MH3510: Regression Analysis Project

Name: Kan Kar Shen   Matriculation Number: U1840283A

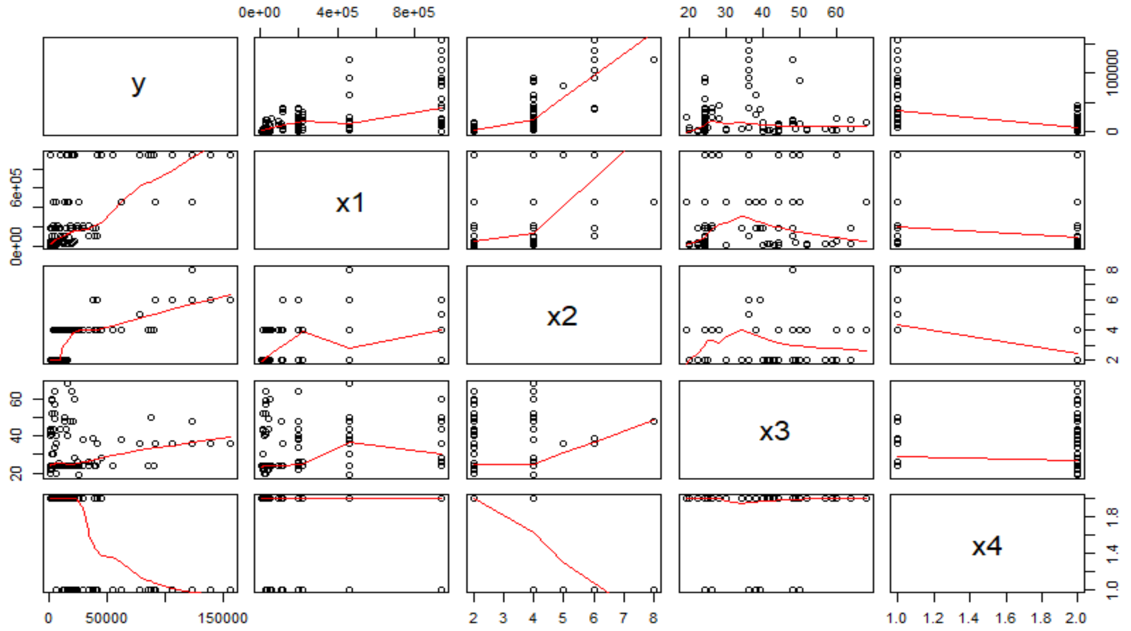
This data are about traffic monitoring. One of the most important traffic monitoring variables is the average annual daily traffic (aadt) for a section of road or highway. It is defined as the average, over a year, of the number of vehicles that pass through a particular section of a road each day.
Consider the first column (aadt) of the data to be the response, and

X1: population of county in which road section in located--the second column of data;
X2: number of lanes in road section-- the third column of data ;
X3: width of road section (in feet)-the fourth column of data;
Control (X4): two-category quality variable indicating whether or not there is control of access to road section (1=access control; 2=no access control)

```
aadt_raw=read.table('C:/Users/karsh/Desktop/Uni/Math Mods/MH3510 Regression Analysis/Project/aadt(2).txt',header=FALSE)
aadt=data.frame(y=aadt_raw$V1, x1=aadt_raw$V2, x2=aadt_raw$V3, x3=aadt_raw$V4, x4=aadt_raw$V5)
plot(aadt)
help("panel.smooth")
plot(aadt[,1:5],panel=panel.smooth)
#
```

1. Scatterplot

The scatterplot between the response and predictor variables is plotted below to observe the relationship between them



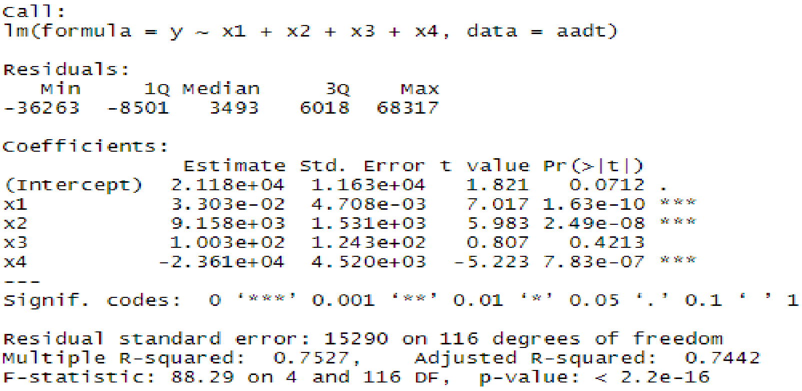
From the scatterplot shown above, there is a relationship between the response variable, y, and the predictor variables, x1, x2, x3 and x4.

2. Multiple Linear Regression Model (MLR)

2.1 MLR1

The model is fitted using the MLR model, $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$

