Shuo Han

MA575 Linear Models

Spring 2022

5/4/2022

Transportation

Deaths

Final Project

Part I.

A. Mean Testing

1) Null Hypothesis: I believe that the average Motor vehicle deaths in U.S. is 300 per year from year 2001 to 2019. (H_0: $\mu = \mu0 = 300$, alpha = 0.05)

2) Population: US population from year 2001 to 2019(Year, Deaths, Crashes, Miles traveled (millions), Motor vehicles)

3) Why this claim: Motor vehicles lead to so many injuries but deaths is not a common result.

4) Dataset reference:

https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year

https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot

```
S <- read.csv("accident-1.csv")

alpha <- 0.05
n <- dim(S)[1]

Y <- S[1:n, 2]
X1Old <- S[1:n, 3]
X2Old <- S[1:n, 4]
X3Old <- S[1:n, 5]

sx1Old  <- sd(X1Old)
sx2Old  <- sd(X2Old)
sx3Old  <- sd(X3Old)

x1barOld <- mean(X1Old)
x2barOld <- mean(X2Old)
x3barOld <- mean(X3Old)

SEOld <- sx1Old/sqrt(n)
```

5) Confidence interval：

```
tcrit <- qt(alpha/2, df = n-1, lower.tail=F)

#margin of error
eps <- tcrit * SEOld

#claimed value of the mean
mu0 <- 300

# Confidence Interval
Low <- x1barOld - eps
Upper <- x1barOld + eps
```

```
> Low
[1] 32757
> Upper <- x1barOld + eps
> Upper
[1] 36094
```

test statistic：

```
> tstat <- (x1barOld - mu0)/SEOld
> tstat
[1] 42.98
```

p-value calculated:

```
> #p-value
> pval <- 2*pt(tstat, df=n-1, lower.tail = F)
> pval
[1] 1.353e-19
```

P = 1.353e-19 < 0.05/2, so we can reject null hypothesis, thus we cannot believe that the average Motor vehicle deaths in U.S. is 300 per year.

metric_name	metric_val
CI.lower	3.275712e+04
CI.upper	3.609352e+04
claimed.mean	3.000000e+02
T.stat	4.297726e+01
p-value	1.353378e-19
alpha	5.000000e-02

6) Potential invalidity:

   In the later Normality section, we can see that in the -2 to -1 quantile, and 1 to 2 quantile, so the data are not normally distributed. Thus, we need to think about potential invalidity carefully when using it.

B. Standard Deviation Testing

1) Null Hypothesis: I believe that the standard deviation of Motor vehicle deaths in U.S from year 2001 to 2019 is 30. (H_0: 0: $\sigma = \sigma 0$ = 30, alpha = 0.05)

```
#Null hypothesis: H0: σ = σ0
alpha <- 0.05

#claimed value of the standard deviation
sd0 = 30
```

2) Dataset reference: same

   Population: US population from year 2001 to 2019(Year, Deaths, Crashes, Miles traveled (millions), Motor vehicles)

3) Confidence interval, test statistic, p-value calculated:

```
# Confidence Interval
LowC<-qchisq(alpha/2, df=n-1, lower.tail = T)
UpperC<-qchisq(alpha/2, df=n-1, lower.tail = F)
LowC
UpperC

# Test Statistics
tstatC<- (n-1)*(sx1Old/sd0)^2
tstatC

#p-value
pvalC <- 2*pt(abs(tstatC), df=n-1, lower.tail = F)
pvalC
```

```
> #Summary
> metric_nameC <-c("CI.lower","CI.upper", "claimed.sd","T.stat", "p-value","alpha")
> metric_valC <- c(LowC, UpperC, sd0, tstatC, pvalC, alpha)
>
> options(digits =7)
> SummaryC <- data.frame(metric_nameC, metric_valC)
> SummaryC
  metric_nameC  metric_valC
1     CI.lower 8.230746e+00
2     CI.upper 3.152638e+01
3   claimed.sd 3.000000e+01
4       T.stat 2.395850e+05
5      p-value 5.438198e-87
6        alpha 5.000000e-02
```

4) P = 5.438198e-87 < 0.05/2, so we can reject null hypothesis, thus we cannot believe that the standard deviation of Motor vehicle deaths in U.S. is 30.

5) Potential invalidity:

   In the later Normality section, we can see that in the -2 to -1 quantile, and 1 to 2 quantile, so the data are not normally distributed. Thus, we need to think about potential invalidity carefully when using it.

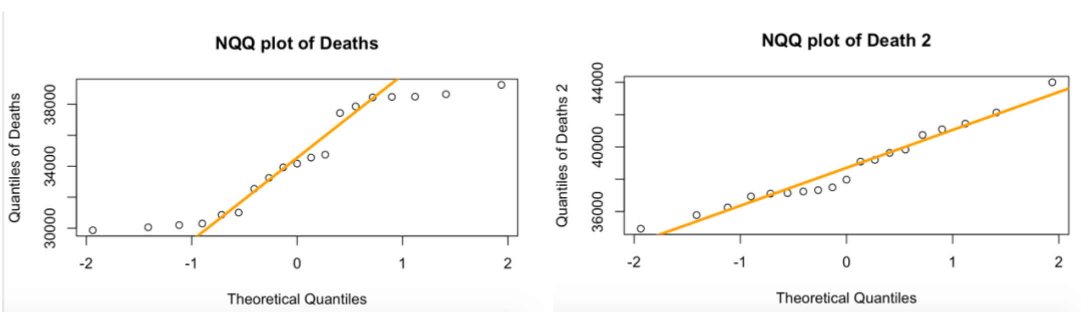
C. Normality Testing

1) Dataset 1: Same as 1.A

Dataset 2: US population from year 1981 to 1999(Year, Deaths, Crashes, Miles traveled (millions), Motor vehicles) from the same source as A.

2) Normal-Quantile Quantile plots:

```
#Part I.C - Normality Testing
Q1 <- qqnorm(X1Old, ylab = "Quantiles of Deaths", main = "NQQ plot of Deaths")
qqline(X1Old, col="orange", lwd=3)

cor(Q1$x,Q1$y)

S2 <- read.csv("accident-2.csv")

n2 <- dim(S2)[1]
X21Old <- S2[1:n2, 3]
sx21Old  <- sd(X21Old)
x21barOld <- mean(X21Old)

Q2 <- qqnorm(X21Old, ylab = "Quantiles of Deaths 2", main = "NQQ plot of Death 2")
qqline(X21Old, col="orange", lwd=3)

cor(Q2$x,Q2$y)
```



3) Correlation coefficient of NQQ Plot of Deaths:

```
> cor(Q1$x,Q1$y)
[1] 0.9477006
```

Correlation coefficient of NQQ Plot of Deaths 2:

```
> cor(Q2$x,Q2$y)
[1] 0.9802995
```

4) Interpretation of the plots and calculations:

In the NQQ Plot of Deaths, from -1 to 1 quantile, the data almost fit the straight line, which means they are normally distributed. While for the -2 to -1 quantile, and 1 to 2 quantile, the data are not normally distributed.

