NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

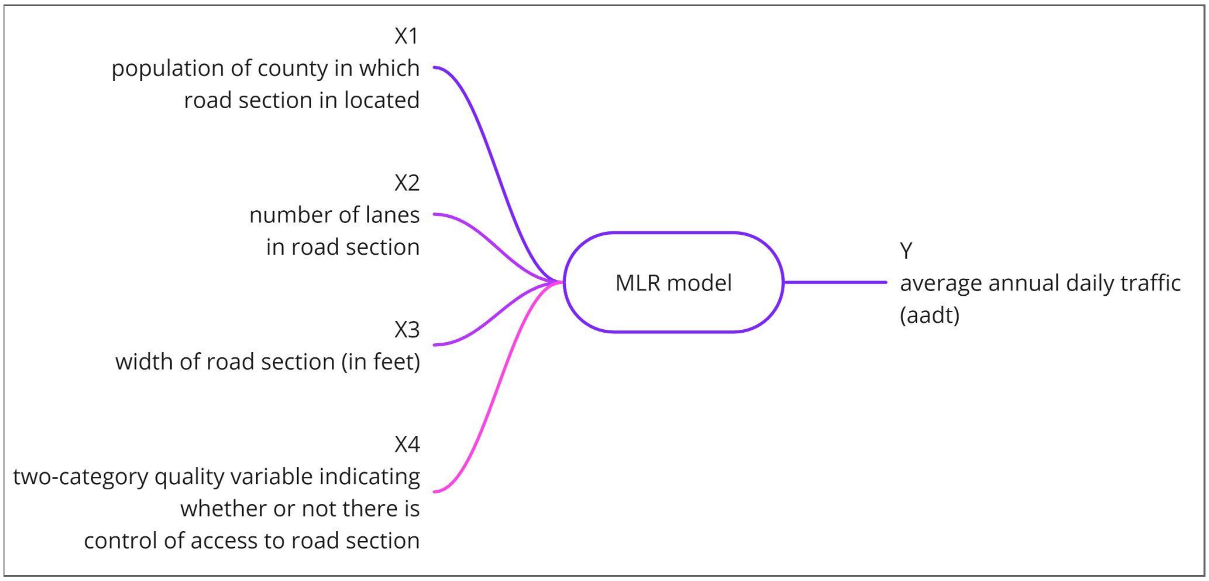
# MH3510 Regression Analysis

# Project Report

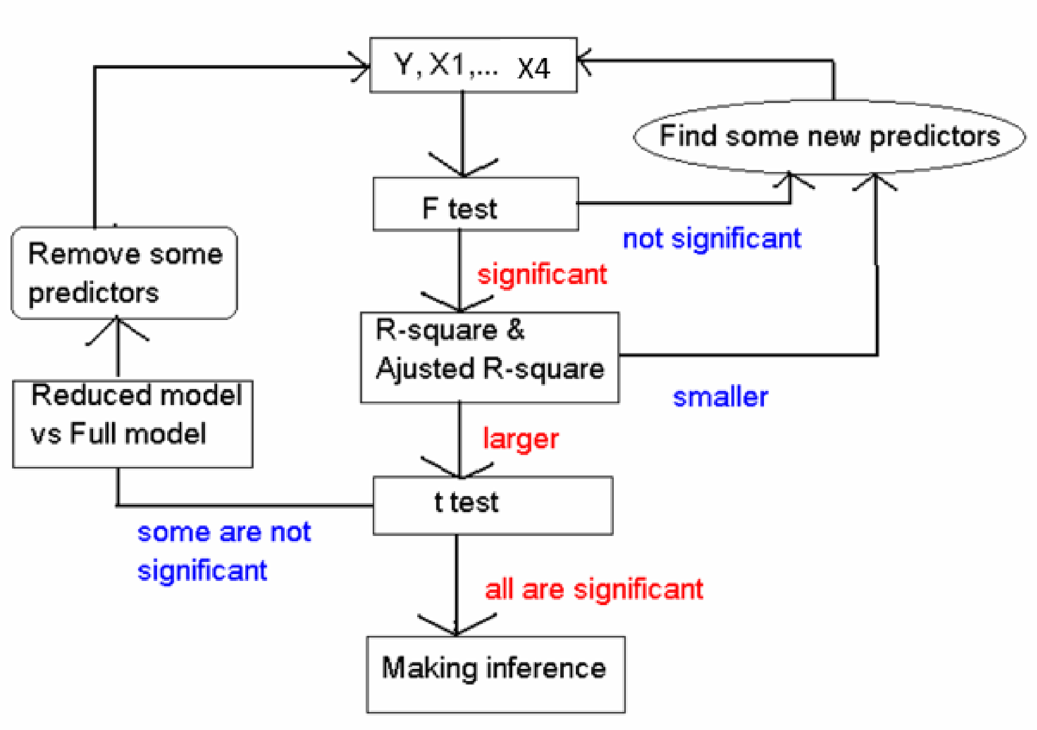
| Group Member |
|---|
| Tiviatis |
| Heen Sunn |
| Geremie |
| Nicholas |
| Jeremy |
| Vaishnavi |
| Harini |
| Danendra |
| Andrew |
| Qi Xiang |

# 1.0 - General Procedure

The inputs and output of the linear base model are as followed:

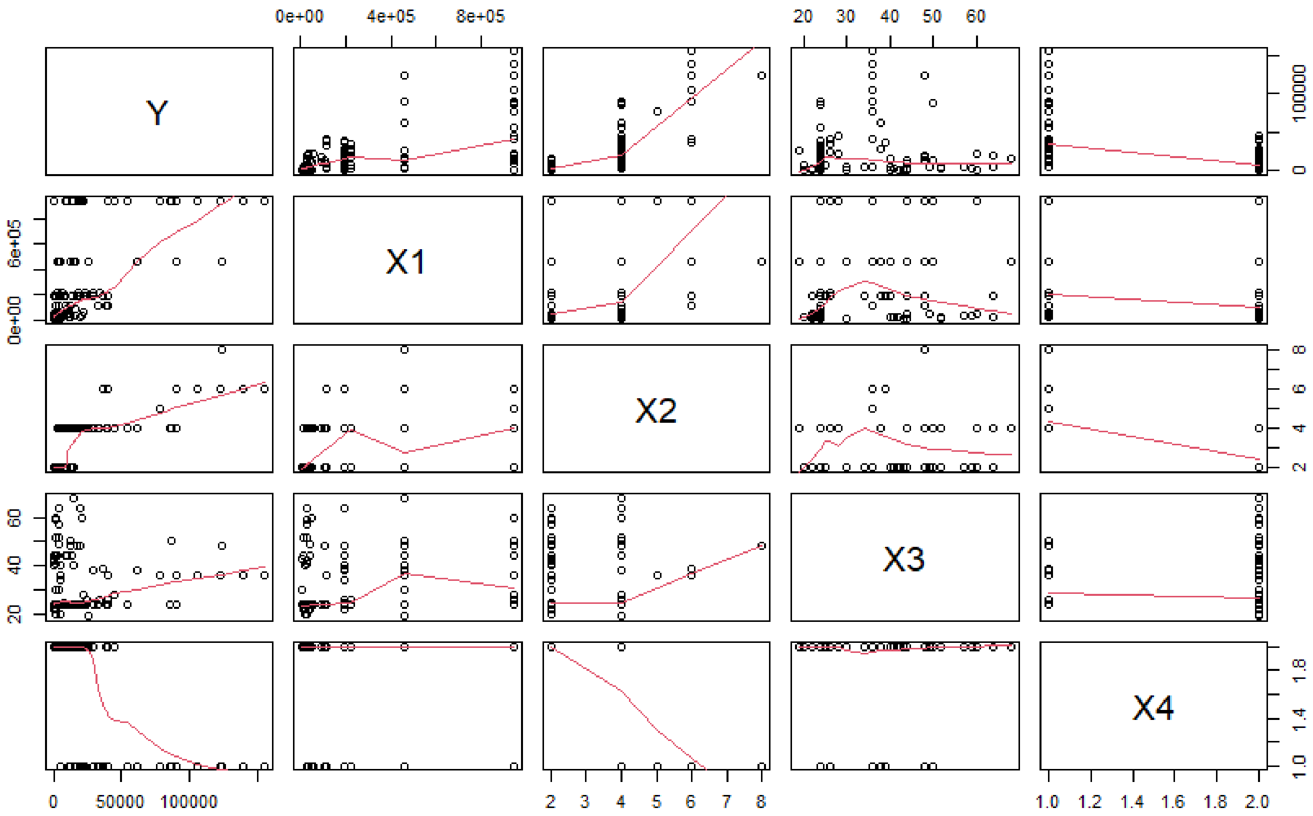

X1
population of county in which road section in located
X2
number of lanes in road section
X3
width of road section (in feet)
X4
two-category quality variable indicating whether or not there is control of access to road section
MLR model
Y
average annual daily traffic (aadt)

Since there are 4 predictors in this MLR model, so p = 4. We will follow the procedure given as followed:



