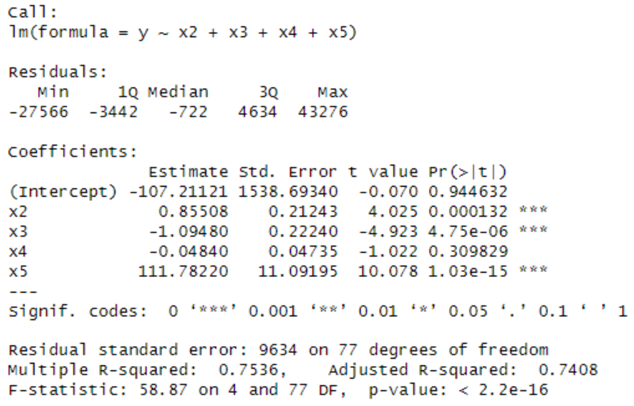
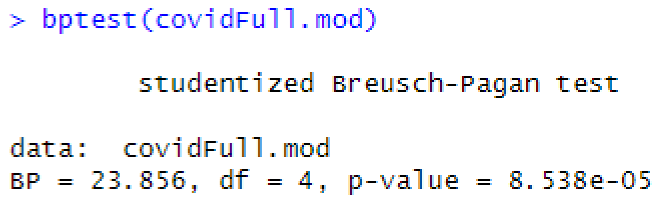
```
# Load all necessary libraries.

library(lmtest)

library(MASS)

library(boot)

library(car)

library(lmridge)

library(caret)

library(fmsb)


# Define Variables

covid <- read.csv("C:/Users/Bella/Desktop/covid.csv")

subset = subset(covid, select = c(Doseone, Series, Booster, death, infection))

data <- subset[apply(subset, 1, function(row) all(row != 0)),]

x2 <- data$Doseone

x3 <- data$Series

x4 <- data$Booster

x5 <- data$death

y <- data$infection


# Define Full Model

covidFull.mod <- lm(y~x2+x3+x4+x5)

summary(covidFull.mod)
```



```
Call:
lm(formula = y ~ x2 + x3 + x4 + x5)

Residuals:
   Min     1Q Median     3Q    Max 
-27566  -3442   -722   4634  43276 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -107.21121 1538.69340  -0.070 0.944632    
x2             0.85508    0.21243   4.025 0.000132 ***
x3            -1.09480    0.22240  -4.923 4.75e-06 ***
x4            -0.04840    0.04735  -1.022 0.309829    
x5           111.78220   11.09195  10.078 1.03e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9634 on 77 degrees of freedom
Multiple R-squared:  0.7536,	Adjusted R-squared:  0.7408 
F-statistic: 58.87 on 4 and 77 DF,  p-value: < 2.2e-16
```

# Check Variance Constancy

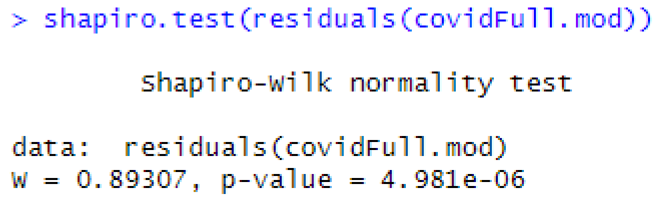
bptest(covidFull.mod)



```
> bptest(covidFull.mod)

	studentized Breusch-Pagan test

data:  covidFull.mod
BP = 23.856, df = 4, p-value = 8.538e-05
```