linear_regression1=lm(y~x1+x2+x3+x4,data=aadt)
summary(linear_regression1)
#

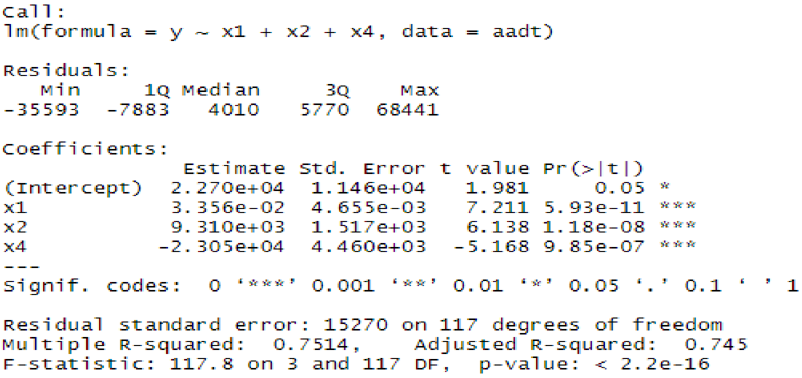


```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-36263  -8501   3493   6018  68317 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.118e+04  1.163e+04   1.821   0.0712 .  
x1           3.303e-02  4.708e-03   7.017 1.63e-10 ***
x2           9.158e+03  1.531e+03   5.983 2.49e-08 ***
x3           1.003e+02  1.243e+02   0.807   0.4213    
x4          -2.361e+04  4.520e+03  -5.223 7.83e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15290 on 116 degrees of freedom
Multiple R-squared:  0.7527,	Adjusted R-squared:  0.7442 
F-statistic: 88.29 on 4 and 116 DF,  p-value: < 2.2e-16
```

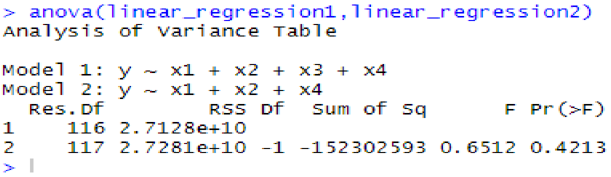
From summary statistics, the hypothesis H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$ cannot be rejected as x3 p-value = 0.4213 > 5%, when tested against 5% level of significance. Hence, x3 will be removed from the model as it does not have a linear relationship with the response variable.

To further confirm the conclusion, F-test is conducted for $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$ and the reduced model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_4$

linear_regression2=lm(y~x1+x2+x4,data=aadt)
summary(linear_regression2)
anova(linear_regression1,linear_regression2)
#



```
Call:
lm(formula = y ~ x1 + x2 + x4, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-35593  -7883   4010   5770  68441 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.270e+04  1.146e+04   1.981     0.05 *  
x1           3.356e-02  4.655e-03   7.211 5.93e-11 ***
x2           9.310e+03  1.517e+03   6.138 1.18e-08 ***
x4          -2.305e+04  4.460e+03  -5.168 9.85e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15270 on 117 degrees of freedom
Multiple R-squared:  0.7514,	Adjusted R-squared:  0.745 
F-statistic: 117.8 on 3 and 117 DF,  p-value: < 2.2e-16
```



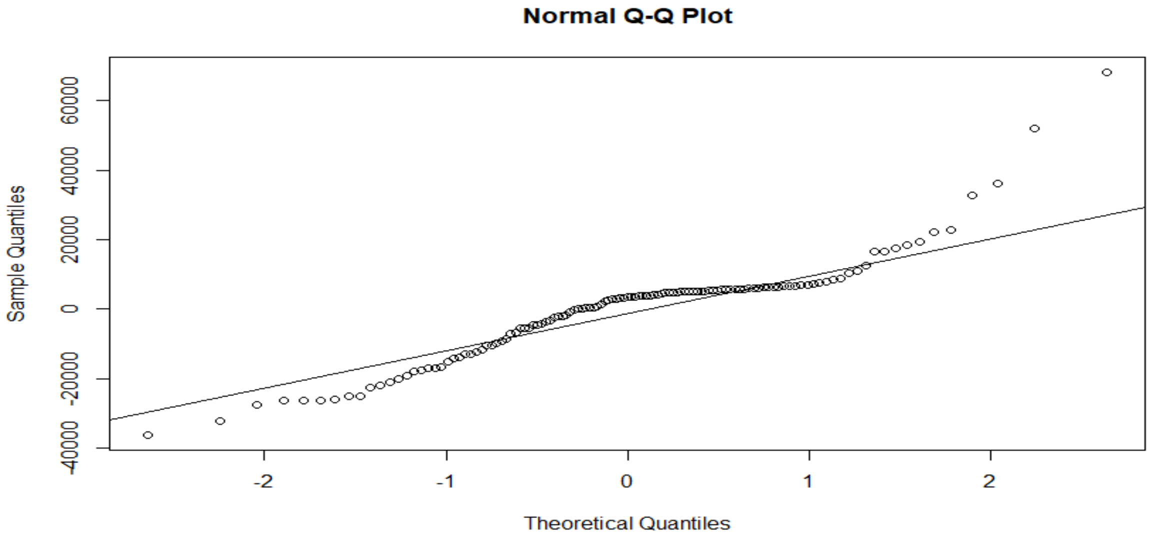
```
> anova(linear_regression1,linear_regression2)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4
Model 2: y ~ x1 + x2 + x4
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1    116 2.7128e+10                            
2    117 2.7281e+10 -1 -152302593 0.6512 0.4213
> |
```

From the F-test, p-value = 0.4213 > 1% hence, null the hypothesis cannot be rejected at 1% level of significance and x3 is removed from the model as there is no significant linear relationship with the response variable.