## 2.0 - Graphic Display

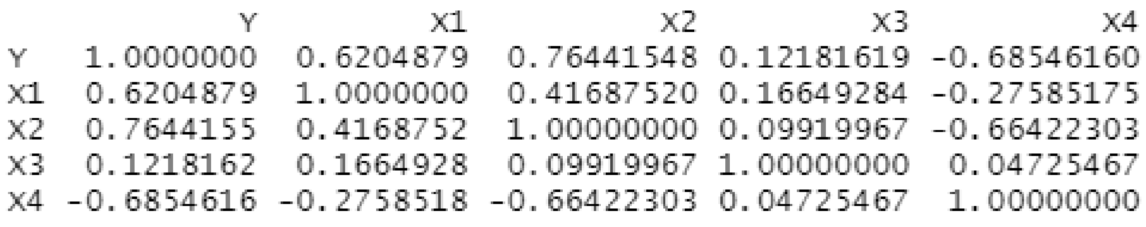
There are more than one predictors in the multiple linear regression. As such, it is not suitable to only have one scatter plot. We need a **scatter plot matrix** instead of a scatter plot. The plots are as followed:



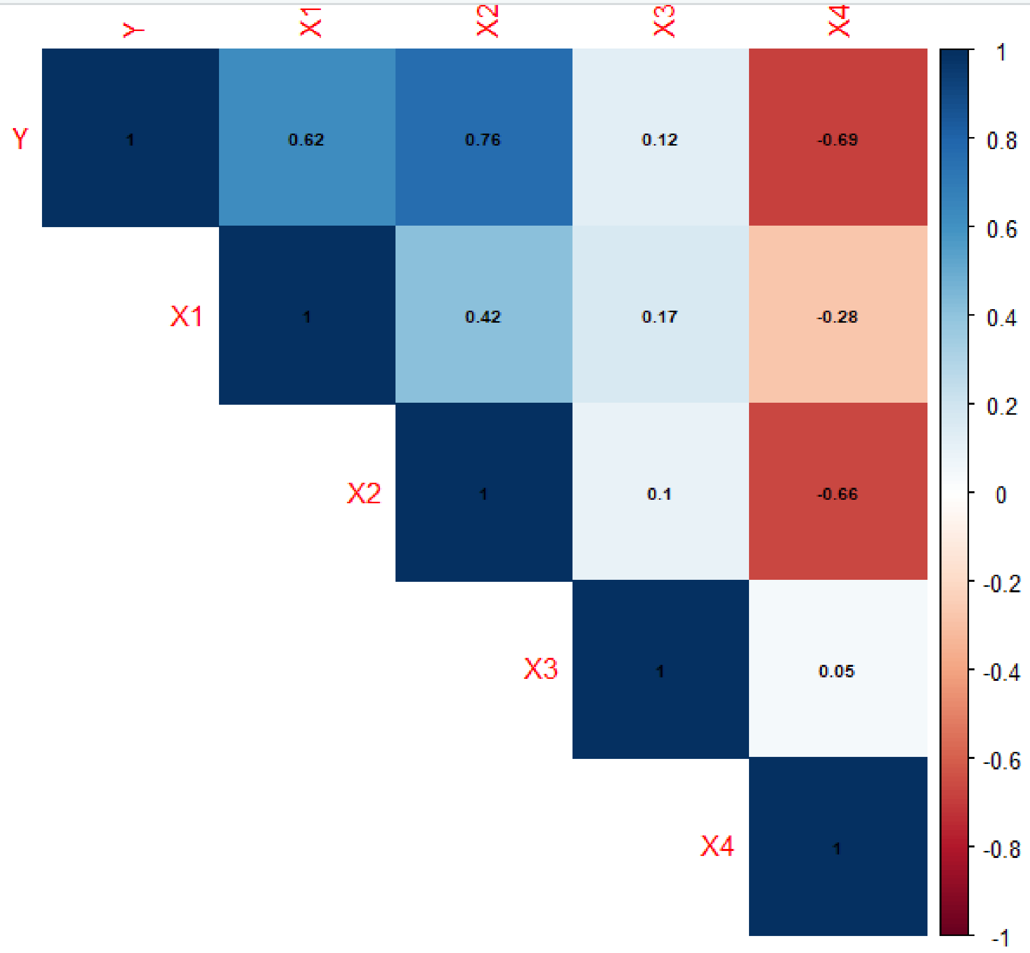
Conclusion: We observe a quadratic curve when plotting Y against X2. As such, we may add the **second order term I(X2$^2$) in the new model**

From the scatter matrix, we observe there is a relationship between the variables, some being linear and some being non-linear. We then study the correlation of the coefficients.

Output:



```
            Y         X1          X2          X3          X4
Y   1.0000000  0.6204879  0.76441548  0.12181619 -0.68546160
X1  0.6204879  1.0000000  0.41687520  0.16649284 -0.27585175
X2  0.7644155  0.4168752  1.00000000  0.09919967 -0.66422303
X3  0.1218162  0.1664928  0.09919967  1.00000000  0.04725467
X4 -0.6854616 -0.2758518 -0.66422303  0.04725467  1.00000000
```

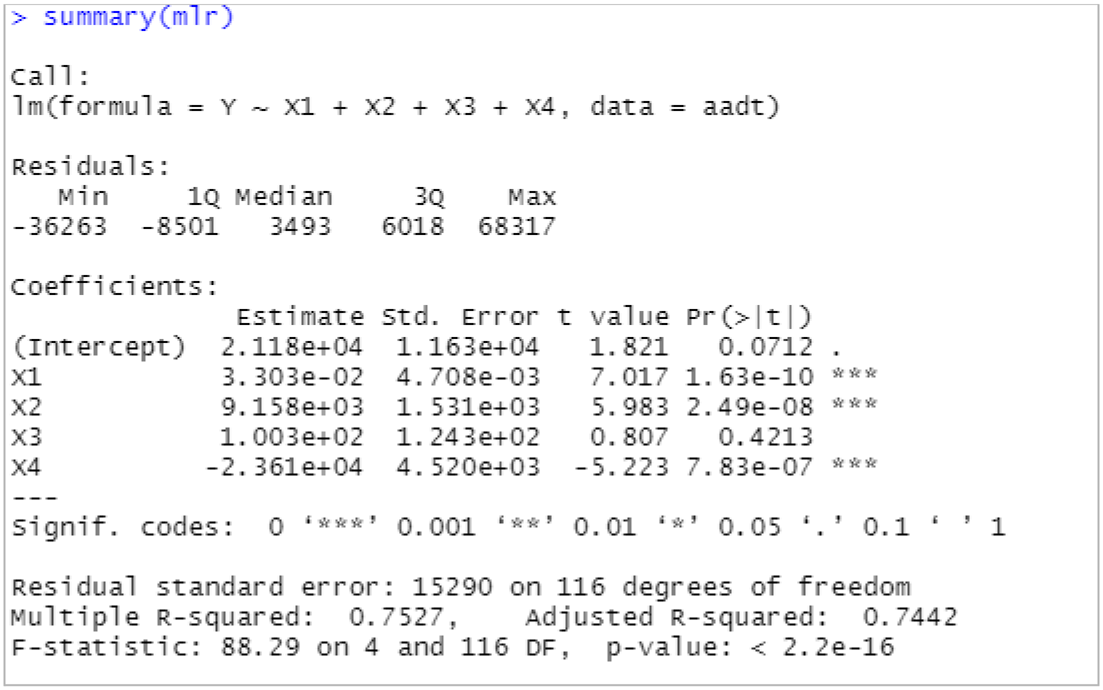


From the corrplot, which shows the correlation between any 2 distinct variables in the model, we observe none of them have absolute value over 0.8. As such, we conclude that **multicollinearity** is not likely to be an issue

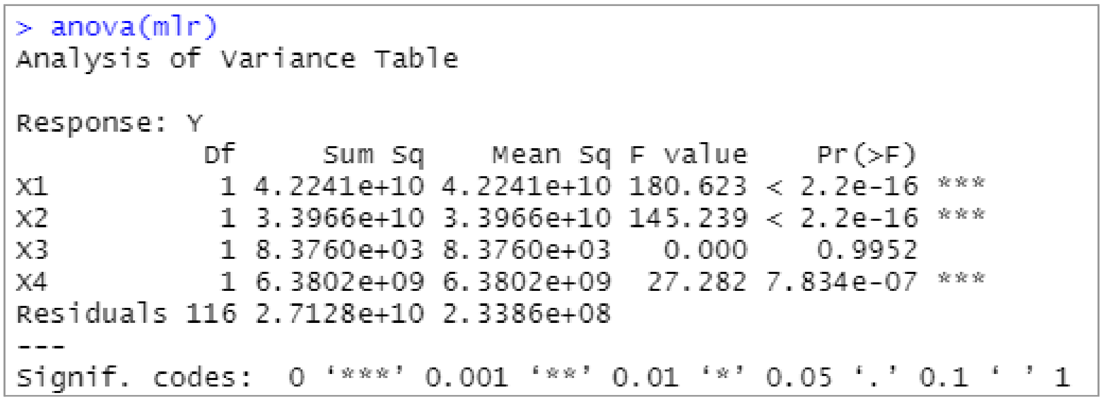
## 3.0 - Standard multiple regression model only using the predictor