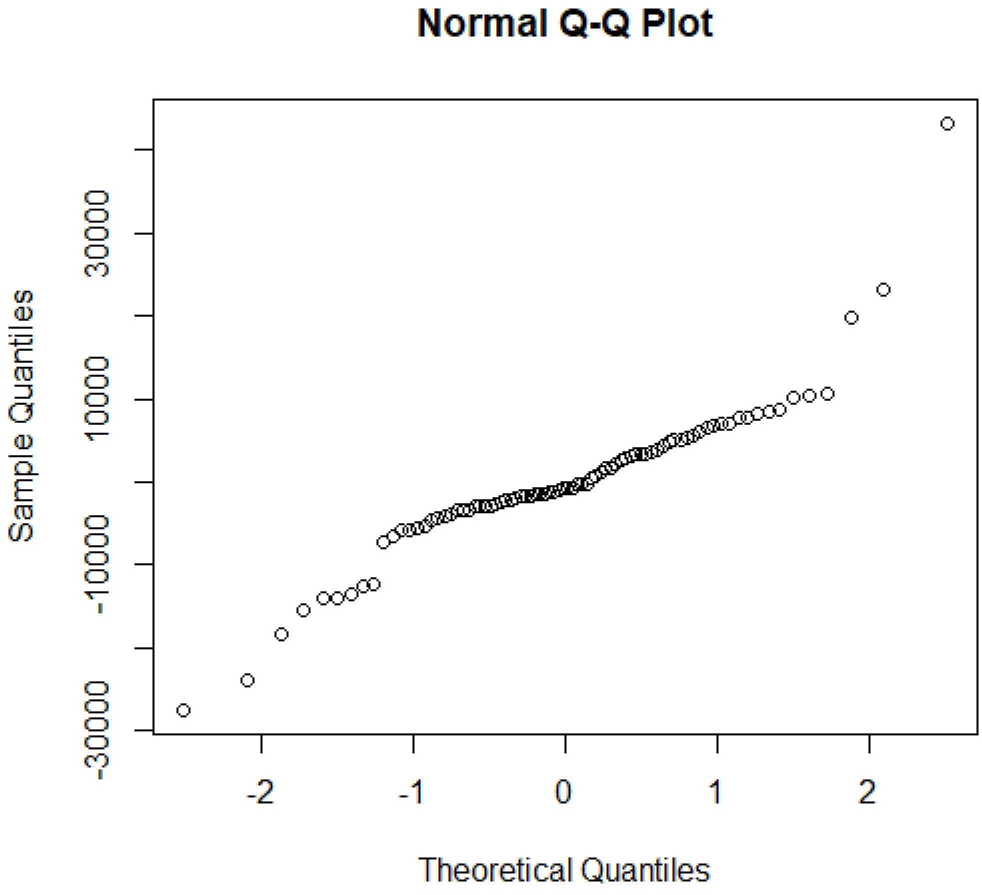
# Check Variance Normality

shapiro.test(residuals(covidFull.mod))



```
> shapiro.test(residuals(covidFull.mod))

	Shapiro-Wilk normality test

data:  residuals(covidFull.mod)
W = 0.89307, p-value = 4.981e-06
```

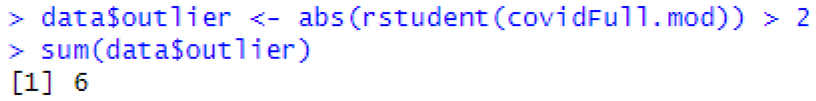
qqnorm(residuals(covidFull.mod))



# Check Outlier

data$outlier <- abs(rstudent(covidFull.mod)) > 2

sum(data$outlier)



```
> data$outlier <- abs(rstudent(covidFull.mod)) > 2
> sum(data$outlier)
[1] 6
```

data <- data[data$outlier == FALSE,]

# Transform X and Y Data

newx2 <- log(data$Doseone)

newx3 <- log(data$Series)