2.1.1 QQ-Plot

We now use the QQ-Plot to check for normality of the MLR1 as shown below

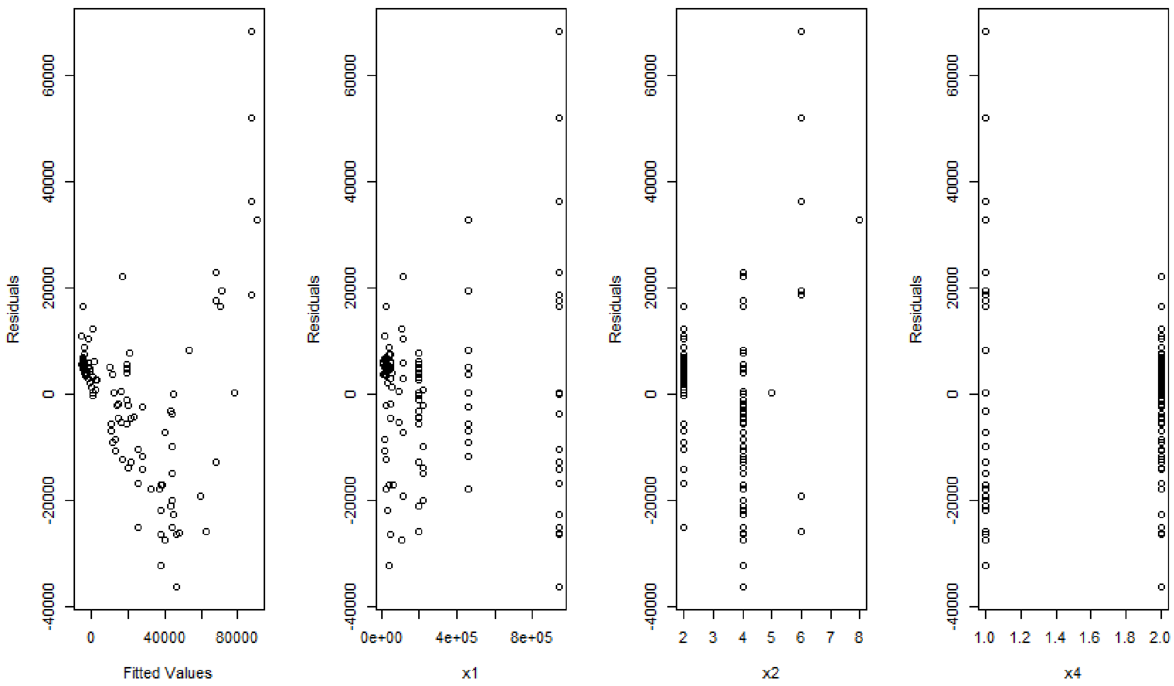
```
qqnorm(residuals(linear_regression,ylab='Residuals'))
qqline(residuals(linear_regression))
```



As shown in the QQ-Plot above, the residuals deviate from linearity hence, the residuals are not normally distributed.

We now plot the residuals against the fitted values and predictor variables, x1, x2, x4. They are shown below.

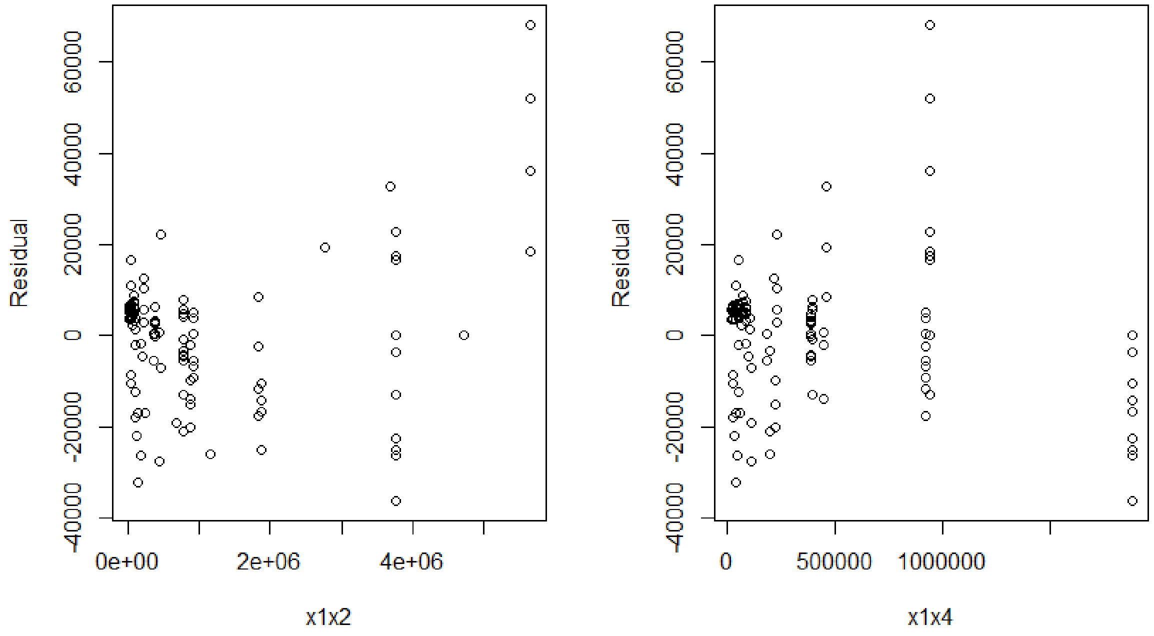
```
par(mfrow=c(1,5))
plot(linear_regression$fitted.values,linear_regression$residuals,ylab="Residuals",xlab="Fitted
Values")
plot(aadt$x1,linear_regression$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,linear_regression$residuals,ylab="Residuals",xlab="x2")
plot(aadt$x4,linear_regression$residuals,ylab="Residuals",xlab="x4")esidu
par(mfrow=c(1,1))
```



From the graphs, residuals plotted against x2 and x4, there is no observable relationship between the residuals and the predictor variables, x2 and x4.
From the graph, residuals plotted against fitted values, the residuals variance increases as the fitted value increases along the graph. Therefore, it is observed that the errors variances are not constant.
From the graph, residuals plotted against x1, it is observed that the residuals and x1 have a linear relationship.

