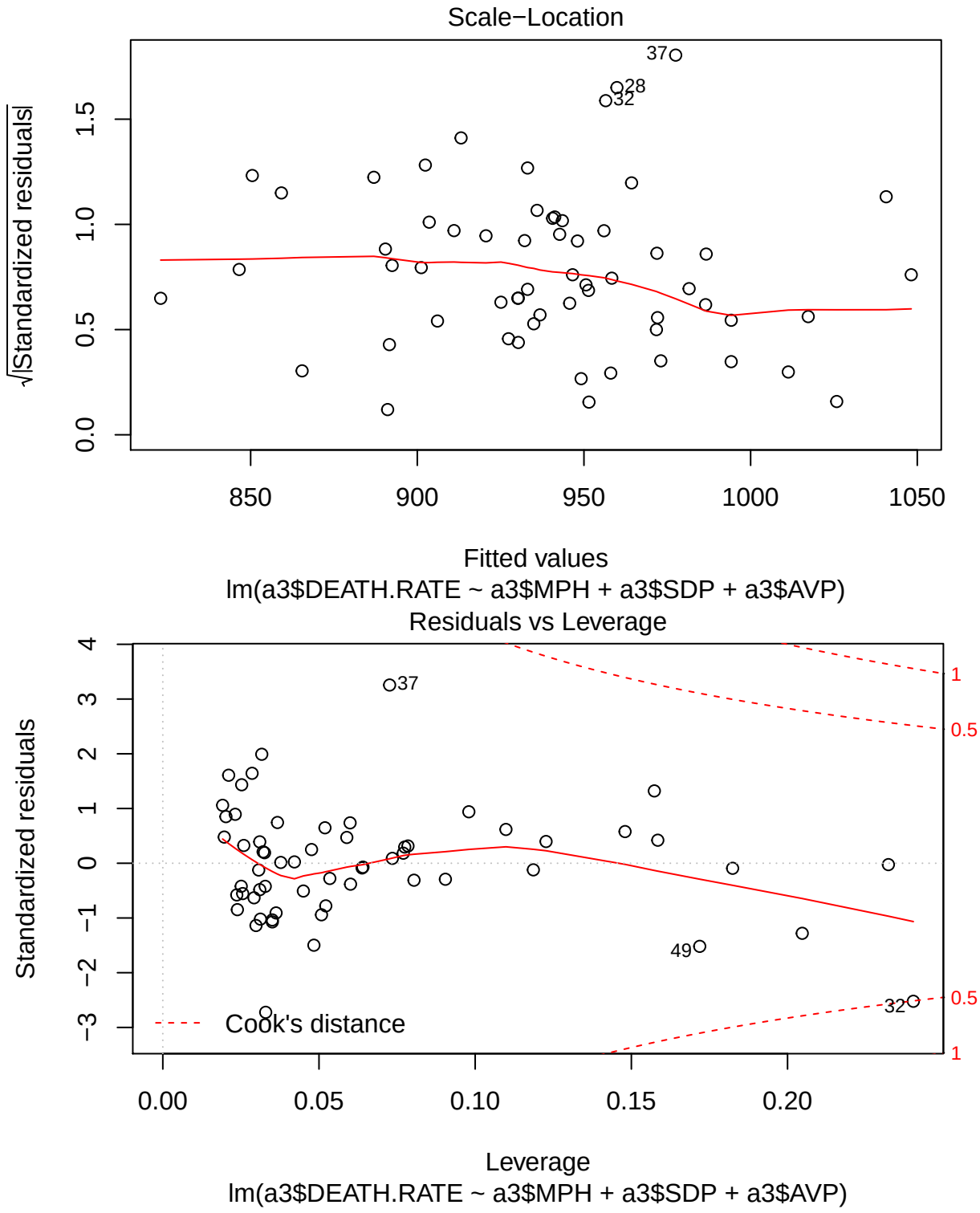
```
#Find Multiple R-squared of SDP,MPH
summary(lm(a3$SDP ~ a3$MPH))
```

```
##
## Call:
## lm(formula = a3$SDP ~ a3$MPH)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.31 -42.38 -23.76  15.28 224.11
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.012    200.987   0.299    0.766
## a3$MPH        -1.914     61.541  -0.031    0.975
##
## Residual standard error: 63.93 on 58 degrees of freedom
```

```
## Multiple R-squared:  1.668e-05,  Adjusted R-squared:  -0.01722
## F-statistic: 0.0009672 on 1 and 58 DF,  p-value: 0.9753
```

```r
#Residual plot
plot(lm_fit9)
```




Scale−Location
√|Standardized residuals|
37
28
32
0.0
0.5
1.0
1.5
850
900
950
1000
1050
Fitted values
lm(a3$DEATH.RATE ~ a3$MPH + a3$SDP + a3$AVP)
Residuals vs Leverage
Standardized residuals
37
49
32
Cook's distance
1
0.5
0.5
1
−3
−2
−1
0
1
2
3
4
0.00
0.05
0.10
0.15
0.20
Leverage
lm(a3$DEATH.RATE ~ a3$MPH + a3$SDP + a3$AVP)