In the NQQ Plot of Deaths 2, the data almost fit the straight line, but there are still some skews in the plot from -1 to 0 quantile and so on, which means the data are not perfectly normally distributed.

D. Parameter Comparisons for Means

1) Null hypothesis: I believe that the mean of Motor vehicle deaths in U.S. from year 2001 to 2019 is the same as that from year 1981 to 1999. ($H0$: $\mu 1 = \mu 2$, alpha = 0.05)

2) test used to test this claim: t-test

   assumptions: t-test assumes random sampling, normality of these dataset distribution, adequacy of sample size(<30), and equality of variance in standard deviation.

3) Confidence interval, test statistic, p-value calculated:

```
#test statistic
xbarD<-x1barOld-x21barOld
SED<-sqrt((sx1Old^2/n)+(sx21Old/n2))

tcritD<-qt(alpha/2, n+n2-2,lower.tail=F)
epsD<-tcritD*SED
tstatD<-(xbarD-0)/SED

#p-value calculated
pvalD<-2*pt(-abs(tstatD), n+n2-2, lower.tail = T)

#Confidence interval
LowD<-xbarD-epsD
UpperD<-xbarD+epsD
```

```
 metric_nameD metric_valD
     CI.lower  -5.888e+03
     CI.upper  -2.667e+03
       T.stat  -5.386e+00
      p-value   4.596e-06
        alpha   5.000e-02
```

5) Interpretation of the calculations:

P = 4.596e-06< 0.05/2, so we can reject null hypothesis, thus we cannot believe that the mean of Motor vehicle deaths in U.S. from year 2001 to 2019 is the same as that from year 1981 to 1999, which means there is a difference between the mean of Motor vehicle deaths in U.S. from year 2001 to 2019 and that from year 1981 to 1999.

6) Potential invalidity:

   In the NQQ Plot of Deaths, from -1 to 1 quantile, the data almost fit the straight line, which means they are normally distributed. While for the -2 to -1 quantile, and 1 to 2 quantile, the data are not normally distributed.

   In the NQQ Plot of Deaths 2, the data almost fit the straight line, but there are still some skews in the plot from -1 to 0 quantile and so on, which means the data are not perfectly normally distributed.

   Thus, we need to think about potential invalidity carefully when we are using the analysis.

E. Parameter Comparisons for Variances

1) Null hypothesis: I believe that the variance of Motor vehicle deaths in U.S. from year 2001 to 2019 is the same as that from year 1981 to 1999. ($H0$: $\sigma 2 = \sigma 2$, alpha = 0.05)
2) test used to test this claim: f-test

   assumptions: An F-test assumes that dataset “accident-1” and “accident-2” are both normally distributed and that they are independent from one another.

3) Confidence interval, test statistic, p-value calculated:

```
#Part I.E – Parameter Comparisons for Variances
source("nemolm2.r")
#Null hypothesis: H : σ2 = σ2
#test statistic
sx21Old  <- sd(X21Old)
fstatV<-sx1Old^2/sx21Old^2

#Confidence interval
fcritLV<-qf(alpha/2, n-1, n2-1,lower.tail = T )
fcritUV<-qf(alpha/2, n-1, n2-1,lower.tail = F )

#p-value calculated
fstatLV<-min(fstatV, 1/fstatV)
fstatUV<-max(fstatV, 1/fstatV)
pvalFV<-pf(fstatLV, n-1, n2-1, lower.tail = T) + pf(fstatUV, n-1, n2-1, lower.tail = F)
```

```
metric_nameFV metric_valFV
     CI.lower    0.3852685
     CI.upper    2.5955922
       T.stat    2.1037119
      p-value    0.1238677
        alpha    0.0500000
```

4) Interpretation of the calculations:

P = 0.1238677< 0.05/2, so we can reject null hypothesis, thus we cannot believe that the variance of Motor vehicle deaths in U.S. from year 2001 to 2019 is the same as that from year 1981 to 1999.

5) Potential invalidity:

In the NQQ Plot of Deaths, from -1 to 1 quantile, the data almost fit the straight line, which means they are normally distributed. While for the -2 to -1 quantile, and 1 to 2 quantile, the data are not normally distributed.

In the NQQ Plot of Deaths 2, the data almost fit the straight line, but there are still some skews in the plot from -1 to 0 quantile and so on, which means the data are not perfectly normally distributed.

Thus, we need to think about potential invalidity carefully when we are using the analysis.

Part II

A. Simple Linear Regression

1) numerical explanatory variable(X): Standardized Crushes

2) response variable(Y): Standardized Deaths

3) why linearly related: The r^2 of the model of 0.9962, which means the model explains 99.62% of the data. Also, the p-value of this model is 0.001063 < 0.05, so we can accept this model. It is a good fit.

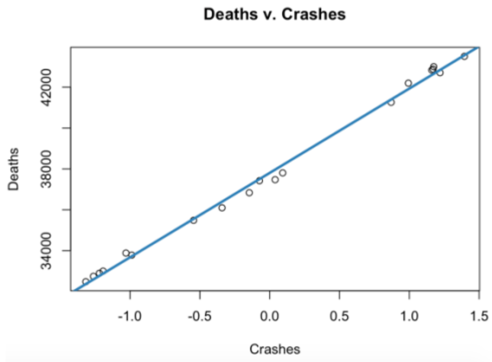
```
 metric.name metric.val
  covariance  4.130e+03
     r value  9.981e-01
   r^2 value  9.962e-01
    beta1hat  4.130e+03
 SE.beta1hat  1.049e+03
    beta0hat  3.781e+04
 SE.beta0hat  6.006e+01
         SSE  1.165e+06
```

```
alpha<-0.05
tcrits<-qt(alpha/2, df=n-2, lower.tail = F)
beta1<-0
epsS <- tcrits*SE.beta1hat
tstats <- (beta1hat - beta1)/SE.beta1hat
CIL <- beta1hat - epsS
CIU <- beta1hat + epsS
pvalS <- 2*pt(abs(tstats), df=n-2, lower.tail = F)
pvalS
```

```
> pvalS
[1] 0.001063
```

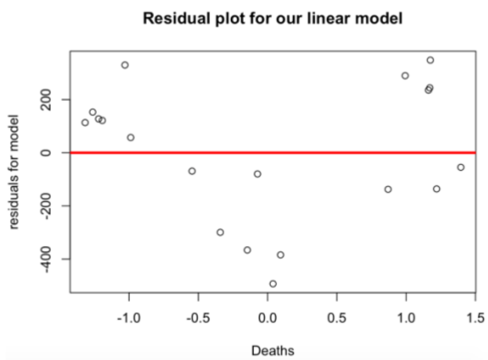
4) Scatter plot of the data

```
# Linear Regression Line
yhat <- lm(Y ~ X1)
abline(yhat, col="steelblue", lwd = 3)
```



5) Standardized residual plot

```
residual <- resid(yhat)
plot(X1, residual, xlab = "Deaths", ylab = "residuals for model",
      main="Residual plot for our linear model")
abline(0, 0, col ="red", lwd=3)
```