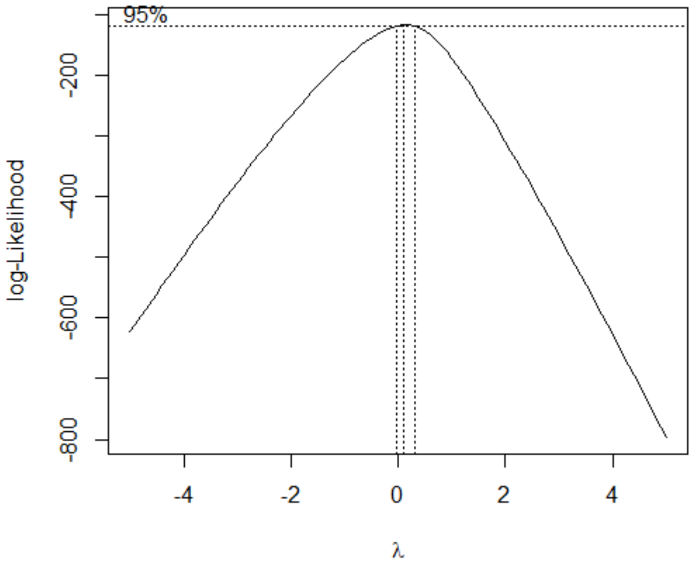
newx4 <- log(data$Booster)

newx5 <- log(data$death)

newy <- data$infection

covidNew.mod <- lm(newy~newx2+newx3+newx4+newx5)

bcmle <- boxcox(covidNew.mod, lambda=seq(-5,5,by=0.1))

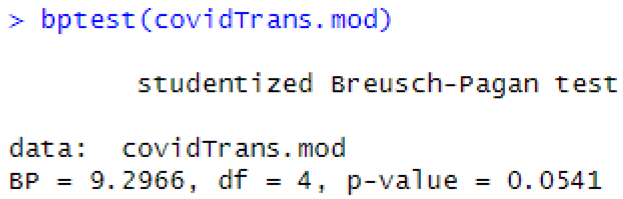


lambda <- bcmle$x[which.max(bcmle$y)]

newy <- data$infection^lambda

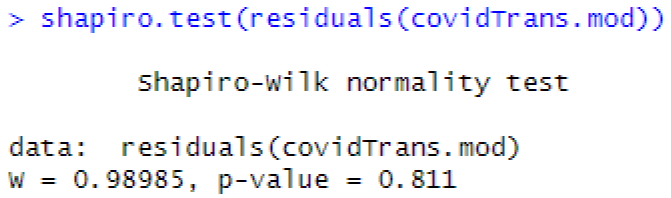
covidTrans.mod <- lm(newy~newx2+newx3+newx4+newx5)

```
# Check Constant Variance

bptest(covidTrans.mod)
```



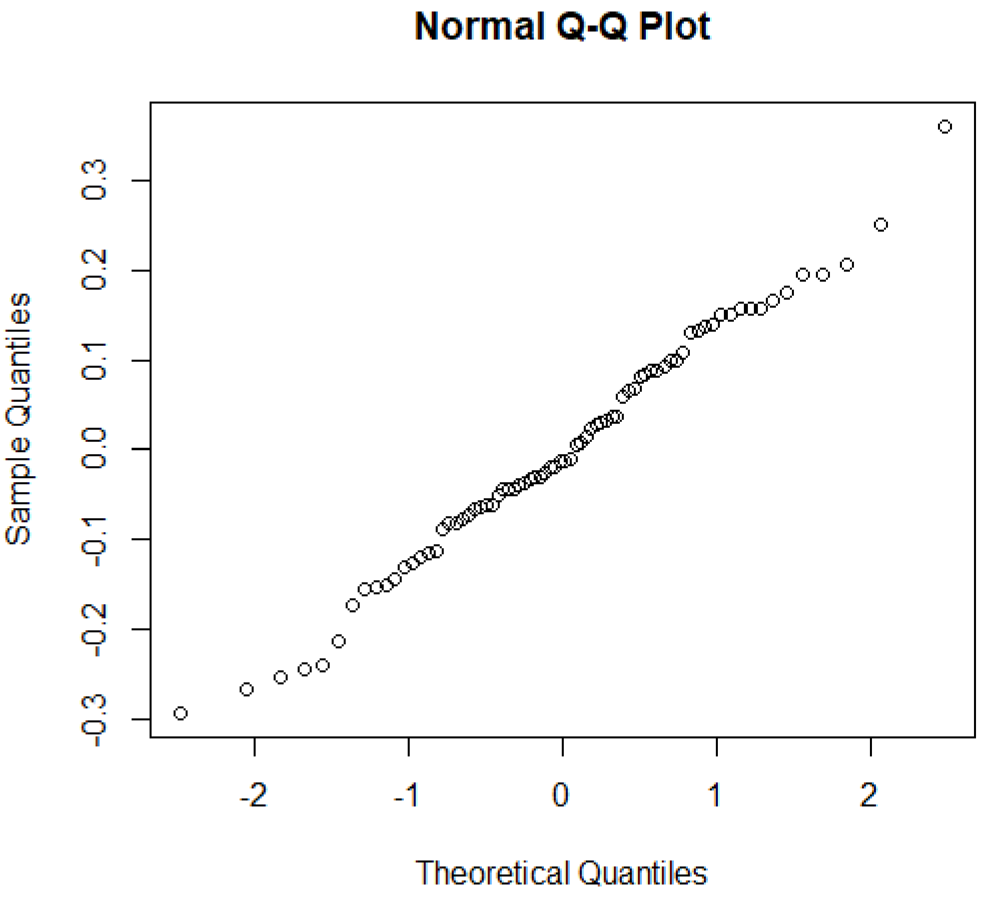
```
> bptest(covidTrans.mod)

	studentized Breusch-Pagan test

data:  covidTrans.mod
BP = 9.2966, df = 4, p-value = 0.0541
```

```
# Check Normality

shapiro.test(residuals(covidTrans.mod))
```