The multiple linear regression equation is as follows:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4$$



```
> summary(mlr)

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
```



```
> anova(mlr)
Analysis of Variance Table

Response: Y
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
X1          1 4.2241e+10 4.2241e+10 180.623 < 2.2e-16 ***
X2          1 3.3966e+10 3.3966e+10 145.239 < 2.2e-16 ***
X3          1 8.3760e+03 8.3760e+03   0.000    0.9952    
X4          1 6.3802e+09 6.3802e+09  27.282 7.834e-07 ***
Residuals 116 2.7128e+10 2.3386e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## 4.0 Adequacy Checking

### 4.1 From the viewpoint of the fitted model

There are three types of tests used to check the adequacy of the fitted model from the viewpoint of the model. We have:

1. T-tests: they are used to check the significance of the fitted parameters.

    - $H_0: \beta_0 = 0$ t-value: 1.821 Pr(>|t|): 0.0712
    - $H_0: \beta_1 = 0$ t-value: 7.017 Pr(>|t|): 1.63e-10
    - $H_0: \beta_2 = 0$ t-value: 5.983 Pr(>|t|): 2.49e-08
    - $H_0: \beta_3 = 0$ t-value: 0.807 Pr(>|t|): 0.4213
    - $H_0: \beta_4 = 0$ t-value: -5.223 Pr(>|t|): 7.83e-07

As the standard of norm, we will reject $H_0$ if Pr(>|t|) < 0.05, at the 95% significance level. Thus we reject $H_0$ for predictors X1, X2 and X4. X1, X2 and X4 significantly improves our model fit. But we do not reject $H_0$ for predictors X3 and so we consider removing X3 from our model.

2. F-ratio: is the regression model significant?

    - F-statistic: 88.29 on 4 and 116 DF p-value:2.2e-16

Since the p-value is extremely small, the MLR model is said to be highly statistically significant.

3. R Statistic: does there exist a strong linear relationship between Y and X1, X2, X3, X4?

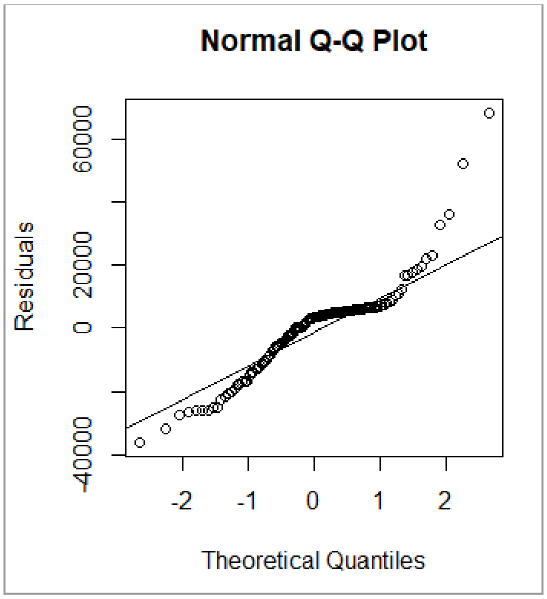
    - Multiple R-squared: 0.7527
    - Adjusted R-squared: 0.7442

According to the lecture notes, generally speaking, a good model should have an adjusted R-squared value not too small (< 60%) or not too large (> 95%). In our case the computed adjusted R-squared is 0.7442, between the two benchmark values aforementioned. Hence, it implies there is a relatively strong linear relationship between Y and the four X variables.

### 4.2 From the viewpoint of residuals

#### 4.2.1 Normality Checking

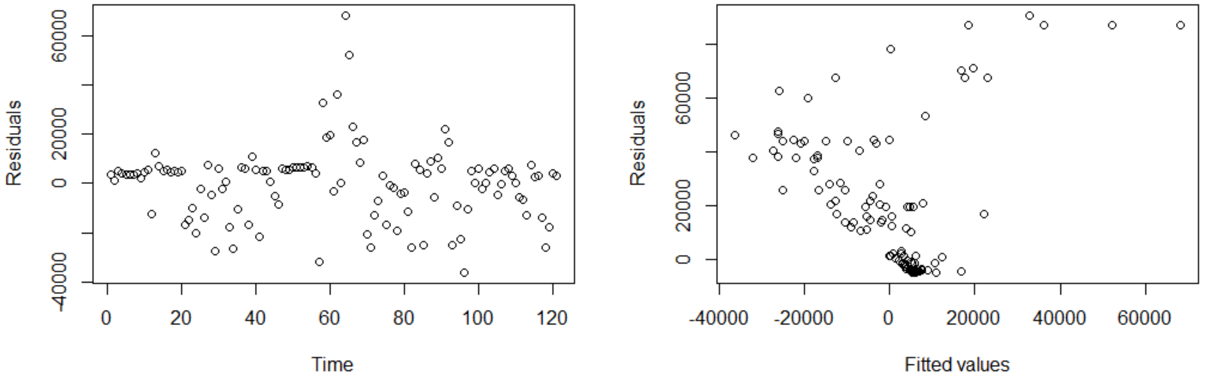
We will also be checking the residuals to obtain useful information for how we should adjust the model.

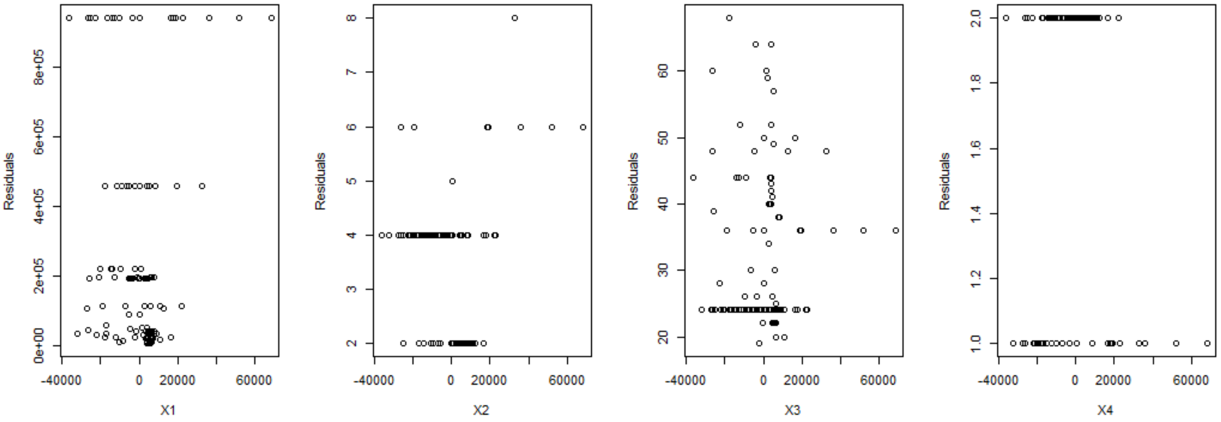


As observed, the residuals are not quite normally distributed since it has a skewed right tail. This corroborates with our earlier observations that the current MLR model does not fully explain the behaviour of the system since we are estimating our model under maximum likelihood estimator which assumes errors to be normally distributed.

#### 4.2.2 Checking for time effects, non-constant variance and higher order curvatures

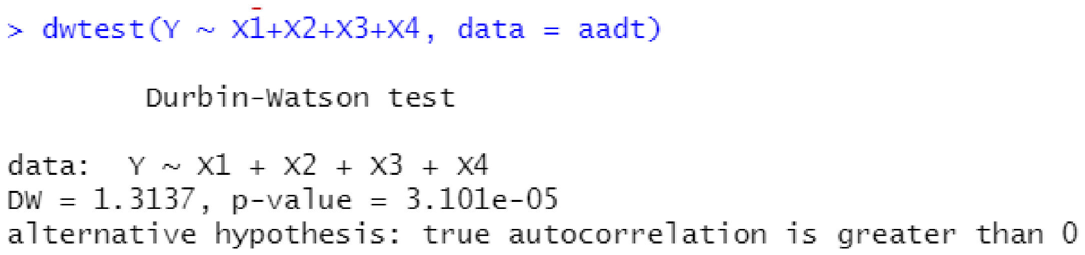
Next we will be checking for time effects if the time order of the data is known, for non-constant variance to see whether we need to take transformation on response, and for curvature of higher order than fitted in the predictors.





Residuals against X2 and X4 seem to contain horizontal bands of points and are relatively unhelpful. However, the bottom left section of the residual plot of X1 might show there is some linear relationship between X1 and the other predictors. The plots of residual against time, fitted and predictor values indicate that the variance increases as the fitted values increase (indicating non-constant variance). The residuals do not appear to have a relationship with time (are independent).

## **4.2.3 Checking for sequential dependence**



```
> dwtest(Y ~ X1+X2+X3+X4, data = aadt)

        Durbin-Watson test

data:  Y ~ X1 + X2 + X3 + X4
DW = 1.3137, p-value = 3.101e-05
alternative hypothesis: true autocorrelation is greater than 0
```

We use the Durbin-Watson test to check the possible sequential dependence, and after we check we have a D value of 1.3137, which is close to 0. Thus, we can conclude that the successive residuals is positively serially correlated

## 5.0 - F-test for the reduced model and full model

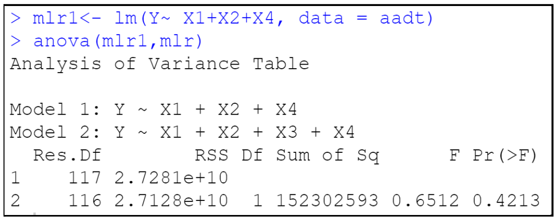
### 5.1 Test for whether some coefficients are zeros

To check the validity of X3 removal from the aforementioned conclusion, i.e: testing the null hypothesis where $H_0 : \beta 3 = 0$ , an F test is performed to compare

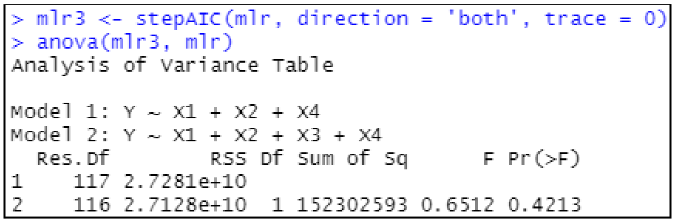
$Y0 = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4$, versus

$Y1 = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 4X4$.

The result of the F-test is shown in the left image below. To corroborate our findings, we also make use of a package to perform stepwise regression, to help us remove predictors that are not significant The result of the F-test is shown in the right image below.