6) Y = 4130*X1+37810

metric.name	metric.val
covariance	4.130e+03
r value	9.981e-01
r^2 value	9.962e-01
beta1hat	4.130e+03
SE.beta1hat	1.049e+03
beta0hat	3.781e+04
SE.beta0hat	6.006e+01
SSE	1.165e+06

```
#Standardize
X1 <- (X1Old-x1barOld)/sx1Old
X2 <- (X2Old-x2barOld)/sx2Old
X3 <- (X3Old-x3barOld)/sx3Old

sx1  <- 1
sx2  <- 1
sx3  <- 1

x1bar <- 0
x2bar <- 0
x3bar <- 0

ybar <- mean(Y)
sy <- sd(Y)
covs <- cov(X1, Y)
rs <- cor(X1, Y)

SE <- sx1/sqrt(n)

plot(X1, Y, xlab = "Crashes", ylab = "Deaths", main = "Deaths v. Crashes")

beta1hat <- rs*sy/sx1
beta0hat <- ybar - beta1hat*x1bar

SSE <-  sy^2*(n-1)*(1-rs^2)
SE.beta1hat <- (1/sx1)*sqrt(SSE/(n-1)*(n-2))
SE.beta0hat <- sqrt(SSE/(n-2))*sqrt(1/n + (x1bar)^2/(sx1^2*(n-1)))
```

B. Simple Quadratic Regression

1) numerical explanatory variable(X1): <u>Standardized</u> Crushes

response variable(Y): <u>Standardized</u> Deaths

```
$predicted
        [,1]
 [1,] 41903
 [2,] 42757
 [3,] 42738
 [4,] 42693
 [5,] 43813
 [6,] 42973
 [7,] 41334
 [8,] 37252
 [9,] 33597
[10,] 33021
[11,] 32594
[12,] 33746
[13,] 32928
[14,] 32781
[15,] 35387
[16,] 37938
[17,] 37712
[18,] 36955
[19,] 36177

$residual
        [,1]
 [1,]  292.67
 [2,]  248.00
 [3,]  146.19
 [4,]  143.40
 [5,] -303.46
 [6,] -264.84
 [7,]  -74.89
 [8,]  171.44
 [9,]  286.28
[10,]  -21.79
[11,] -114.81
[12,]   36.47
[13,]  -33.53
[14,]  -36.90
[15,]   97.78
[16,] -131.59
[17,] -239.05
[18,] -119.91
[19,]  -81.47

$sres
         [,1]
 [1,]  1.6121
 [2,]  1.3973
 [3,]  0.8230
 [4,]  0.8058
 [5,] -1.8463
 [6,] -1.5076
 [7,] -0.4112
 [8,]  0.9751
 [9,]  1.5835
[10,] -0.1243
[11,] -0.6851
[12,]  0.2009
[13,] -0.1927
[14,] -0.2152
[15,]  0.5417
[16,] -0.7491
[17,] -1.3613
[18,] -0.6805
[19,] -0.4577

$condition
[1] 10.69

$leverage
 [1] 0.1074 0.1468 0.1454 0.1423 0.2683 0.1642 0.1015 0.1628 0.1148 0.1674 0.2396 0.1076 0.1804
[14] 0.2038 0.1177 0.1642 0.1648 0.1592 0.1418

$sse
[1] 590788

$mse
[1] 36924

$ssm
[1] 307616142

$msm
[1] 153808071

$pval
[1] 1.823e-22

$betahat
        [,1]
[1,] 37551.4
[2,]  4116.9
[3,]   267.8

$SEbetahat
[1] 77.97 45.41 67.88

$r2
[1] 0.9981

$r2adj
[1] 0.9978
```

```
M2 <- nemolm2(Y, cbind(X1, X1^2))

#Standardized residual plot
plot(X1, M2$residual, xlab = "Crashes", ylab = "residuals for Simple Quadratic model",
     main="Residual plot for Simple Quadratic model")
abline(0, 0, col ="red", lwd=3)
M2
```

2) Y = 4116.9*X1+267.8*X1^2+37551.4

   Why good predictors of Y:

   The r^2 of the model of 0.9981, which means the model explains 99.81% of the data.

   Also, the p-value of this model is 0.001063, so we can accept this model. It is a good fit.

Part III Multiple Linear Regression

1) numerical explanatory variable(X): Standardized Crushes(X1), Miles traveled (millions)(X2), Motor vehicles(X3)

   response variable(Y): Standardized Deaths

```
#Part III - Multiple Linear Regression
M3 <- nemolm2(Y, cbind(X1, X2, X3))
M3
```

```
$predicted
       [,1]
 [1,] 42359
 [2,] 42908
 [3,] 42853
 [4,] 42660
 [5,] 43514
 [6,] 42724
 [7,] 41270
 [8,] 37535
 [9,] 33654
[10,] 33029
[11,] 32564
[12,] 33854
[13,] 32903
[14,] 32660
[15,] 35470
[16,] 37878
[17,] 37630
[18,] 36834
[19,] 36000

$residual
         [,1]
 [1,] -163.015
 [2,]   96.929
 [3,]   30.746
 [4,]  176.074
 [5,]   -3.970
 [6,]  -15.652
 [7,]  -10.517
 [8,] -111.651
 [9,]  228.794
[10,]  -30.193
[11,]  -84.662
[12,]  -71.692
[13,]   -9.132
[14,]   84.500
[15,]   15.288
[16,]  -71.793
[17,] -157.372
[18,]    1.417
[19,]   95.900

$sres
          [,1]
 [1,] -1.96435
 [2,]  0.95400
 [3,]  0.30917
 [4,]  1.67805
 [5,] -0.03903
 [6,] -0.16570
 [7,] -0.10552
 [8,] -1.09629
 [9,]  2.21534
[10,] -0.29108
[11,] -0.83274
[12,] -0.68171
[13,] -0.08782
[14,]  0.80859
[15,]  0.14318
[16,] -0.69097
[17,] -1.56380
[18,]  0.01463
[19,]  1.01344

$betahat
        [,1]
[1,] 37805.1
[2,]  3663.2
[3,]  -263.8
[4,]   408.6

$SEbetahat
[1]  26.02 306.16  34.32 302.03

$r2
[1] 0.9994

$r2adj
[1] 0.9992

$condition
[1] 535.3

$leverage
 [1] 0.4645 0.1973 0.2309 0.1439 0.1959 0.3062 0.2275 0.1934 0.1706 0.1633 0.1962 0.1400 0.1592
[14] 0.1508 0.1134 0.1605 0.2125 0.2703 0.3037

$sse
[1] 192896

$mse
[1] 12860

$ssm
[1] 3.08e+08

$msm
[1] 102671345

$pval
[1] 3.05e-24
```

1) Y = 3663.2*X1-263.8*X2+408.6*X3+37805.1

   Why good predictors of Y:

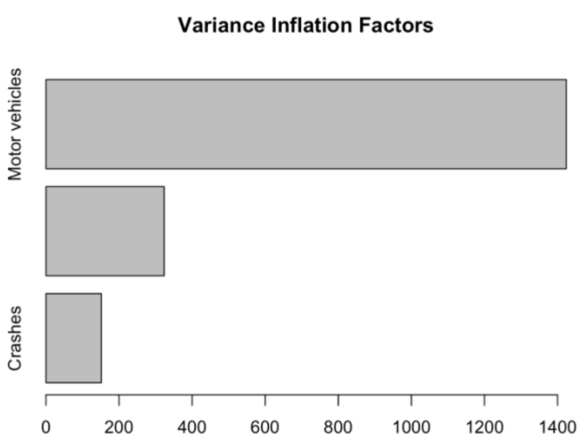
   The r^2 of the model of 0.9994, which means the model explains 99.94% of the data. Also, the p-value of this model is 0.001063, so we can accept this model. It is a good fit. But we still need to figure out collinearity later.