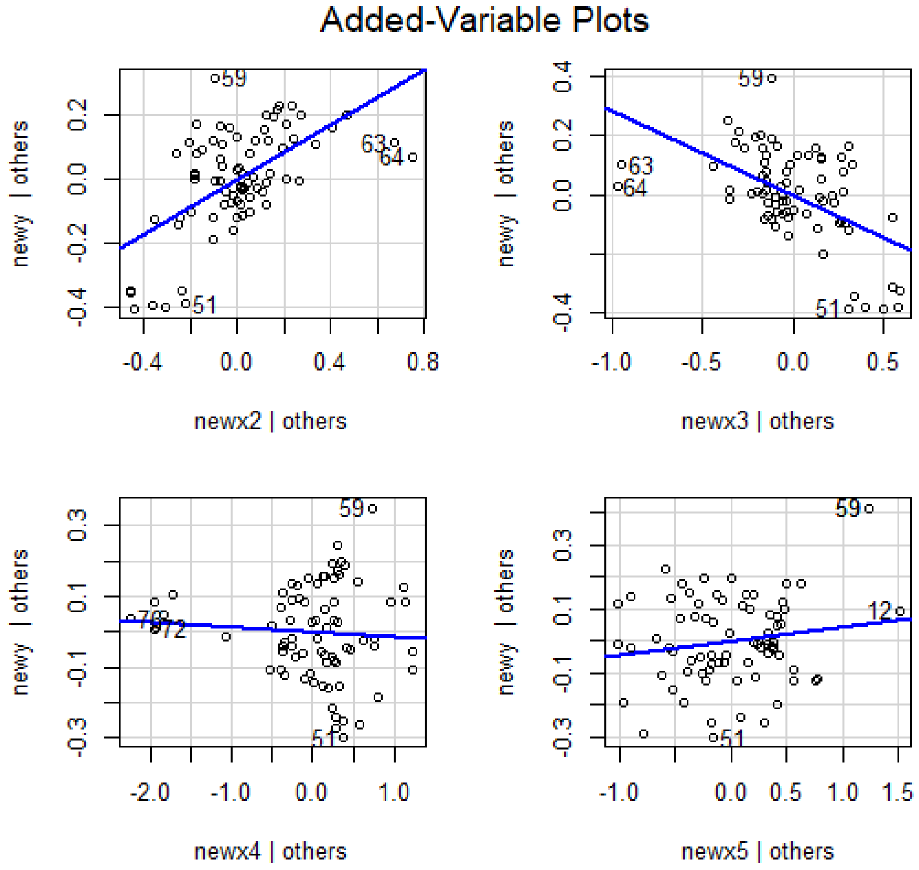


```
> shapiro.test(residuals(covidTrans.mod))

	Shapiro-Wilk normality test

data:  residuals(covidTrans.mod)
W = 0.98985, p-value = 0.811
```