Since there is a linear relationship between the residuals and x1, the residuals will be checked against x1 with the other predictor variables (x2 and x4).

```
par(mfrow=c(1,2))
plot(aadt$x1*aadt$x2,linear_regression$residuals,ylab="Residual",xlab="x1x2")
plot(aadt$x1*aadt$x4,linear_regression$residuals,ylab="Residual",xlab="x1x4")
```



From the graph shown above, there is no observable relationship between the residuals and x1x2 and x1x4. Hence both x1x2 and x1x4 will be added into the model for further testing. Therefore, the model to be tested will be as follow, Y = $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_4 + \beta_4I(X_1X_2) + \beta_5I(X_1X_4)$.

2.2 MLR Model to Test

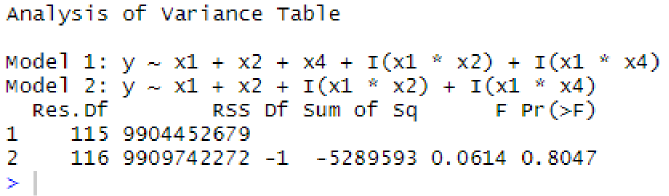
The model Y = $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_4 + \beta_4I(X_1X_2) + \beta_5I(X_1X_4)$ will be tested to find the appropriate fit.

```
mlr=lm(y~x1+x2+x4+I(x1*x2)+I(x1*x4),data=aadt)
summary(mlr)
#
Call:
lm(formula = y ~ x1 + x2 + x4 + I(x1 * x2) + I(x1 * x4), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-33661  -2001   -674   2412  34514 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.602e+03  9.744e+03  -0.780 0.436900    
x1           9.195e-02  2.108e-02   4.362 2.83e-05 ***
x2           5.968e+03  1.288e+03   4.634 9.53e-06 ***
x4          -9.456e+02  3.816e+03  -0.248 0.804712    
I(x1 * x2)   1.089e-02  2.780e-03   3.915 0.000154 ***
I(x1 * x4)  -5.765e-02  7.766e-03  -7.424 2.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9280 on 115 degrees of freedom
Multiple R-squared:  0.9097,	Adjusted R-squared:  0.9058 
F-statistic: 231.8 on 5 and 115 DF,  p-value: < 2.2e-16
```

From summary statistics, the hypothesis H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$ cannot be rejected as x4 p-value = 0.8368 > 1%, when tested against 1% level of significance.

To further confirm the conclusion, F-test is conducted for $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_4 + \beta_4I(X_1X_2) + \beta_5I(X_1X_4)$ and the reduced model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3I(X_1X_2) + \beta_4I(X_1X_4)$

```
mlr_reduced=lm(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt)
anova(mlr,mlr_reduced)
#
```



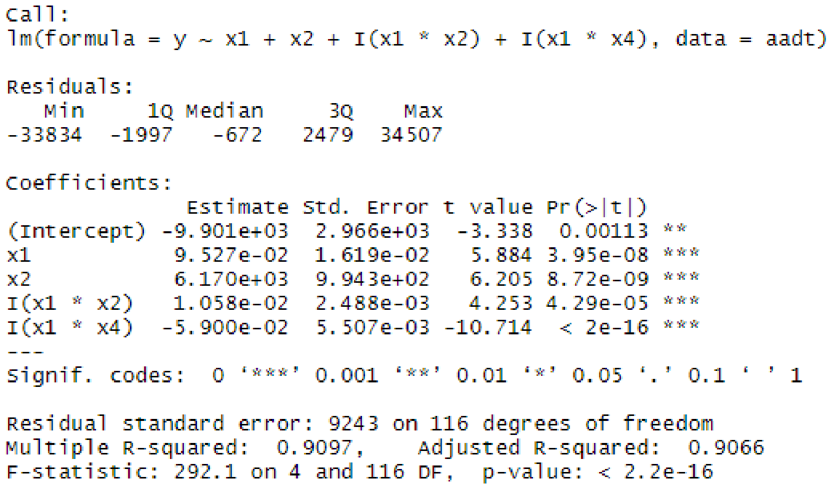
```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x4 + I(x1 * x2) + I(x1 * x4)
Model 2: y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    115 9904452679                            
2    116 9909742272 -1  -5289593 0.0614 0.8047
> |
```

From the F-test, the p-value = 0.8047 > 1%, hence the null hypothesis cannot be rejected at 1% level of significance. Hence x_4 is removed from the model as there is no significant relationship with the response variable.

2.3 MLR Reduced Model

Therefore, the reduced model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3I(X_1X_2) + \beta_4I(X_1X_4)$ will be used.

```
summary(mlr_reduced)
#
```



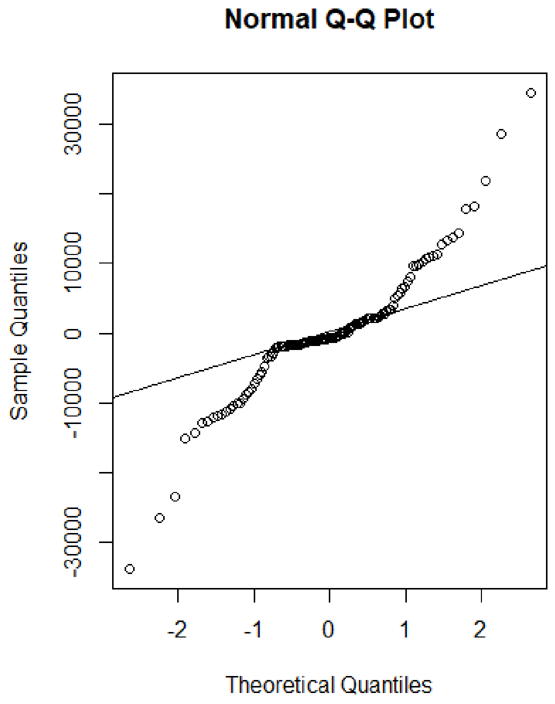
```
Call:
lm(formula = y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-33834  -1997   -672   2479  34507 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.901e+03  2.966e+03  -3.338  0.00113 ** 
x1           9.527e-02  1.619e-02   5.884 3.95e-08 ***
x2           6.170e+03  9.943e+02   6.205 8.72e-09 ***
I(x1 * x2)   1.058e-02  2.488e-03   4.253 4.29e-05 ***
I(x1 * x4)  -5.900e-02  5.507e-03 -10.714  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9243 on 116 degrees of freedom
Multiple R-squared:  0.9097,	Adjusted R-squared:  0.9066 
F-statistic: 292.1 on 4 and 116 DF,  p-value: < 2.2e-16
```