2) ANOVA:

```
> #ANOVA table
> metric_name_A <-c("SST", "MST", "SSM", "MSM", "SSE", "MSE", "Fstat", "p-value")
> metric_val_A <-c(M3$sst, M3$mst, M3$ssm, M3$msm, M3$sse, M3$mse, M3$fstat, M3$pval)
> Summary_A <- data.frame(metric_name_A, metric_val_A)
> Summary_A
  metric_name_A metric_val_A
1           SST    3.082e+08
2           MST    1.712e+07
3           SSM    3.080e+08
4           MSM    1.027e+08
5           SSE    1.929e+05
6           MSE    1.286e+04
7         Fstat    7.984e+03
8       p-value    3.050e-24
```

Variance inflation factors:

```
#Variance inflation factors calculated for each variable with barplot
# Y regressed on X1, X2, and X3
MYvX1c <- nemolm2(Y, cbind(X2, X3))
MX1vX1c <- nemolm2(X1, cbind(X2, X3))

MYvX2c <- nemolm2(Y, cbind(X1, X3))
MX2vX2c <- nemolm2(X2, cbind(X1, X3))

MYvX3c <- nemolm2(Y, cbind(X1, X2))
MX3vX3c <- nemolm2(X3, cbind(X1, X2))

vif1 <- 1/(1-MYvX1c$r2)
vif2 <- 1/(1-MYvX2c$r2)
vif3 <- 1/(1-MYvX3c$r2)

vif <- c(vif1, vif2, vif3)
barplot(vif, horiz=T, main="Variance Inflation Factors",
        names.arg = c('Crashes', 'Miles traveled (millions)', 'Motor vehicles'),
        xlim=c(0,1425))
```




```
> vif
[1]  151.5  323.5 1424.0
```

Interpretation:

The VIF shows there is strong multicollinearity since the vifs are all larger than 100. Thus, I choose to add an interaction term.

New model: Y = 3946.36*X1-225*X2+98.12*X3-72.64*X2*X3+37793.78

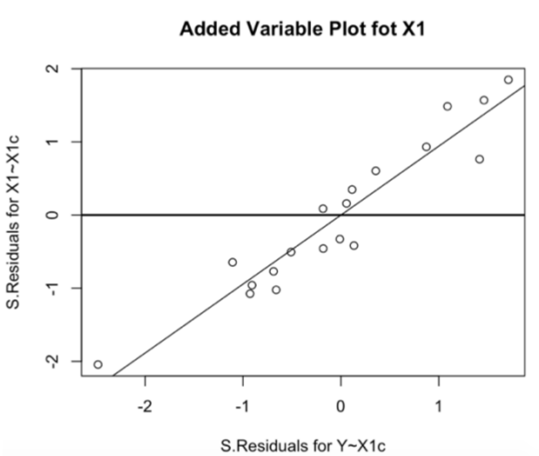
The r^2 of this model is 0.9994, which means the model explains 99.94% of the data. Also, the p-value of this model is 2.398e-22<0.05, so we can accept this model. It is a good fit.

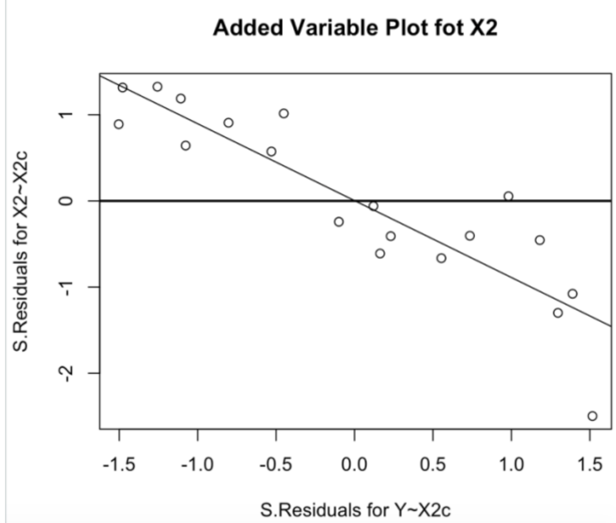
```
> #new fits
> M4 <- nemolm2(Y, cbind(X1, X2, X3, X2*X3))
> M4
```

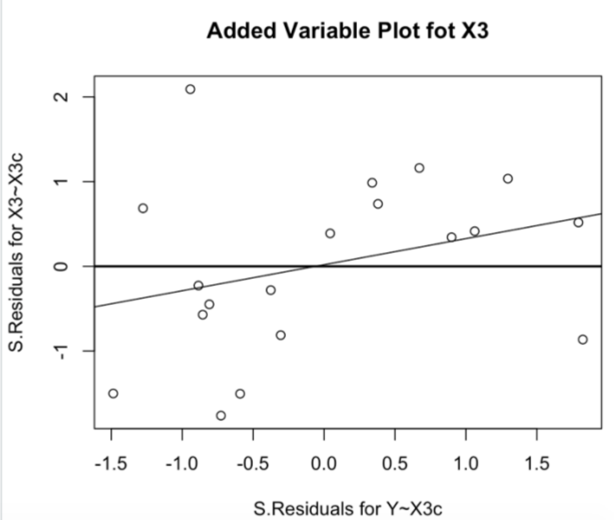
```
$predicted
       [,1]
 [1,] 42373
 [2,] 42933
 [3,] 42846
 [4,] 42630
 [5,] 43500
 [6,] 42725
 [7,] 41270
 [8,] 37556
 [9,] 33665
[10,] 33017
[11,] 32531
[12,] 33848
[13,] 32893
[14,] 32683
[15,] 35484
[16,] 37889
[17,] 37605
[18,] 36825
[19,] 36025

$residual
           [,1]
 [1,] -177.3524
 [2,]   72.1888
 [3,]   38.0361
 [4,]  206.4495
 [5,]   10.1296
 [6,]  -17.2614
 [7,]  -10.5167
 [8,] -132.6749
 [9,]  217.7220
[10,]  -17.9910
[11,]  -51.8939
[12,]  -65.8935
[13,]    0.9335
[14,]   61.0533
[15,]    0.8515
[16,]  -83.2478
[17,] -131.5287
[18,]    9.6476
[19,]   71.3483

$sres
           [,1]
 [1,] -2.167125
 [2,]  0.745389
 [3,]  0.377665
 [4,]  2.127385
 [5,]  0.099833
 [6,] -0.179480
 [7,] -0.103601
 [8,] -1.340241
 [9,]  2.095235
[10,] -0.172820
[11,] -0.566522
[12,] -0.617167
[13,]  0.008903
[14,]  0.606466
[15,]  0.007984
[16,] -0.796876
[17,] -1.382206
[18,]  0.098533
[19,]  0.798701

$sst
[1] 308206930

$mst
[1] 17122607

$fstat
[1] 5773

$condition
[1] 1875

$leverage
 [1] 0.4979 0.2969 0.2396 0.2940 0.2282 0.3066 0.2275 0.2654 0.1905 0.1876 0.3710 0.1454 0.1757 0.2403
[15] 0.1473 0.1819 0.3212 0.2813 0.4018

$sse
[1] 186751

$mse
[1] 13339

$ssm
[1] 3.08e+08

$msm
[1] 77005045

$pval
[1] 2.398e-22

$betahat
         [,1]
[1,] 37793.78
[2,]  3946.36
[3,]  -225.00
[4,]    98.12
[5,]   -72.64

$SEbetahat
[1]  31.31 520.86  66.98 551.32 107.02

$r2
[1] 0.9994

$r2adj
[1] 0.9992
```