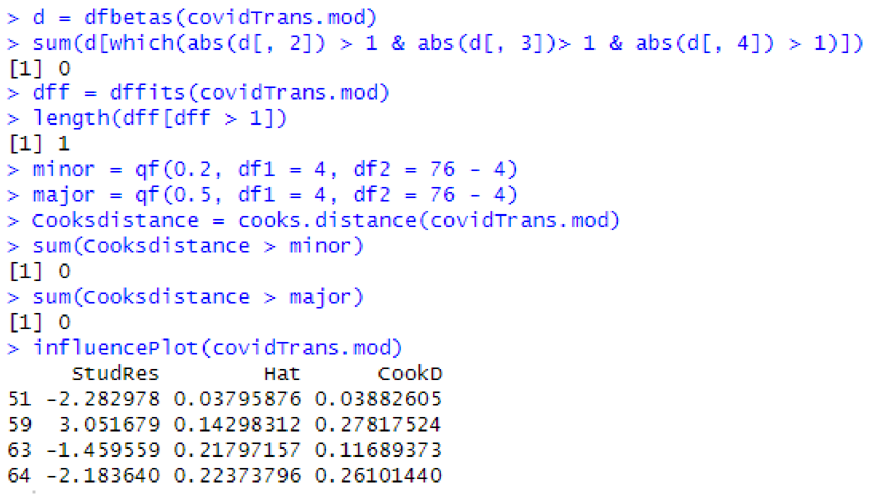
```
qqnorm(residuals(covidTrans.mod))
```



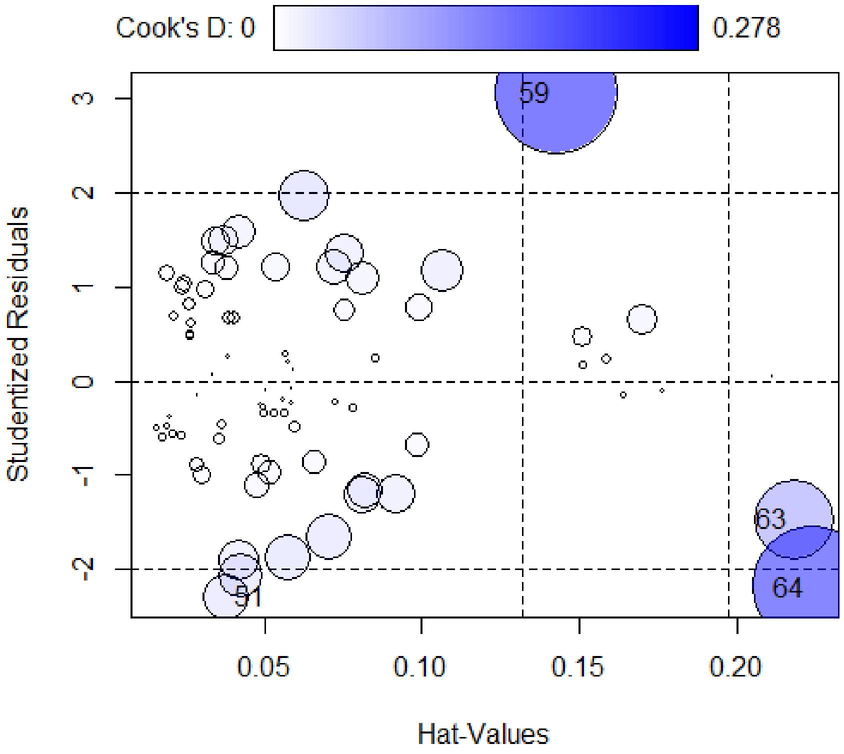
```
# Check Marginal Effect

avPlots(covidTrans.mod)
```



```
# Check Influential Points

d = dfbetas(covidTrans.mod)

sum(d[which(abs(d[, 2]) > 1 & abs(d[, 3])> 1 & abs(d[, 4]) > 1)])

dff = dffits(covidTrans.mod)

length(dff[dff > 1])

minor = qf(0.2, df1 = 4, df2 = 76 - 4)

major = qf(0.5, df1 = 4, df2 = 76 - 4)

Cooksdistance = cooks.distance(covidTrans.mod)

sum(Cooksdistance > minor)

sum(Cooksdistance > major)

influencePlot(covidTrans.mod)
```



```
> d = dfbetas(covidTrans.mod)
> sum(d[which(abs(d[, 2]) > 1 & abs(d[, 3])> 1 & abs(d[, 4]) > 1)])
[1] 0
> dff = dffits(covidTrans.mod)
> length(dff[dff > 1])
[1] 1
> minor = qf(0.2, df1 = 4, df2 = 76 - 4)
> major = qf(0.5, df1 = 4, df2 = 76 - 4)
> Cooksdistance = cooks.distance(covidTrans.mod)
> sum(Cooksdistance > minor)
[1] 0
> sum(Cooksdistance > major)
[1] 0
> influencePlot(covidTrans.mod)
     StudRes        Hat      CookD
51 -2.282978 0.03795876 0.03882605
59  3.051679 0.14298312 0.27817524
63 -1.459559 0.21797157 0.11689373
64 -2.183640 0.22373796 0.26101440
```