From the summary statistic, the p-values of all the predictor variables are < 1% when tested against 1% level of significance. Hence, the null hypothesis is rejected and can conclude that there is a significant linear relationship between the response variable and predictor variables.

2.3.1 QQ-Plot

We now use the QQ-Plot to check for normality of the MLR Reduced Model as shown below

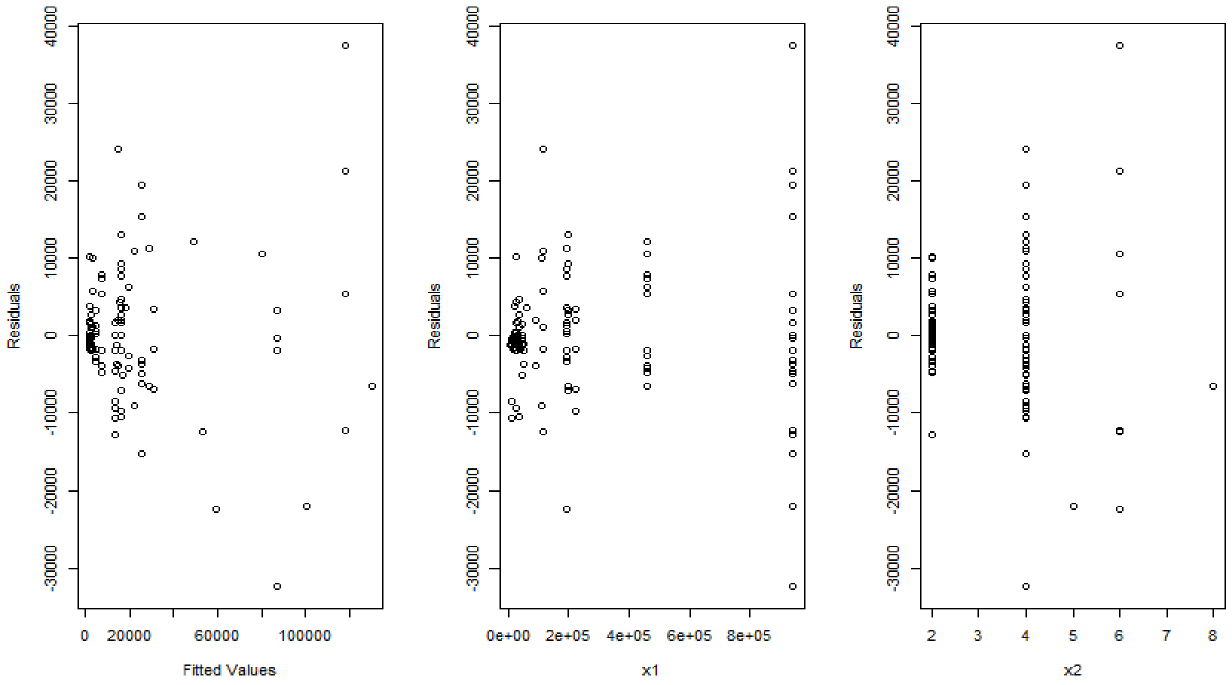
```
qqnorm(residuals(mlr_reduced,ylab="Residuals"))
qqline(residuals(mlr_reduced))
#
```



As shown in the QQ-Plot above, the residuals of the MLR Reduced Model deviate from linearity hence, the residuals are not normally distributed.

We now plot the residuals against the fitted values and predictor variables, x1, x2. They are shown below.

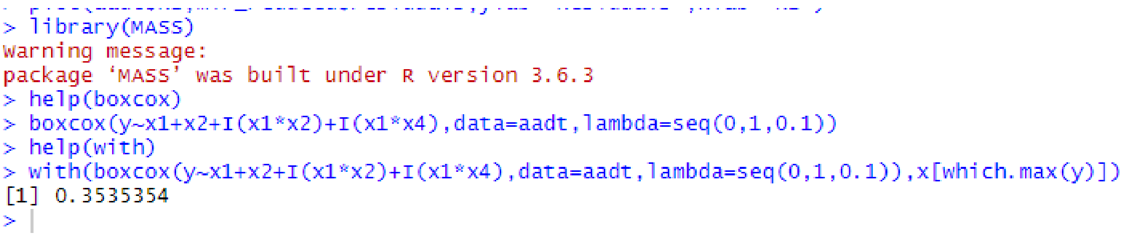
```
par(mfrow=c(1,3))
plot(mlr_reduced$fitted.values,mlr_reduced$residuals,ylab="Residuals",xlab="Fitted Values")
plot(aadt$x1,mlr_reduced$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,mlr_reduced$residuals,ylab="Residuals",xlab="x2")
#
```