```
> mlr1<- lm(Y~ X1+X2+X4, data = aadt)
> anova(mlr1,mlr)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    117 2.7281e+10
2    116 2.7128e+10  1 152302593 0.6512 0.4213
```



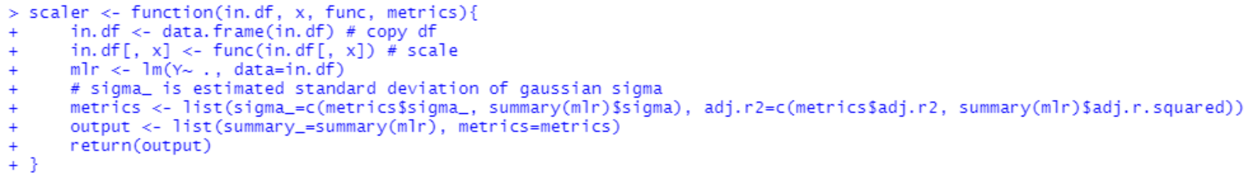
```
> mlr3 <- stepAIC(mlr, direction = 'both', trace = 0)
> anova(mlr3, mlr)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    117 2.7281e+10
2    116 2.7128e+10  1 152302593 0.6512 0.4213
```

**Fig: F-test for fitted model after removing X3 from observations (left) and F-test for fitted model after removing X3 using stepwise regression (right)**

Where $H_0: \beta_3 = 0$    F-value: 0.6512    Pr(>F): 0.4213

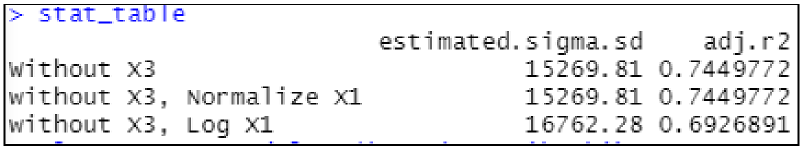
If Pr(>F) < 0.01: we reject $H_0$. In conclusion, from both stepwise regression and our observation of the p-values of the T-tests, we will not reject the null hypothesis, $H_0: \beta_3 = 0$. Thus we accept $H_0$ and **we will be removing X3 in the new model.**

Subsequently, we will conduct the following experiments with scaling of the data to see if the new model can fit the data better. To do that, we wrote a scaler function which will help us scale the data.



```
> scaler <- function(in.df, x, func, metrics){
+     in.df <- data.frame(in.df) # copy df
+     in.df[, x] <- func(in.df[, x]) # scale
+     mlr <- lm(Y~ ., data=in.df)
+     # sigma_ is estimated standard deviation of gaussian sigma
+     metrics <- list(sigma_=c(metrics$sigma_, summary(mlr)$sigma), adj.r2=c(metrics$adj.r2, summary(mlr)$adj.r.squared))
+     output <- list(summary_=summary(mlr), metrics=metrics)
+     return(output)
+ }
```

We conducted 2 additional scaling of the data, normalization and log transformation. The table below summarizes the standard error and the adjusted R$^2$ for the model when fitted with the reduced predictors.



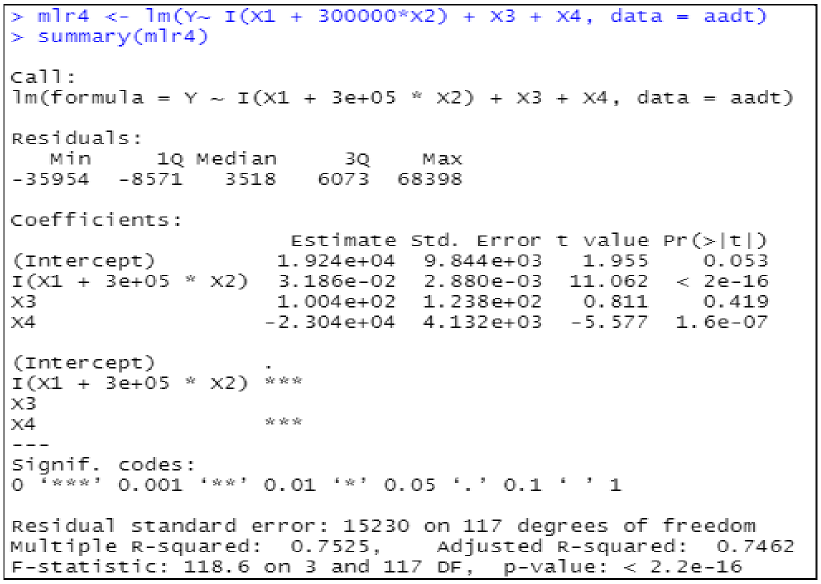
```
> stat_table
                          estimated.sigma.sd    adj.r2
Without X3                          15269.81 0.7449772
without X3, Normalize X1            15269.81 0.7449772
without X3, Log X1                  16762.28 0.6926891
```

From the results shown below, we observe that the normalization did not provide the model with any improvements and the log transformation caused the performance of the model to worsen.

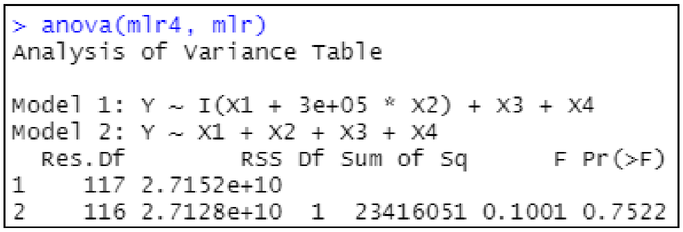
**5.2 Test for more complicated relationship**

From the fitted regression line, we see that the estimated coefficients of X1, X2, X3 and X4 are 3.303x10$^{-2}$, 9.158x10$^3$, 1.003x10$^2$ and -2.361x10$^4$ respectively. Therefore, we can conduct a test.

Null hypothesis, H$_0$: 300000$\beta_1 = \beta_2$



```
> mlr4 <- lm(Y~ I(X1 + 300000*X2) + X3 + X4, data = aadt)
> summary(mlr4)

Call:
lm(formula = Y ~ I(X1 + 3e+05 * X2) + X3 + X4, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-35954  -8571   3518   6073  68398 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.924e+04  9.844e+03   1.955    0.053
I(X1 + 3e+05 * X2)  3.186e-02  2.880e-03  11.062  < 2e-16
X3                  1.004e+02  1.238e+02   0.811    0.419
X4                 -2.304e+04  4.132e+03  -5.577  1.6e-07
                      
(Intercept)        .  
I(X1 + 3e+05 * X2) ***
X3                    
X4                 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15230 on 117 degrees of freedom
Multiple R-squared:  0.7525,	Adjusted R-squared:  0.7462 
F-statistic: 118.6 on 3 and 117 DF,  p-value: < 2.2e-16
```



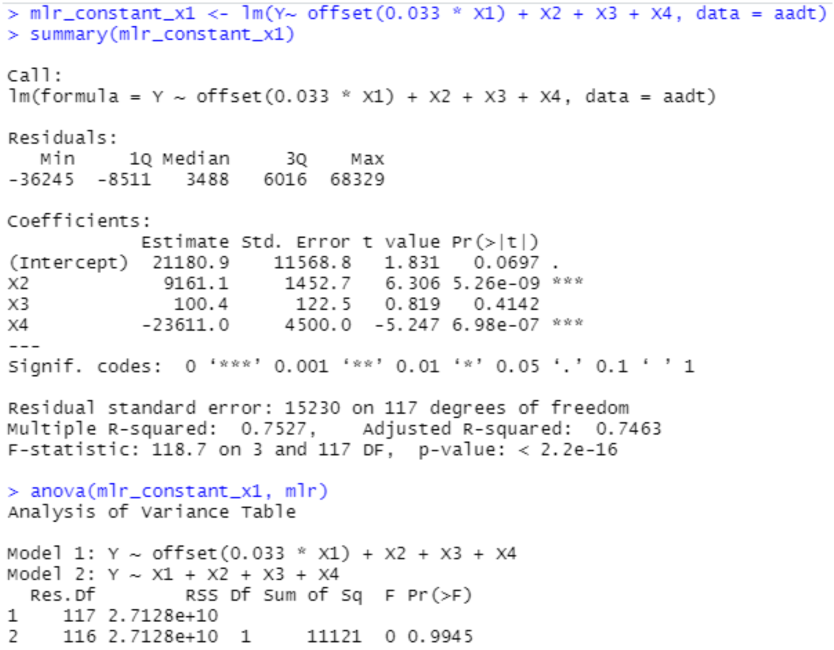
```
> anova(mlr4, mlr)
Analysis of Variance Table