Added variable plots:

```r
#Added variable plots for each variable
plot(MYvX1c$sres, MX1vX1c$sres,
     main="Added Variable Plot fot X1",
     xlab = "S.Residuals for Y~X1c",
     ylab = "S.Residuals for X1~X1c")
abline(0,0, lwd=2)
abline(mean(MX1vX1c$sres)-cor(MYvX1c$sres,
                              MX1vX1c$sres)*sd(MX1vX1c$sres)/sd(MYvX1c$sres)*mean(MYvX1c$sres),
       cor(MYvX1c$sres, MX1vX1c$sres)*sd(MX1vX1c$sres)/sd(MYvX1c$sres))

plot(MYvX2c$sres, MX2vX2c$sres,
     main="Added Variable Plot fot X2",
     xlab = "S.Residuals for Y~X2c",
     ylab = "S.Residuals for X2~X2c")
abline(0,0, lwd=2)
abline(mean(MX2vX2c$sres)-cor(MYvX2c$sres,
                              MX2vX2c$sres)*sd(MX2vX2c$sres)/sd(MYvX2c$sres)*mean(MYvX2c$sres),
       cor(MYvX2c$sres, MX2vX2c$sres)*sd(MX2vX2c$sres)/sd(MYvX2c$sres))

plot(MYvX3c$sres, MX3vX3c$sres,
     main="Added Variable Plot fot X3",
     xlab = "S.Residuals for Y~X3c",
     ylab = "S.Residuals for X3~X3c")
abline(0,0, lwd=2)
abline(mean(MX3vX3c$sres)-cor(MYvX3c$sres,
                              MX3vX3c$sres)*sd(MX3vX3c$sres)/sd(MYvX3c$sres)*mean(MYvX3c$sres),
       cor(MYvX3c$sres, MX3vX3c$sres)*sd(MX3vX3c$sres)/sd(MYvX3c$sres))
```







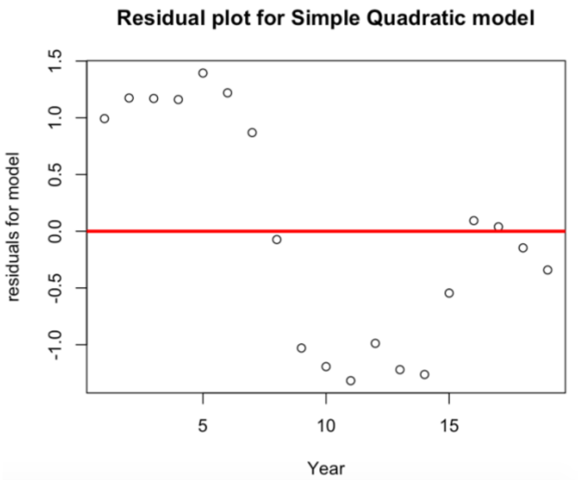

```r
#Standardized residual plot with title and axis labels
plot(X1, M2$std.residual, xlab = "Year", ylab = "residuals for model",
     main="Residual plot for Simple Quadratic model")
abline(0, 0, col ="red", lwd=3)
```

Interpretation:

In the added variable plot, these slopes are all not equal to 0, thus their coefficients are of significance and influential in model after adjusting for the other variables.

Standardized residual plot:

```
#Standardized residual plot with title and axis labels
plot(X1, M2$std.residual, xlab = "Year", ylab = "residuals for model",
     main="Residual plot for Simple Quadratic model")
abline(0, 0, col ="red", lwd=3)
```



Interpretation:

It shows the variance is not constant, thus the constant variance assumption does not hold.

correlation matrix:

```
> cor(S)
                             Year  Deaths Crashes Miles.traveled..millions. Motor.vehicles
Year                       1.0000 -0.6825 -0.6409                    0.8742        -0.5784
Deaths                    -0.6825  1.0000  0.9981                   -0.2840         0.9891
Crashes                   -0.6409  0.9981  1.0000                   -0.2305         0.9939
Miles.traveled..millions.  0.8742 -0.2840 -0.2305                    1.0000        -0.1645
Motor.vehicles            -0.5784  0.9891  0.9939                   -0.1645         1.0000
```

3) An example of prediction:

The the Motor vehicle Deaths in U.S. in a year from 2001 to 2019 with 37860 Crushes(X1), 2781460 Miles traveled (millions)(X2), and 57920 Motor vehicles(X3).

Y = 3663.2*X1-263.8*X2+408.6*X3+37805.1 = 3663.2*(37860-34425)/3461 - 263.8*(2781460-3018857)/127871 + 408.6*(57920-51991)/ 5674 + 37805.1 = 42357.49821

Thus, the Motor vehicle Deaths in U.S. in a year from 2001 to 2019 with 37860 Crushes, 2781460 Miles traveled (millions), and 57920 Motor vehicles is expected to be 42357.49821.

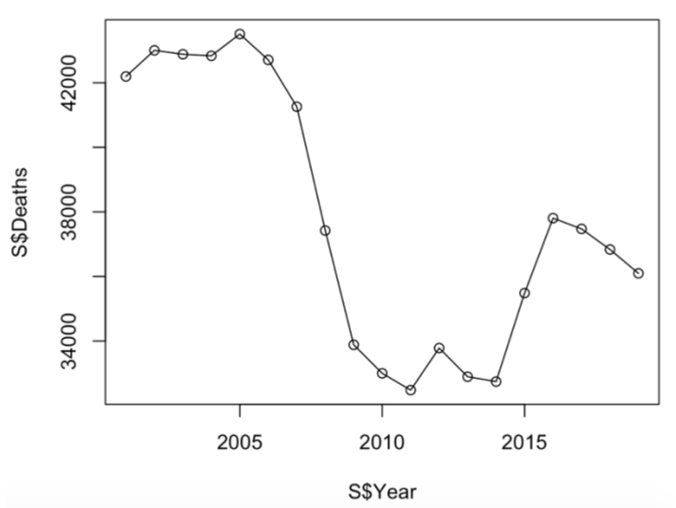
Part IV Time Series Fundamentals

1) numerical explanatory variable(X): Year

   response variable(Y): Deaths

   Scatter plot:

```
#Part IV – Time Series Fundamentals
plot(S$Year, S$Deaths, type='o')
```



2) Quantitative observations on trends and seasonality (or lack thereof) and if it is stationary (or not):

There is no obvious seasonality in this plot. It is not stationary since there are Trends and changing levels in this plot since it is a quartic pattern.