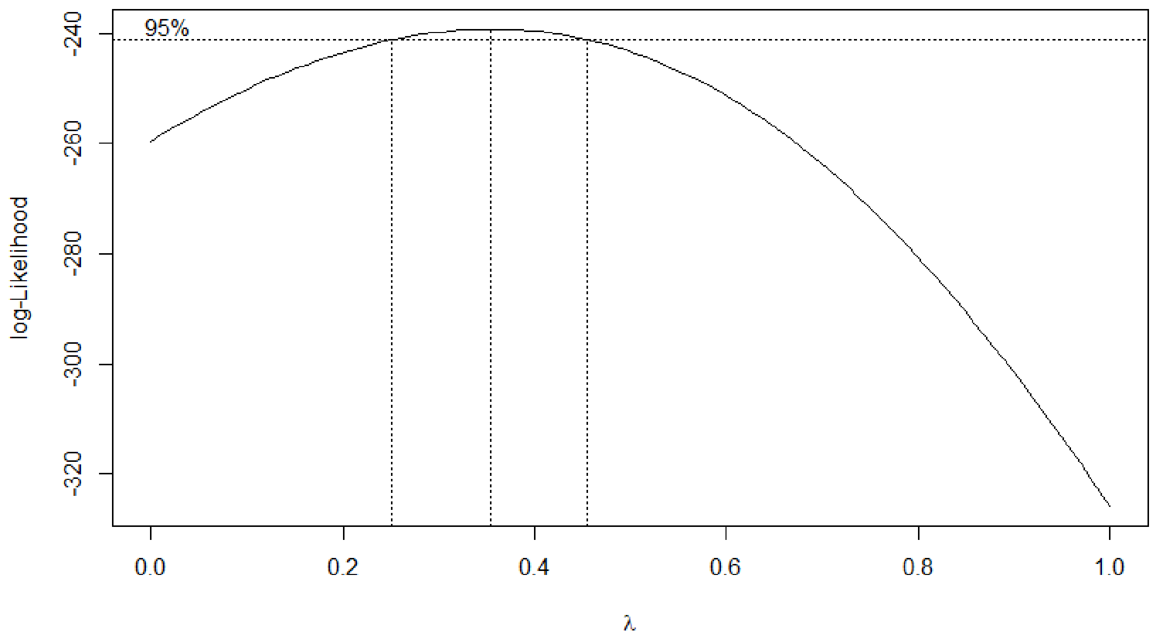
From the residual plots shown above,
From the residuals plotted against x1 and x2 respectively, there is no observable relationship between the residuals and x1 and x2.
From the residuals plotted against fitted values, the variance appears to increase as fitted values increases hence, the error variances are not constant.

To transform the MLR Reduced Model and QQ-plot into a normal shape, box-cox transformation, is used. The graph and value of lambda is as shown below.

```
library(MASS)
help(boxcox)
boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))
help(with)
with(boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1)),x[which.max(y)])
#
```

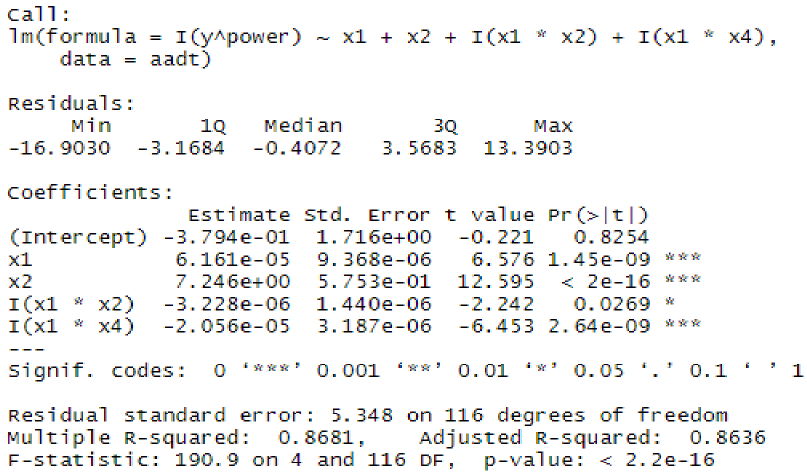


```
> library(MASS)
Warning message:
package 'MASS' was built under R version 3.6.3
> help(boxcox)
> boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))
> help(with)
> with(boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1)),x[which.max(y)])
[1] 0.3535354
> 
```



As shown, the value of lambda = 0.3535354 is the optimal fit. Hence, the reduced model will be transformed. The new reduced model is $Y^{0.3535354} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3I(X_1X_2) + \beta_4I(X_1X_4)$. Similarly, to check the reduced model.