Model 1: Y ~ I(X1 + 3e+05 * X2) + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    117 2.7152e+10                           
2    116 2.7128e+10  1  23416051 0.1001 0.7522
```

From the F-test above, we see that the Pr(>F) = 0.7522. Thus, we cannot reject the null hypothesis, and we can observe that $\beta_1$ is a factor of $\beta_2$.

### 5.3 Test for whether coefficients are constant

Using the reduced model, we now test to see if each of the coefficients are constant. From the fitted regression line, we shall conduct 3 tests.

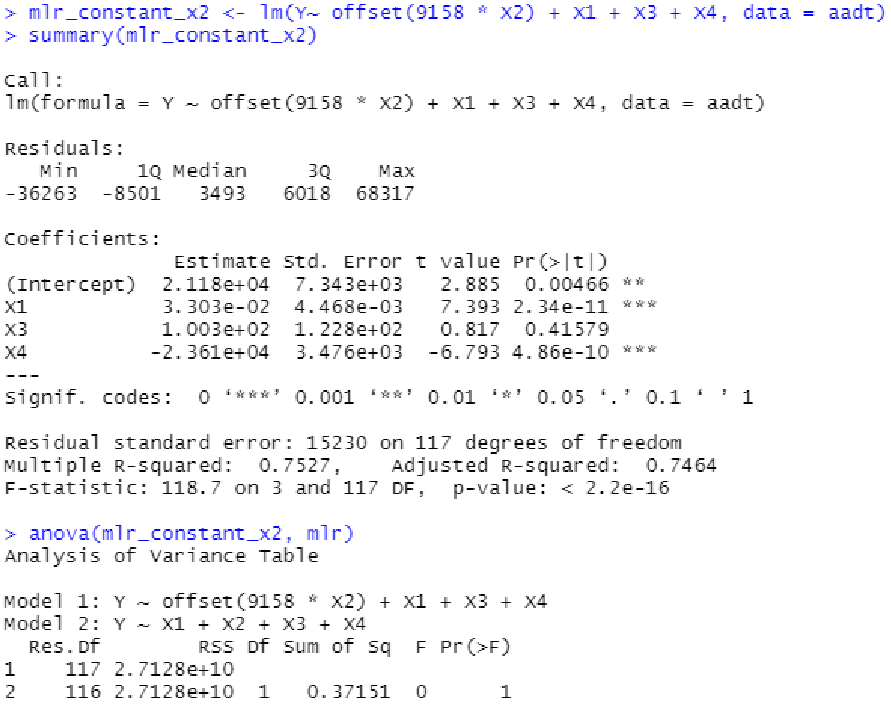
Null hypothesis for test 1, $H_0$: $\beta_1 = 0.033$



```
> mlr_constant_x1 <- lm(Y~ offset(0.033 * X1) + X2 + X3 + X4, data = aadt)
> summary(mlr_constant_x1)

Call:
lm(formula = Y ~ offset(0.033 * X1) + X2 + X3 + X4, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-36245  -8511   3488   6016  68329 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21180.9    11568.8   1.831   0.0697 .  
X2            9161.1     1452.7   6.306 5.26e-09 ***
X3             100.4      122.5   0.819   0.4142    
X4          -23611.0     4500.0  -5.247 6.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15230 on 117 degrees of freedom
Multiple R-squared:  0.7527,	Adjusted R-squared:  0.7463 
F-statistic: 118.7 on 3 and 117 DF,  p-value: < 2.2e-16

> anova(mlr_constant_x1, mlr)
Analysis of Variance Table

Model 1: Y ~ offset(0.033 * X1) + X2 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq  F Pr(>F)
1    117 2.7128e+10                       
2    116 2.7128e+10  1     11121  0 0.9945
```

For the first test, the Pr(>F) = 0.9945. We cannot reject the null hypothesis at a significance level of 0.1.

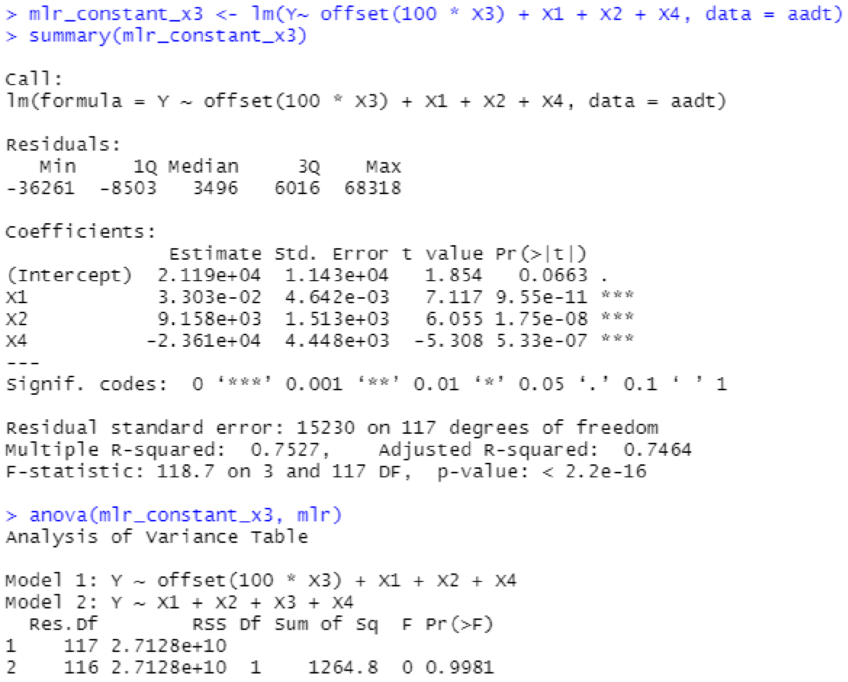
Null hypothesis for test 2, $H_0$: $\beta_2 = 9158$



```
> mlr_constant_x2 <- lm(Y~ offset(9158 * X2) + X1 + X3 + X4, data = aadt)
> summary(mlr_constant_x2)

Call:
lm(formula = Y ~ offset(9158 * X2) + X1 + 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  7.343e+03   2.885  0.00466 ** 
X1           3.303e-02  4.468e-03   7.393 2.34e-11 ***
X3           1.003e+02  1.228e+02   0.817  0.41579    
X4          -2.361e+04  3.476e+03  -6.793 4.86e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15230 on 117 degrees of freedom
Multiple R-squared:  0.7527,	Adjusted R-squared:  0.7464 
F-statistic: 118.7 on 3 and 117 DF,  p-value: < 2.2e-16

> anova(mlr_constant_x2, mlr)
Analysis of Variance Table

Model 1: Y ~ offset(9158 * X2) + X1 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq F Pr(>F)
1    117 2.7128e+10                      
2    116 2.7128e+10  1   0.37151 0      1
```

For the second test, we should not reject the null hypothesis at a significance level of 0.1.

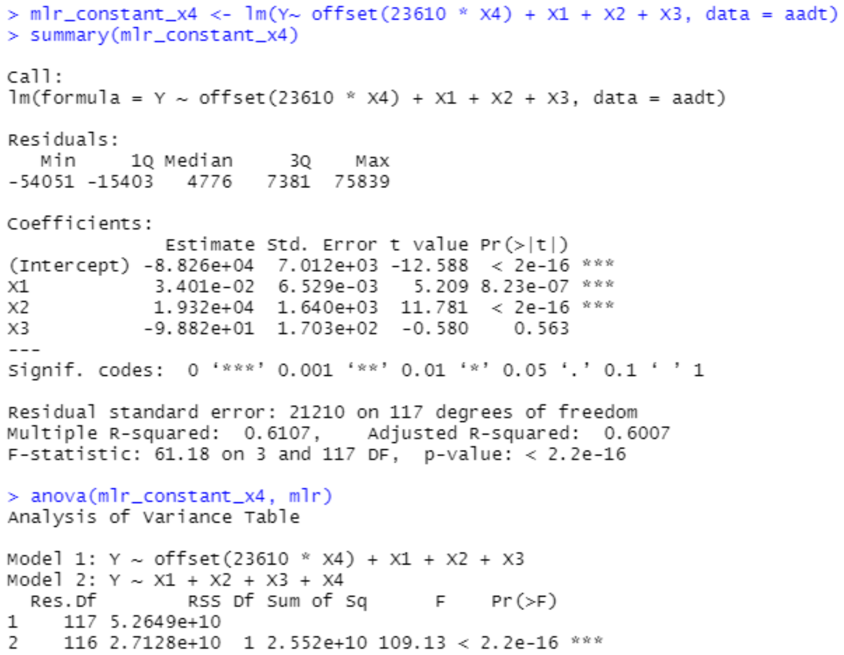
Null hypothesis for test 3, $H_0$: $\beta_3$ = 100



```
> mlr_constant_x3 <- lm(Y~ offset(100 * X3) + X1 + X2 + X4, data = aadt)
> summary(mlr_constant_x3)

Call:
lm(formula = Y ~ offset(100 * X3) + X1 + X2 + X4, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-36261  -8503   3496   6016  68318 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.119e+04  1.143e+04   1.854   0.0663 .  
X1           3.303e-02  4.642e-03   7.117 9.55e-11 ***
X2           9.158e+03  1.513e+03   6.055 1.75e-08 ***
X4          -2.361e+04  4.448e+03  -5.308 5.33e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15230 on 117 degrees of freedom
Multiple R-squared:  0.7527,	Adjusted R-squared:  0.7464 
F-statistic: 118.7 on 3 and 117 DF,  p-value: < 2.2e-16