# Check Multicollinearity

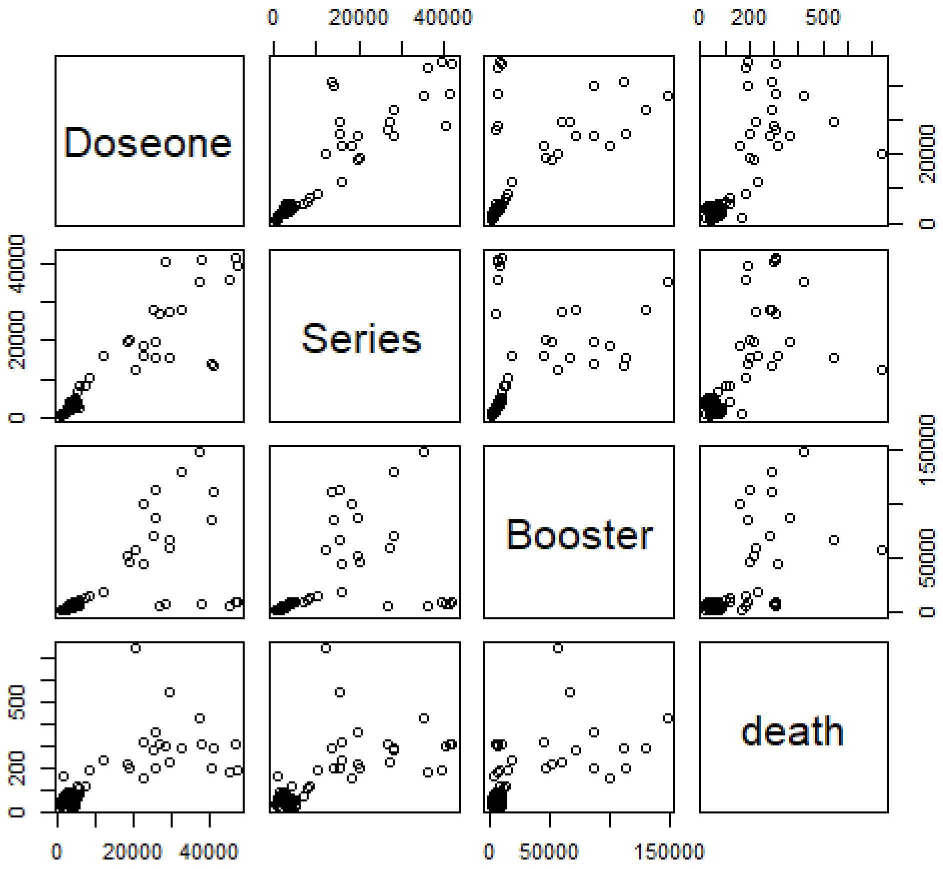
pairs(data[, c("Doseone", "Series", "Booster", "death")])