3) Polynomial regression model on maximum interval with min. degree for peak/trough matching(whole model):

```r
#Standardize
XT1 <- S$Year
YT1 <- S$Deaths

sxT1  <- sd(XT1)
xT1 <- mean(XT1)
XT <- (XT1-xT1)/sxT1

syT1  <- sd(YT1)
yT1 <- mean(YT1)
YT <- (YT1-yT1)/syT1

MT<-nemolm2(YT, cbind(XT, XT^2, XT^3, XT^4))
plot(XT, YT, type='o')
lines(XT, MT$predicted, lwd=3, col='green')
```

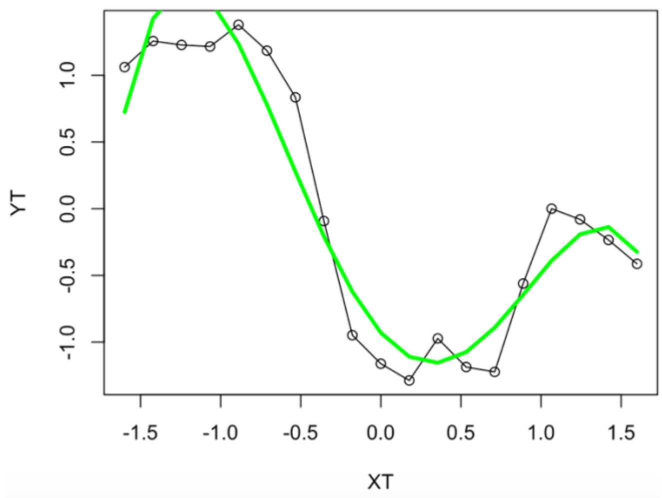
```
$predicted
        [,1]
 [1,]  0.7248
 [2,]  1.4257
 [3,]  1.6627
 [4,]  1.5623
 [5,]  1.2361
 [6,]  0.7806
 [7,]  0.2774
 [8,] -0.2072
 [9,] -0.6215
[10,] -0.9293
[11,] -1.1090
[12,] -1.1542
[13,] -1.0736
[14,] -0.8909
[15,] -0.6446
[16,] -0.3885
[17,] -0.1912
[18,] -0.1365
[19,] -0.3232
```

```
$residual
          [,1]
 [1,]  0.33637
 [2,] -0.16906
 [3,] -0.43529
 [4,] -0.34652
 [5,]  0.14256
 [6,]  0.40422
 [7,]  0.55730
 [8,]  0.11481
 [9,] -0.32632
[10,] -0.23221
[11,] -0.17817
[12,]  0.18197
[13,] -0.11321
[14,] -0.33222
[15,]  0.08390
[16,]  0.38868
[17,]  0.11094
[18,] -0.09791
[19,] -0.08984
```

```
$sres
        [,1]
 [1,]  2.0397
 [2,] -0.6070
 [3,] -1.5225
 [4,] -1.2205
 [5,]  0.4941
 [6,]  1.3675
 [7,]  1.8612
 [8,]  0.3843
 [9,] -1.1024
[10,] -0.7883
[11,] -0.6019
[12,]  0.6091
[13,] -0.3781
[14,] -1.1239
[15,]  0.2908
[16,]  1.3690
[17,]  0.3880
[18,] -0.3516
[19,] -0.5448
```

```
$sst
[1] 18

$mst
[1] 1

$fstat
[1] 38.63
```

```
$condition
[1] 204.5

$leverage
 [1] 0.7454 0.2738 0.2347 0.2453 0.2207 0.1820 0.1605 0.1643 0.1796 0.1875 0.1796 0.1643 0.1605 0.1820
[15] 0.2207 0.2453 0.2347 0.2738 0.7454

$sse
[1] 1.495

$mse
[1] 0.1068

$ssm
[1] 16.5

$msm
[1] 4.126

$pval
[1] 2.025e-07

$betahat
        [,1]
[1,] -0.9293
[2,] -1.3845
[3,]  2.0472
[4,]  0.4132
[5,] -0.6276

$SEbetahat
[1] 0.1415 0.1945 0.3177 0.1051 0.1264

$r2
[1] 0.9169

$r2adj
[1] 0.8932
```



Interpretations:

Since there is an obvious quartic pattern in this plot, I chose quartic regression to fit this model. And, the model is quartic fitted well for the peak and trough, so we can accept this model to fit the data.