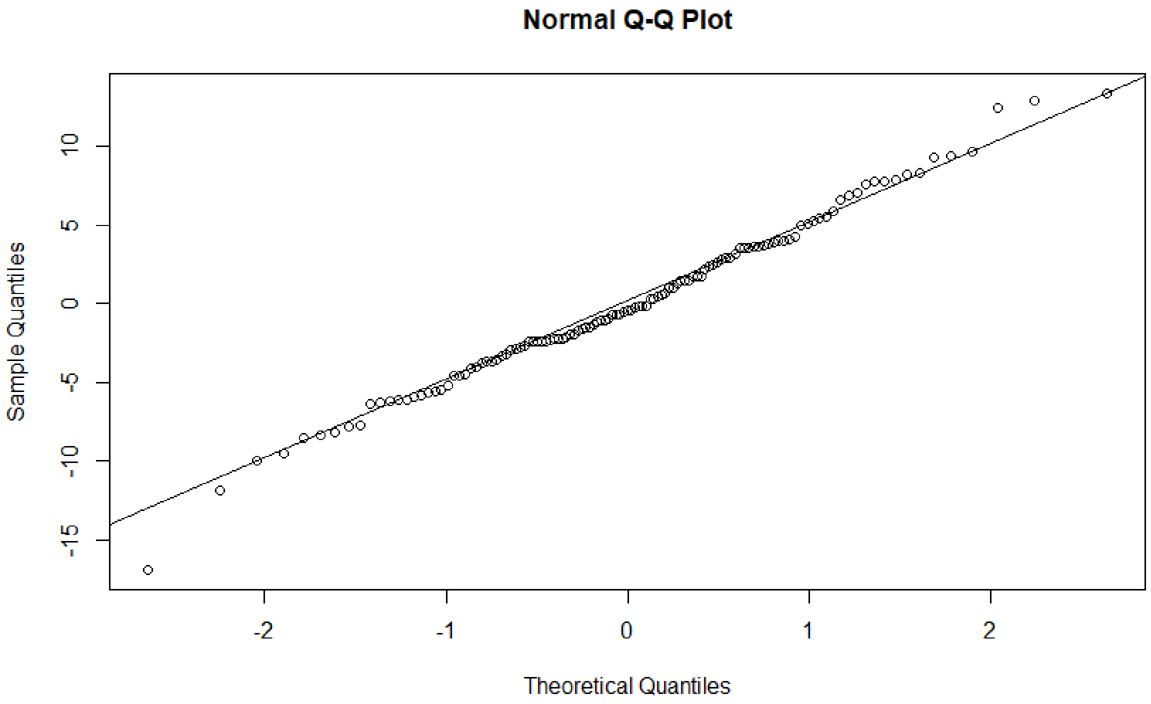
power=boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))$x[which.max(boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))$y)]
mlr_reduced_transformed=lm(I(y^power)~x1+x2+I(x1*x2)+I(x1*x4),data=aadt)
summary(mlr_reduced_transformed)
#



```
Call:
lm(formula = I(y^power) ~ x1 + x2 + I(x1 * x2) + I(x1 * x4), 
    data = aadt)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.9030  -3.1684  -0.4072   3.5683  13.3903 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.794e-01  1.716e+00  -0.221   0.8254    
x1           6.161e-05  9.368e-06   6.576 1.45e-09 ***
x2           7.246e+00  5.753e-01  12.595  < 2e-16 ***
I(x1 * x2)  -3.228e-06  1.440e-06  -2.242   0.0269 *  
I(x1 * x4)  -2.056e-05  3.187e-06  -6.453 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.348 on 116 degrees of freedom
Multiple R-squared:  0.8681,	Adjusted R-squared:  0.8636 
F-statistic: 190.9 on 4 and 116 DF,  p-value: < 2.2e-16
```

The QQ-Plot of the transformed model is as follow,

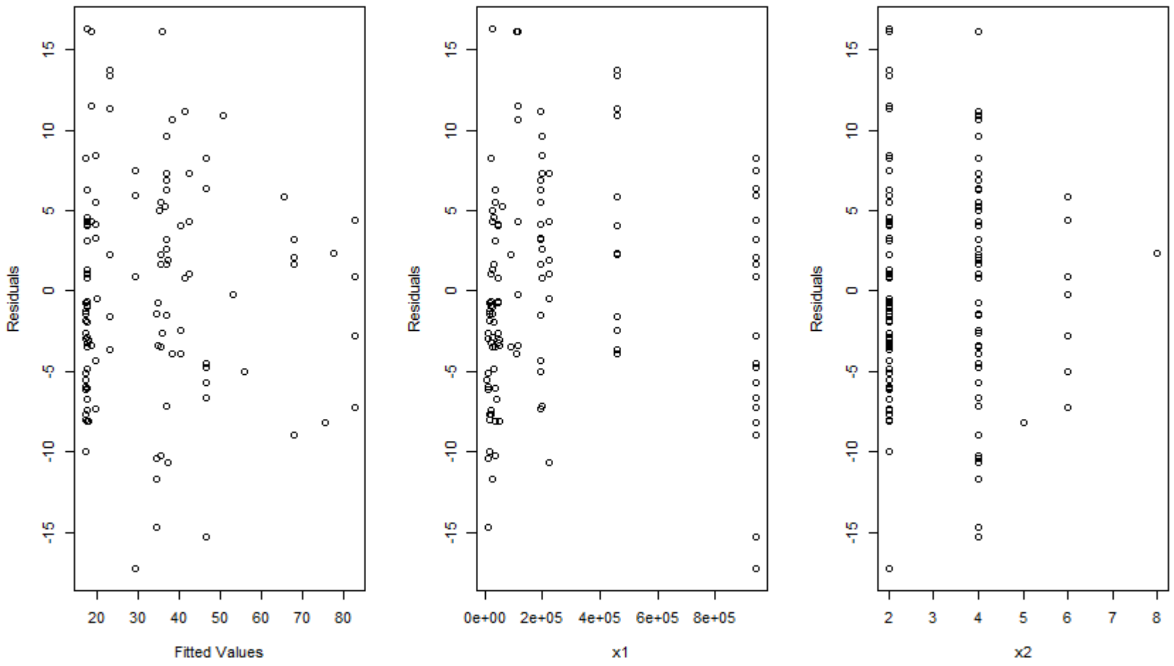
```
qqnorm(residuals(mlr_reduced_transformed,ylab="Residuals"))
is.recursive(residuals(mlr_reduced_transformed))
is.atomic(residuals(mlr_reduced_transformed))
class(residuals(mlr_reduced_transformed))
class(residuals(mlr_reduced))
qqline(as.numeric(residuals(mlr_reduced_transformed)))
#
```



As shown In the Q-Q Plot above, the residuals of the **transformed** MLR Reduced Model is normally distributed and do not deviate from linearity.