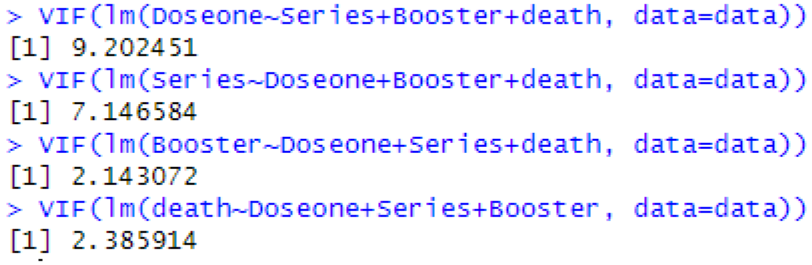
VIF(lm(Doseone~Series+Booster+death, data=data))

VIF(lm(Series~Doseone+Booster+death, data=data))

VIF(lm(Booster~Doseone+Series+death, data=data))

VIF(lm(death~Doseone+Series+Booster, data=data))





```
> VIF(lm(Doseone~Series+Booster+death, data=data))
[1] 9.202451
> VIF(lm(Series~Doseone+Booster+death, data=data))
[1] 7.146584
> VIF(lm(Booster~Doseone+Series+death, data=data))
[1] 2.143072
> VIF(lm(death~Doseone+Series+Booster, data=data))
[1] 2.385914
```

# Ridge Regression

newdata <- data.frame(newx2, newx3, newx4, newx5, newy)

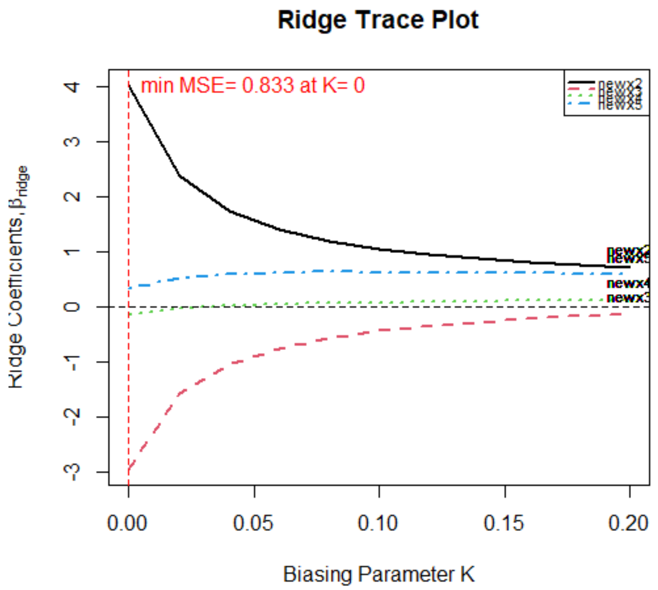
Ridge.mod <- lmridge(newy~newx2+newx3+newx4+newx5, data =newdata, K=seq(0,0.2,0.02))

plot(Ridge.mod)

vif(Ridge.mod)

covidRidge.mod <- lmridge(newy~newx2+newx3+newx4+newx5, data =newdata, K=0.12)