Appendix I. – R code

```
#Shuo Han U09953590
#MA575 Linear Models Spring2022
#05/03/2022

#Part I.A - Univariate Data Analysis - Mean Testing
#Null Hypothesis: H0:  μ  =  μ0 = 30
#The averCrashes fatalities per 100,000 population is 30
S <- read.csv("accident-1.csv")

alpha <- 0.05
n <- dim(S)[1]

Y <- S[1:n, 2]
X1Old <- S[1:n, 3]
X2Old <- S[1:n, 4]
X3Old <- S[1:n, 5]

sx1Old  <- sd(X1Old)
sx2Old  <- sd(X2Old)
sx3Old  <- sd(X3Old)

x1barOld <- mean(X1Old)
x2barOld <- mean(X2Old)
x3barOld <- mean(X3Old)

SEOld <- sx1Old/sqrt(n)

tcrit <- qt(alpha/2, df = n-1, lower.tail=F)

#margin of error
eps <- tcrit * SEOld

#claimed value of the mean
mu0 <- 300

# Confidence Interval
Low <- x1barOld - eps
```

```
Low
Upper <- x1barOld + eps
Upper

# Test Statistics
tstat <- (x1barOld - mu0)/SEOld
tstat

#p-value
pval <- 2*pt(tstat, df=n-1, lower.tail = F)
pval

#Summary
metric_name <-c("CI.lower", "CI.upper", "claimed.mean","T.stat", "p-value","alpha")
metric_val <- c(Low, Upper, mu0, tstat, pval, alpha)
options(digits =7)
Summary <- data.frame(metric_name, metric_val)
Summary


#Part I.B – Univariate Data Analysis – Standard Deviation Testing
#Null hypothesis: H0: σ = σ0
alpha <- 0.05

#claimed value of the standard deviation
sd0 = 30

# Confidence Interval
LowC<-qchisq(alpha/2, df=n-1, lower.tail = T)
UpperC<-qchisq(alpha/2, df=n-1, lower.tail = F)
LowC
UpperC

# Test Statistics
tstatC<- (n-1)*(sx1Old/sd0)^2
tstatC


#p-value
pvalC <- 2*pt(abs(tstatC), df=n-1, lower.tail = F)
pvalC

#Summary
metric_nameC <-c("CI.lower","CI.upper", "claimed.sd","T.stat", "p-value","alpha")
metric_valC <- c(LowC, UpperC, sd0, tstatC, pvalC, alpha)

options(digits =7)
SummaryC <- data.frame(metric_nameC, metric_valC)
SummaryC



#Part I.C – Normality Testing
Q1 <- qqnorm(X1Old, ylab = "Quantiles of Deaths", main = "NQQ plot of Deaths")
qqline(X1Old, col="orange", lwd=3)

cor(Q1$x,Q1$y)

S2 <- read.csv("accident-2.csv")

n2 <- dim(S2)[1]
X21Old <- S2[1:n2, 3]
sx21Old  <- sd(X21Old)
x21barOld <- mean(X21Old)

Q2 <- qqnorm(X21Old, ylab = "Quantiles of Deaths 2", main = "NQQ plot of Death 2")
qqline(X21Old, col="orange", lwd=3)

cor(Q2$x,Q2$y)

#Part I.D – Parameter Comparisons for Means
#Null hypothesis: H0: μ1 = μ2
#test statistic
xbarD<-x1barOld-x21barOld
SED<-sqrt((sx1Old^2/n)+(sx21Old/n2))
```

```
tcritD<-qt(alpha/2, n+n2-2,lower.tail=F)
epsD<-tcritD*SED
tstatD<-(xbarD-0)/SED

#p-value calculated
pvalD<-2*pt(-abs(tstatD), n+n2-2, lower.tail = T)

#Confidence interval
LowD<-xbarD-epsD
UpperD<-xbarD+epsD

#Summary
metric_nameD <-c("CI.lower", "CI.upper","T.stat","p-value", "alpha")
metric_valD<- c(LowD, UpperD, tstatD, pvalD, alpha)
options(digits =4)
DataSummaryD <- data.frame(metric_nameD, metric_valD)
DataSummaryD


#Part I.E – Parameter Comparisons for Variances
source("nemolm2.r")
#Null hypothesis: H : σ2 = σ2
#test statistic
sx21Old  <- sd(X21Old)
fstatV<-sx1Old^2/sx21Old^2

#Confidence interval
fcritLV<-qf(alpha/2, n-1, n2-1,lower.tail = T )
fcritUV<-qf(alpha/2, n-1, n2-1,lower.tail = F )

#p-value calculated
fstatLV<-min(fstatV, 1/fstatV)
fstatUV<-max(fstatV, 1/fstatV)
pvalFV<-pf(fstatLV, n-1, n2-1, lower.tail = T) + pf(fstatUV, n-1, n2-1, lower.tail = F)

#Summary
metric_nameFV <-c("CI.lower","CI.upper", "T.stat", "p-value","alpha")


metric_valFV <- c(fcritLV, fcritUV, fstatV, pvalFV, alpha)
options(digits =7)
SummaryFV <- data.frame(metric_nameFV, metric_valFV)
SummaryFV



#Part II.A – Simple Linear Regression
#Standardize
X1 <- (X1Old-x1barOld)/sx1Old
X2 <- (X2Old-x2barOld)/sx2Old
X3 <- (X3Old-x3barOld)/sx3Old

sx1  <- 1
sx2  <- 1
sx3  <- 1

x1bar <- 0
x2bar <- 0
x3bar <- 0

ybar <- mean(Y)
sy <- sd(Y)
covs <- cov(X1, Y)
rs <- cor(X1, Y)

SE <- sx1/sqrt(n)

plot(X1, Y, xlab = "Crashes", ylab = "Deaths", main = "Deaths v. Crashes")

beta1hat <- rs*sy/sx1
beta0hat <- ybar - beta1hat*x1bar

SSE <-  sy^2*(n-1)*(1-rs^2)
SE.beta1hat <- (1/sx1)*sqrt(SSE/(n-1)*(n-2))
SE.beta0hat <- sqrt(SSE/(n-2))*sqrt(1/n + (x1bar)^2/(sx1^2*(n-1)))
```

```r
metric.name <-c("covariance", "r value", "r^2 value", "beta1hat", "SE.beta1hat", "beta0hat",
                "SE.beta0hat", "SSE")
metric.val <- c(covs, rs, rs^2, beta1hat, SE.beta1hat, beta0hat, SE.beta0hat, SSE)

D <- data.frame(metric.name, metric.val)
options(digits = 4)
D

# Linear Regression Line
yhat <- lm(Y ~ X1)
abline(yhat, col="steelblue", lwd = 3)

residual <- resid(yhat)
plot(X1, residual, xlab = "Deaths", ylab = "residuals for model",
     main="Residual plot for our linear model")
abline(0, 0, col ="red", lwd=3)

alpha<-0.05
tcrits<-qt(alpha/2, df=n-2, lower.tail = F)
beta1<-0
epsS <- tcrits*SE.beta1hat
tstats <- (beta1hat - beta1)/SE.beta1hat
CIL <- beta1hat - epsS
CIU <- beta1hat + epsS
pvalS <- 2*pt(abs(tstats), df=n-2, lower.tail = F)
pvalS


#Part II.B - Simple Quadratic Regression
M2 <- nemolm2(Y, cbind(X1, X1^2))

#Standardized residual plot
plot(X1, M2$residual, xlab = "Crashes", ylab = "residuals for Simple Quadratic model",
     main="Residual plot for Simple Quadratic model")
abline(0, 0, col ="red", lwd=3)
M2


#Part III - Multiple Linear Regression
M3 <- nemolm2(Y, cbind(X1, X2, X3))
M3

#ANOVA table
metric_name_A <-c("SST", "MST", "SSM", "MSM", "SSE", "MSE", "Fstat", "p-value")
metric_val_A <-c(M3$sst, M3$mst, M3$ssm, M3$msm, M3$sse, M3$mse, M3$fstat, M3$pval)
Summary_A <- data.frame(metric_name_A, metric_val_A)
Summary_A

#Variance inflation factors calculated for each variable with barplot
# Y regressed on X1, X2, and X3
MYvX1c <- nemolm2(Y, cbind(X2, X3))
MX1vX1c <- nemolm2(X1, cbind(X2, X3))

MYvX2c <- nemolm2(Y, cbind(X1, X3))
MX2vX2c <- nemolm2(X2, cbind(X1, X3))

MYvX3c <- nemolm2(Y, cbind(X1, X2))
MX3vX3c <- nemolm2(X3, cbind(X1, X2))

vif1 <- 1/(1-MYvX1c$r2)
vif2 <- 1/(1-MYvX2c$r2)
vif3 <- 1/(1-MYvX3c$r2)

vif <- c(vif1, vif2, vif3)
barplot(vif, horiz=T, main="Variance Inflation Factors",
        names.arg = c('Crashes', 'Miles traveled (millions)', 'Motor vehicles'),
        xlim=c(0,1425))

#new fits
M4 <- nemolm2(Y, cbind(X1, X2, X3, X2*X3))
M4

#Added variable plots for each variable
plot(MYvX1c$sres, MX1vX1c$sres,
     main="Added Variable Plot fot X1",
```

```
     xlab = "S.Residuals for Y~X1c",
     ylab = "S.Residuals for X1~X1c")
abline(0,0, lwd=2)
abline(mean(MX1vX1c$sres)-cor(MYvX1c$sres,
                              MX1vX1c$sres)*sd(MX1vX1c$sres)/sd(MYvX1c$sres)*mean(MYvX1c$sres),
       cor(MYvX1c$sres, MX1vX1c$sres)*sd(MX1vX1c$sres)/sd(MYvX1c$sres))

plot(MYvX2c$sres, MX2vX2c$sres,
     main="Added Variable Plot fot X2",
     xlab = "S.Residuals for Y~X2c",
     ylab = "S.Residuals for X2~X2c")
abline(0,0, lwd=2)
abline(mean(MX2vX2c$sres)-cor(MYvX2c$sres,
                              MX2vX2c$sres)*sd(MX2vX2c$sres)/sd(MYvX2c$sres)*mean(MYvX2c$sres),
       cor(MYvX2c$sres, MX2vX2c$sres)*sd(MX2vX2c$sres)/sd(MYvX2c$sres))

plot(MYvX3c$sres, MX3vX3c$sres,
     main="Added Variable Plot fot X3",
     xlab = "S.Residuals for Y~X3c",
     ylab = "S.Residuals for X3~X3c")
abline(0,0, lwd=2)
abline(mean(MX3vX3c$sres)-cor(MYvX3c$sres,
                              MX3vX3c$sres)*sd(MX3vX3c$sres)/sd(MYvX3c$sres)*mean(MYvX3c$sres),
       cor(MYvX3c$sres, MX3vX3c$sres)*sd(MX3vX3c$sres)/sd(MYvX3c$sres))

#Standardized residual plot with title and axis labels
plot(X1, M2$std.residual, xlab = "Year", ylab = "residuals for model",
     main="Residual plot for Simple Quadratic model")
abline(0, 0, col ="red", lwd=3)

#Construction of the correlation matrix between Y and all three variables
cor(S)

#Part IV - Time Series Fundamentals
plot(S$Year, S$Deaths, type='o')
```

```
#Standardize
XT1 <- S$Year
YT1 <- S$Deaths

sxT1  <- sd(XT1)
xT1 <- mean(XT1)
XT <- (XT1-xT1)/sxT1

syT1  <- sd(YT1)
yT1 <- mean(YT1)
YT <- (YT1-yT1)/syT1

MT<-nemolm2(YT, cbind(XT, XT^2, XT^3, XT^4))
MT
plot(XT, YT, type='o')
lines(XT, MT$predicted, lwd=3, col='green')
```