We now plot the residuals against the fitted values and predictor variables, x1, x2. They are shown below.

```
par(mfrow=c(1,3))
plot(mlr_reduced_transformed$fitted.values,mlr_reduced_transformed$residuals,ylab="Residuals",
xlab="Fitted Values")
plot(aadt$x1,mlr_reduced_transformed$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,mlr_reduced_transformed$residuals,ylab="Residuals",xlab="x2")
#
```



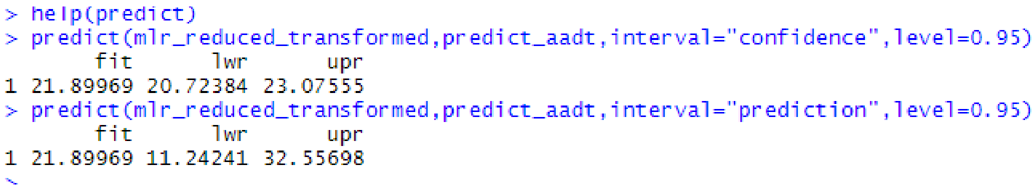
From the residual plots shown above,
From the residuals plotted against x1 and x2 respectively, there is no observable relationship between the residuals and x1 and x2.
From the residuals plotted against fitted values, the variances of errors appear to be constant.

Conclusion

The model, $Y^{0.3535354} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3I(X_1X_2) + \beta_4I(X_1X_4)$ is a good fit of the dataset. In addition, the adjusted R-squared value = 0.8636 is above 0.60 and below 0.95 which implies that the model is a good fit.

Prediction

The values, x_1=50000 x_2=3 x_3=60, x_4=2 are used for prediction.



```
> help(predict)
> predict(mlr_reduced_transformed,predict_aadt,interval="confidence",level=0.95)
       fit      lwr      upr
1 21.89969 20.72384 23.07555
> predict(mlr_reduced_transformed,predict_aadt,interval="prediction",level=0.95)
       fit      lwr      upr
1 21.89969 11.24241 32.55698
>
```

The predicted values are as shown above.

Full R Code

```
aadt_raw=read.table('C:/Users/karsh/Desktop/Uni/Math Mods/MH3510 Regression
Analysis/Project/aadt(2).txt',header=FALSE)
aadt=data.frame(y=aadt_raw$V1, x1=aadt_raw$V2, x2=aadt_raw$V3, x3=aadt_raw$V4,
x4=aadt_raw$V5)
plot(aadt)
help("panel.smooth")
plot(aadt[,1:5],panel=panel.smooth)
```

```
#

linear_regression1=lm(y~x1+x2+x3+x4,data=aadt)
summary(linear_regression1)
#

linear_regression2=lm(y~x1+x2+x4,data=aadt)
summary(linear_regression2)
anova(linear_regression1,linear_regression2)
#

qqnorm(residuals(linear_regression,ylab='Residuals'))
qqline(residuals(linear_regression))
#

par(mfrow=c(1,4))
plot(linear_regression$fitted.values,linear_regression$residuals,ylab="Residuals",xlab="Fitted
Values")
plot(aadt$x1,linear_regression$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,linear_regression$residuals,ylab="Residuals",xlab="x2")
plot(aadt$x4,linear_regression$residuals,ylab="Residuals",xlab="x4")
par(mfrow=c(1,1))
#

par(mfrow=c(1,2))
plot(aadt$x1*aadt$x2,linear_regression$residuals,ylab="Residual",xlab="x1x2")
plot(aadt$x1*aadt$x4,linear_regression$residuals,ylab="Residual",xlab="x1x4")
#

mlr=lm(y~x1+x2+x4+I(x1*x2)+I(x1*x4),data=aadt)
summary(mlr)
#

mlr_reduced=lm(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt)
anova(mlr,mlr_reduced)
#

summary(mlr_reduced)
#

qqnorm(residuals(mlr_reduced,ylab="Residuals"))
qqline(residuals(mlr_reduced))
#

par(mfrow=c(1,3))
plot(mlr_reduced$fitted.values,mlr_reduced$residuals,ylab="Residuals",xlab="Fitted Values")
plot(aadt$x1,mlr_reduced$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,mlr_reduced$residuals,ylab="Residuals",xlab="x2")
```

```
#

library(MASS)
help(boxcox)
boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))
help(with)
with(boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1)),x[which.max(y)])
#

power=boxcox(y~x1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))$x[which.max(boxcox(y~x
1+x2+I(x1*x2)+I(x1*x4),data=aadt,lambda=seq(0,1,0.1))$y)]
mlr_reduced_transformed=lm(I(y^power)~x1+x2+I(x1*x2)+I(x1*x4),data=aadt)
summary(mlr_reduced_transformed)
#

qqnorm(residuals(mlr_reduced_transformed,ylab="Residuals"))
is.recursive(residuals(mlr_reduced_transformed))
is.atomic(residuals(mlr_reduced_transformed))
class(residuals(mlr_reduced_transformed))
class(residuals(mlr_reduced))
qqline(as.numeric(residuals(mlr_reduced_transformed)))
#

par(mfrow=c(1,3))
plot(mlr_reduced_transformed$fitted.values,mlr_reduced_transformed$residuals,ylab="Residuals",
xlab="Fitted Values")
plot(aadt$x1,mlr_reduced_transformed$residuals,ylab="Residuals",xlab="x1")
plot(aadt$x2,mlr_reduced_transformed$residuals,ylab="Residuals",xlab="x2")
#

predict_aadt=data.frame(x1=50000,x2=3,x3=60,x4=2)
help(predict)
predict(mlr_reduced_transformed,predict_aadt,interval="confidence",level=0.95)
predict(mlr_reduced_transformed,predict_aadt,interval="prediction",level=0.95)
#
```