summary(covidRidge.mod)



```
> vif(Ridge.mod)
            newx2    newx3   newx4   newx5
k=0      23.34991 16.96300 3.08289 2.96793
k=0.02    7.77698  6.19320 2.60475 2.33742
k=0.04    3.99406  3.49207 2.26019 2.02151
k=0.06    2.49969  2.37503 1.98702 1.79395
k=0.08    1.75025  1.78400 1.76388 1.61221
k=0.1     1.31640  1.42205 1.57839 1.46094
k=0.12    1.03997  1.17829 1.42217 1.33226
k=0.14    0.85136  1.00301 1.28921 1.22129
k=0.16    0.71594  0.87088 1.17498 1.12460
k=0.18    0.61479  0.76772 1.07606 1.03969
k=0.2     0.53683  0.68496 0.98978 0.96462

> summary(covidRidge.mod)

Call:
lmridge.default(formula = newy ~ newx2 + newx3 + newx4 + newx5,
    data = newdata, K = 0.12)


Coefficients: for Ridge parameter K= 0.12
           Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept   1.4229        -6.4575      2.6821      -2.4076   0.0186 *
newx2       0.1015         0.9483      0.1570       6.0422   <2e-16 ***
newx3      -0.0337        -0.3469      0.1671      -2.0763   0.0414 *
newx4       0.0085         0.0924      0.1835       0.5035   0.6161
newx5       0.0845         0.6244      0.1776       3.5148   0.0008 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
        R2      adj-R2    DF ridge           F          AIC         BIC
   0.43320     0.40960     2.53802    24.81952 -282.57748    52.47370
Ridge minimum MSE= 16.43541 at K= 0.12
P-value for F-test ( 2.53802 , 72.86528 ) = 3.333628e-10
-------------------------------------------------------------------
```

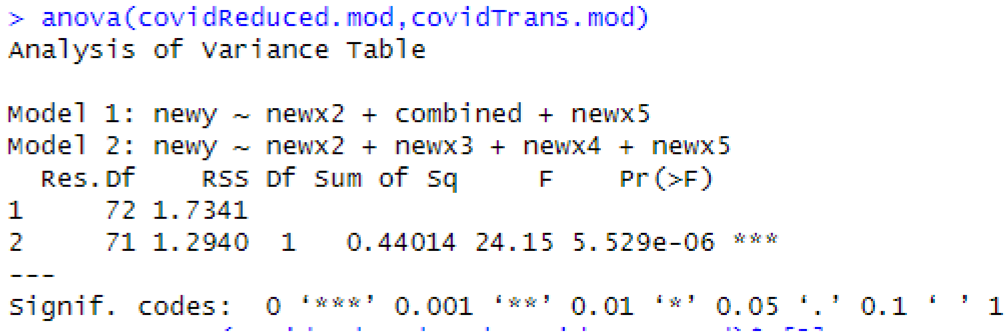
# Reduced Model

combined <- log(data$Series+data$Booster)

covidReduced.mod <- lm(newy~newx2+combined+newx5)

# F test

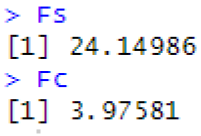
anova(covidReduced.mod,covidTrans.mod)



```
> anova(covidReduced.mod,covidTrans.mod)
Analysis of Variance Table

Model 1: newy ~ newx2 + combined + newx5
Model 2: newy ~ newx2 + newx3 + newx4 + newx5
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1     72 1.7341                                 
2     71 1.2940  1   0.44014 24.15 5.529e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fs <- anova(covidReduced.mod,covidTrans.mod)$F[2]

Fc <- qf(1-0.05, covidReduced.mod$df.residual-covidTrans.mod$df.residual, covidTrans.mod$df.residual)



```
> Fs
[1] 24.14986
> Fc
[1] 3.97581
```

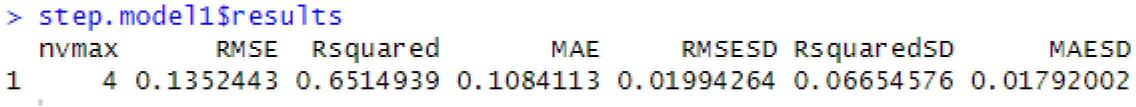
# K-fold Cross Validation

set.seed(123)

train.control <- trainControl(method = 'cv', number = 5)

step.model1 = train(newy ~ newx2 + newx3 + newx4 + newx5, data = newdata, method="leapBackward", tuneGrid = data.frame(nvmax = 4), trControl = train.control)

step.model1$results



```
> step.model1$results
  nvmax      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1     4 0.1352443 0.6514939 0.1084113 0.01994264 0.06654576 0.01792002
```