Nemolm2:

```r
nemolm2 <- function(Y, Xk, ridge=0){
  #ridge = lambda >= 0 (defaulting to 0 results in OLS)
  n <- length(Y)
  v1s <- rep(1, n)
  X <- cbind(v1s,Xk)
  p <- dim(X)[2]-1

  # Ridge Regression If-Statement
  if(ridge != 0){
    lambda = ridge
    S <- svd(t(X)%*%X + lambda^2*diag(p+1))
  }
  else{
    S <- svd(t(X)%*%X)
  }

  # Singular Value Decomposition
  U <- S$u
  D <- diag(S$d)
  V <- S$v

  # Condition Number for XtX
  kappa <- max(S$d)/min(S$d)

  betahat <- V%*%solve(D)%*%t(U)%*%t(X)%*%Y
  Yhat <- X%*%betahat
  H <- X%*%V%*%solve(D)%*%t(U)%*%t(X)
  lv <- diag(H)

  res <- Y - Yhat

  SSE <- sum(res^2)
  MSE <- SSE/(n-p-1)
  SST <- sd(Y)^2*(n-1)
  MST <- SST/(n-1)
  SSM <- SST - SSE
  MSM <- SSM/p



  sres <- res/(sqrt(MSE)*sqrt(1-lv))
  SEbetahat <- sqrt(MSE)*sqrt(diag(V%*%solve(D)%*%t(U)))

  Fstat <- MSM/MSE
  pval <- pf(Fstat, p, n-p-1, lower.tail = F)

  r2 <- 1-SSE/SST
  r2adj <- 1- MSE/MST

  results <- list("predicted" = Yhat,
                  "residual" = res,
                  "sres" = sres,
                  "condition" = kappa,
                  "leverage"= lv,
                  "sse" = SSE,
                  "mse" = MSE,
                  "ssm" = SSM,
                  "msm" = MSM,
                  "pval" = pval,
                  "betahat" = betahat,
                  "SEbetahat" = SEbetahat,
                  "r2" = r2,
                  "r2adj" = r2adj,
                  "sst" = SST,
                  "mst" = MST,
                  "fstat" = Fstat
  )

  return(results)
}
```

Appendix II. dataset

accident-1. csv:

```
Year,Deaths,Crashes, Miles traveled (millions), Motor vehicles,,
1981,49301,44000,1550271,62699
1982,43945,39092,1592481,56455
1983,42589,37976,1649106,55106
1984,44257,39631,1716768,57972
1985,43825,39196,1774762,58272
1986,46087,41090,1838240,60792
1987,46390,41438,1924327,61836
1988,47087,42130,2025586,62703
1989,45582,40741,2107040,60870
1990,44599,39836,2147501,59292
1991,41508,36937,2172214,54795
1992,39250,34942,2239828,52227
1993,40150,35780,2296585,53777
1994,40716,36254,2357588,54911
1995,41817,37241,2422775,56524
1996,42065,37494,2482202,57347
1997,42013,37324,2560373,57060
1998,41501,37107,2625367,56922
1999,41717,37140,2691335,56820
```

accident-2. csv:

```
Year,Deaths,Crashes,Miles traveled (millions),Motor vehicles
2001.00,42196.00,37862.00,2781462.00,57918.00
2002.00,43005.00,38491.00,2855756.00,58426.00
2003.00,42884.00,38477.00,2890893.00,58877.00
2004.00,42836.00,38444.00,2962513.00,58729.00
2005.00,43510.00,39252.00,2989807.00,59495.00
2006.00,42708.00,38648.00,3014116.00,58094.00
2007.00,41259.00,37435.00,3032399.00,56253.00
2008.00,37423.00,34172.00,2973509.00,50660.00
2009.00,33883.00,30862.00,2977591.00,45540.00
2010.00,32999.00,30296.00,2966506.00,44862.00
2011.00,32479.00,29867.00,2946131.00,44119.00
2012.00,33782.00,31006.00,2969433.00,45960.00
2013.00,32894.00,30203.00,2988280.00,45102.00
2014.00,32744.00,30056.00,3025656.00,44950.00
2015.00,35485.00,32539.00,3095373.00,49477.00
2016.00,37806.00,34748.00,3174408.00,52714.00
2017.00,37473.00,34560.00,3212347.00,53128.00
2018.00,36835.00,33919.00,3240327.00,52286.00
2019.00,36096.00,33244.00,3261772.00,51247.00
```