> anova(mlr_constant_x3, mlr)
Analysis of Variance Table

Model 1: Y ~ offset(100 * X3) + X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq  F Pr(>F)
1    117 2.7128e+10                       
2    116 2.7128e+10  1    1264.8  0 0.9981
```

For the second test, we should not reject the null hypothesis at a significance level of 0.1.

Null hypothesis for test 4, $H_{0:}$ $\beta_4$ = 23610



```
> mlr_constant_x4 <- lm(Y~ offset(23610 * X4) + X1 + X2 + X3, data = aadt)
> summary(mlr_constant_x4)

Call:
lm(formula = Y ~ offset(23610 * X4) + X1 + X2 + X3, data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-54051 -15403   4776   7381  75839 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.826e+04  7.012e+03 -12.588  < 2e-16 ***
X1           3.401e-02  6.529e-03   5.209 8.23e-07 ***
X2           1.932e+04  1.640e+03  11.781  < 2e-16 ***
X3          -9.882e+01  1.703e+02  -0.580    0.563    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21210 on 117 degrees of freedom
Multiple R-squared:  0.6107,	Adjusted R-squared:  0.6007 
F-statistic: 61.18 on 3 and 117 DF,  p-value: < 2.2e-16

> anova(mlr_constant_x4, mlr)
Analysis of Variance Table

Model 1: Y ~ offset(23610 * X4) + X1 + X2 + X3
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)    
1    117 5.2649e+10                                  
2    116 2.7128e+10  1 2.552e+10 109.13 < 2.2e-16 ***
```

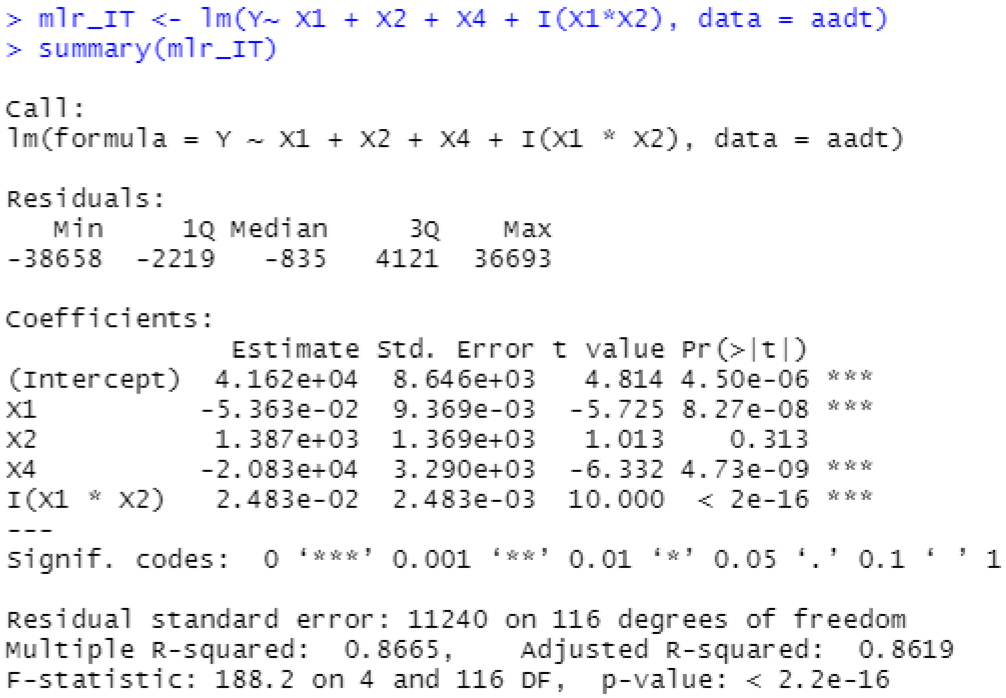
For the third test, Pr(>F) < 0.1. We should reject the null hypothesis at a significance level of 0.1 in favour of the alternative hypothesis.

In conclusion, from the three tests where we see if the coefficients could be constants, we do not reject the null hypothesis for $H_0$: $\beta_1$ = 0.033, $H_0$: $\beta_2$ = 9158 and $H_0$: $\beta_3$ = 100. This shows that the coefficients for X1, X2 and X3 are most likely constants. On the other hand, we reject the null hypothesis $H_{0:}$ $\beta_4$ = 23610 in favour of the alternative hypothesis, where the coefficient of X4 is not a constant.

**5.4 Test for interaction terms**

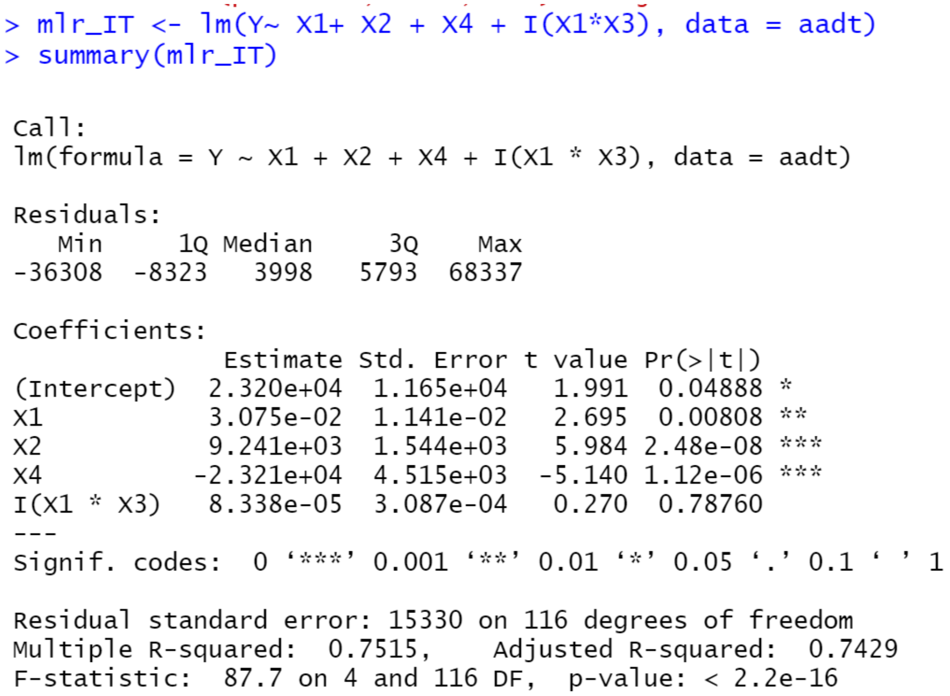
From the previous tests, we have observed that X3 is not significant to the performance of the model. Moreover, from the regression analysis in 4.2.2, we deduced that there may be linear relationship between X1 and the other predictors, therefore, in this section, we experiment with some interaction terms involving X1 and the other predictors to see if we can use the underlying information in creating a more significant predictor for our model.

<u>Test 1: Multiplying X1 and X2 to form X1X2</u>



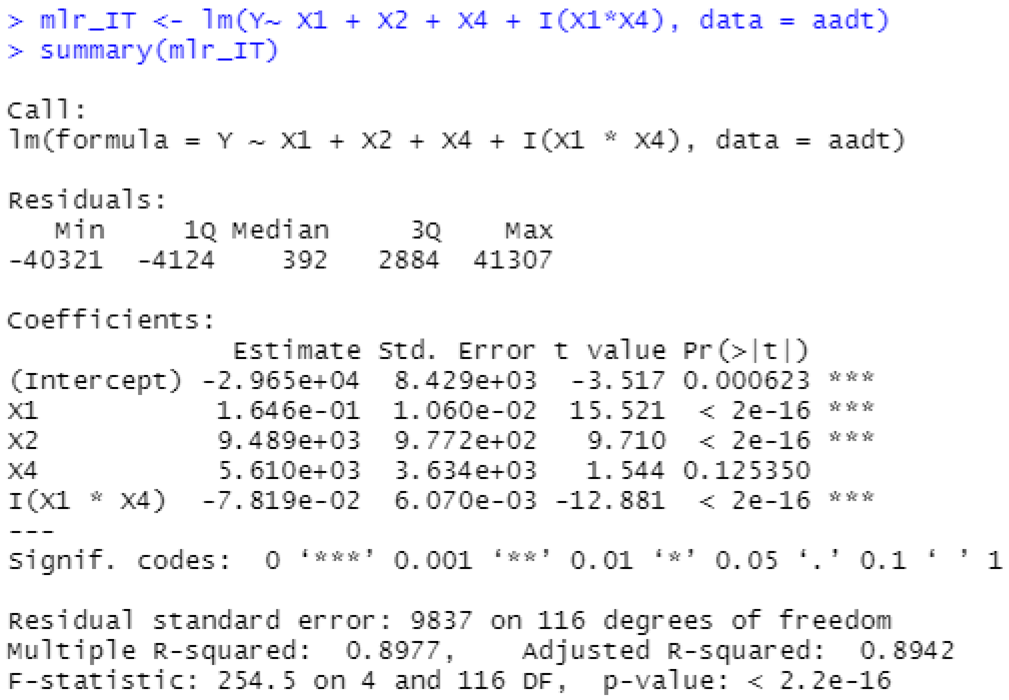
```
> mlr_IT <- lm(Y~ X1 + X2 + X4 + I(X1*X2), data = aadt)
> summary(mlr_IT)

Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X1 * X2), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-38658  -2219   -835   4121  36693 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.162e+04  8.646e+03   4.814 4.50e-06 ***
X1          -5.363e-02  9.369e-03  -5.725 8.27e-08 ***
X2           1.387e+03  1.369e+03   1.013    0.313    
X4          -2.083e+04  3.290e+03  -6.332 4.73e-09 ***
I(X1 * X2)   2.483e-02  2.483e-03  10.000  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11240 on 116 degrees of freedom
Multiple R-squared:  0.8665,    Adjusted R-squared:  0.8619 
F-statistic: 188.2 on 4 and 116 DF,  p-value: < 2.2e-16
```

## Test 2: Multiplying X1 and X3 to form X1X3



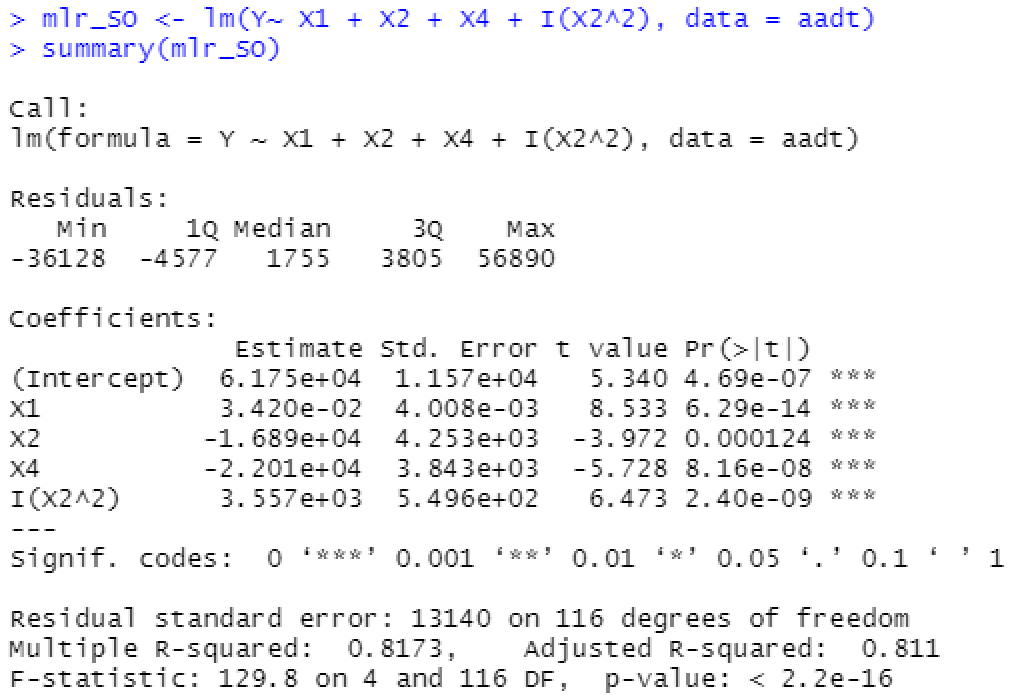
```
> mlr_IT <- lm(Y~ X1+ X2 + X4 + I(X1*X3), data = aadt)
> summary(mlr_IT)

Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X1 * X3), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-36308  -8323   3998   5793  68337 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.320e+04  1.165e+04   1.991  0.04888 *  
X1           3.075e-02  1.141e-02   2.695  0.00808 ** 
X2           9.241e+03  1.544e+03   5.984 2.48e-08 ***
X4          -2.321e+04  4.515e+03  -5.140 1.12e-06 ***
I(X1 * X3)   8.338e-05  3.087e-04   0.270  0.78760    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15330 on 116 degrees of freedom
Multiple R-squared:  0.7515,    Adjusted R-squared:  0.7429 
F-statistic:  87.7 on 4 and 116 DF,  p-value: < 2.2e-16
```

## Test 3: Multiplying X1 and X4 to form X1X4



```
> mlr_IT <- lm(Y~ X1 + X2 + X4 + I(X1*X4), data = aadt)
> summary(mlr_IT)

Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X1 * X4), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-40321  -4124    392   2884  41307 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.965e+04  8.429e+03  -3.517 0.000623 ***
X1           1.646e-01  1.060e-02  15.521  < 2e-16 ***
X2           9.489e+03  9.772e+02   9.710  < 2e-16 ***
X4           5.610e+03  3.634e+03   1.544 0.125350    
I(X1 * X4)  -7.819e-02  6.070e-03 -12.881  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9837 on 116 degrees of freedom
Multiple R-squared:  0.8977,    Adjusted R-squared:  0.8942 
F-statistic: 254.5 on 4 and 116 DF,  p-value: < 2.2e-16
```

Test 4: Adding a second order term I(X2$^2$)



```
> mlr_so <- lm(Y~ X1 + X2 + X4 + I(X2^2), data = aadt)
> summary(mlr_so)

Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X2^2), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-36128  -4577   1755   3805  56890 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.175e+04  1.157e+04   5.340 4.69e-07 ***
X1           3.420e-02  4.008e-03   8.533 6.29e-14 ***
X2          -1.689e+04  4.253e+03  -3.972 0.000124 ***
X4          -2.201e+04  3.843e+03  -5.728 8.16e-08 ***
I(X2^2)      3.557e+03  5.496e+02   6.473 2.40e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13140 on 116 degrees of freedom
Multiple R-squared:  0.8173,	Adjusted R-squared:  0.811 
F-statistic: 129.8 on 4 and 116 DF,  p-value: < 2.2e-16
```

In conclusion, from the 4 tests above, we can see that the interaction terms X1X2 and X1X4, as well as the addition of the second order term I(X2$^2$) are significant to the model. However, the interaction term X1X3 is not significant. Furthermore, with the addition of the interaction terms, the adjusted R$^2$ value of the models increased significantly. This showed that the addition of the interaction terms are not only significant, but also helped the model fit better to the data.

**5.5 Proposed model**

In this section, we are using a **significance level of 0.01**

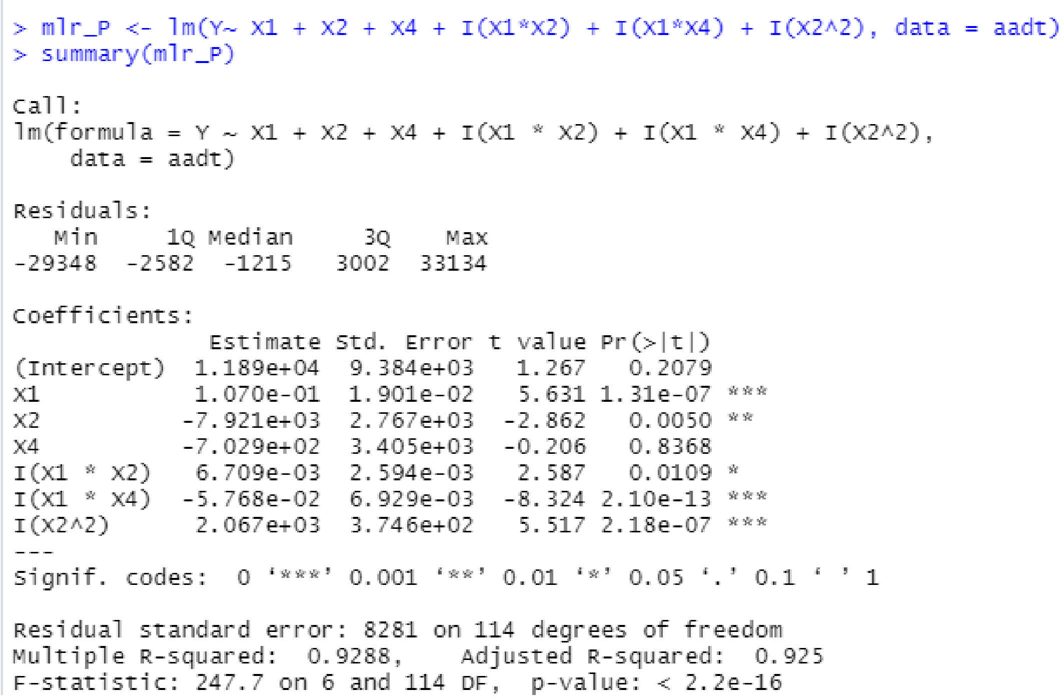
According to these 3 identified improvements,

1. A second order term I(X2$^2$) was added to the new model as there is a quadratic relationship between Y and X2. (from 2.0)

2. Second Order terms I(X1*X2), I(X1*X4) [The residual plot of X1 might show there is some linear relationship between X1 and the other predictors]. We decide to not include I(X1*X3) since it is not significant (from 5.4)
3. X3 was removed (from 5.0)

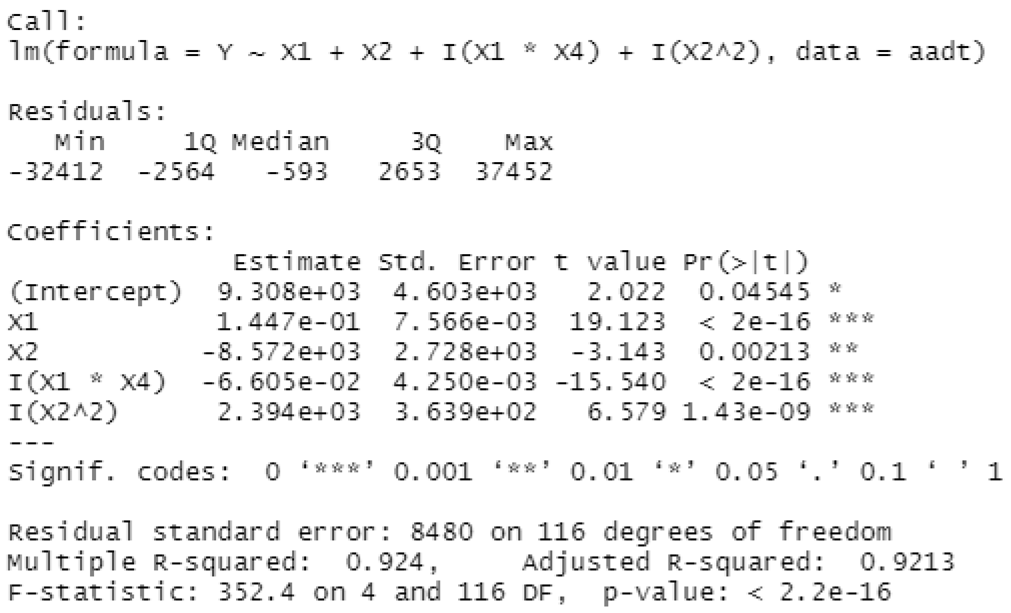
We propose the following model

$$Y = \beta 0 + \beta 1X1 + \beta 2X1X2 + \beta 3X1X4 + \beta 4X2 + \beta 5X4 + \beta 6X2^{2}$$



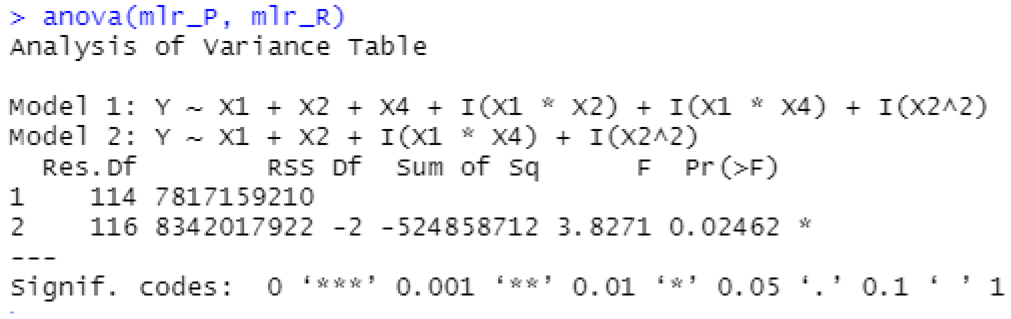
```
> mlr_P <- lm(Y~ X1 + X2 + X4 + I(X1*X2) + I(X1*X4) + I(X2^2), data = aadt)
> summary(mlr_P)

Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X1 * X2) + I(X1 * X4) + I(X2^2),
    data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-29348  -2582  -1215   3002  33134 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.189e+04  9.384e+03   1.267   0.2079    
X1           1.070e-01  1.901e-02   5.631 1.31e-07 ***
X2          -7.921e+03  2.767e+03  -2.862   0.0050 ** 
X4          -7.029e+02  3.405e+03  -0.206   0.8368    
I(X1 * X2)   6.709e-03  2.594e-03   2.587   0.0109 *  
I(X1 * X4)  -5.768e-02  6.929e-03  -8.324 2.10e-13 ***
I(X2^2)      2.067e+03  3.746e+02   5.517 2.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8281 on 114 degrees of freedom
Multiple R-squared:  0.9288,	Adjusted R-squared:  0.925 
F-statistic: 247.7 on 6 and 114 DF,  p-value: < 2.2e-16
```

From the output of summary(mlr_P), we find that the p-values of X4 and X1*X2 is greater than 0.01. As such we may consider to remove it from the proposed model (mlr_P)



```
Call:
lm(formula = Y ~ X1 + X2 + I(X1 * X4) + I(X2^2), data = aadt)

Residuals:
   Min     1Q Median     3Q    Max 
-32412  -2564   -593   2653  37452 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.308e+03  4.603e+03   2.022  0.04545 *  
X1           1.447e-01  7.566e-03  19.123  < 2e-16 ***
X2          -8.572e+03  2.728e+03  -3.143  0.00213 ** 
I(X1 * X4)  -6.605e-02  4.250e-03 -15.540  < 2e-16 ***
I(X2^2)      2.394e+03  3.639e+02   6.579 1.43e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8480 on 116 degrees of freedom
Multiple R-squared:  0.924,	Adjusted R-squared:  0.9213 
F-statistic: 352.4 on 4 and 116 DF,  p-value: < 2.2e-16
```

Now, we examine whether the removal of the 2 predictors is significant
We test H0: mlr_R(reduced model) vs H1: mlr_P (full model)



```
> anova(mlr_P, mlr_R)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X4 + I(X1 * X2) + I(X1 * X4) + I(X2^2)
Model 2: Y ~ X1 + X2 + I(X1 * X4) + I(X2^2)
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)  
1    114 7817159210                               
2    116 8342017922 -2 -524858712 3.8271 0.02462 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of the above test is 0.02462 which is greater than 0.01. As such, it is not small enough to justify the inclusion of the predictors X1*X2 and X4 in the model

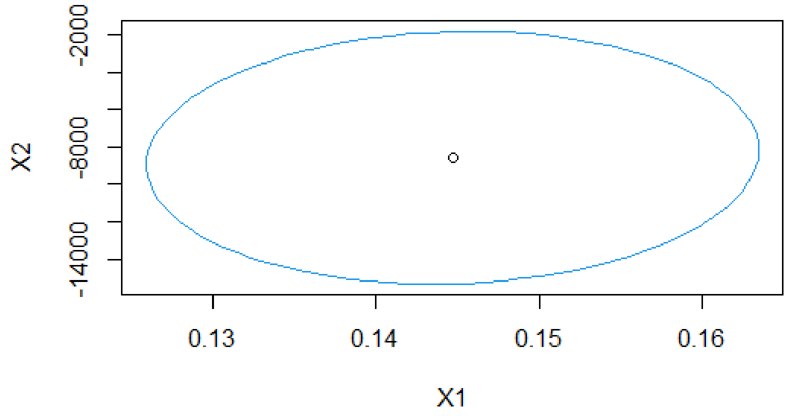
Thus, our **final proposed model is**

$$Y = \beta 0 + \beta 1 X1 + \beta 2 X1X4 + \beta 3 X2 + \beta 4 X2^{2}$$

# 6.0 - Making Inference

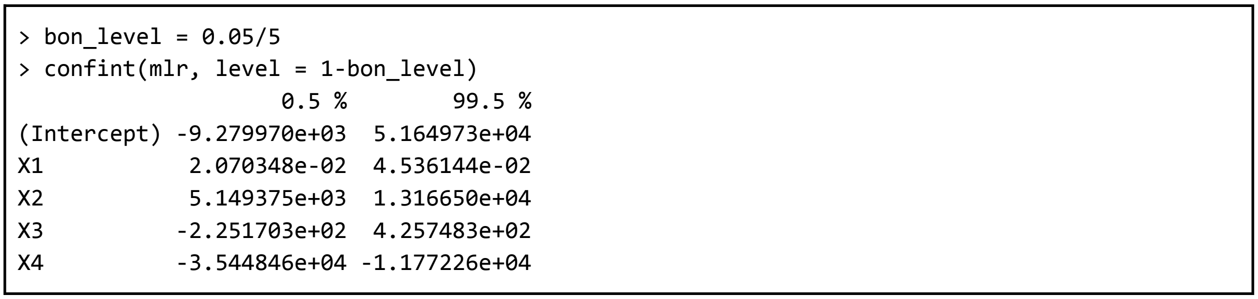
## 6.1 Confidence interval

We plot the 95% confidence interval for X1 and X2 and it resulted in the below graph.



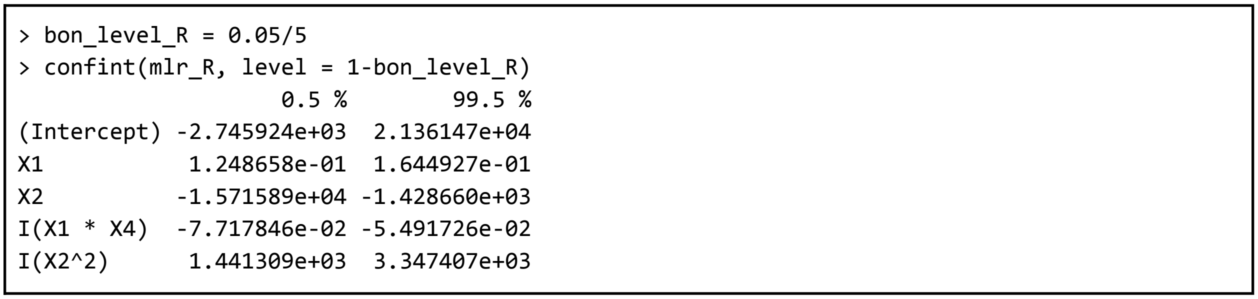
We then use the Bonferroni limit to get the confidence interval for each coefficient.

Confidence interval for each model in the full model:



```
> bon_level = 0.05/5
> confint(mlr, level = 1-bon_level)
                    0.5 %        99.5 %
(Intercept) -9.279970e+03  5.164973e+04
X1           2.070348e-02  4.536144e-02
X2           5.149375e+03  1.316650e+04
X3          -2.251703e+02  4.257483e+02
X4          -3.544846e+04 -1.177226e+04
```

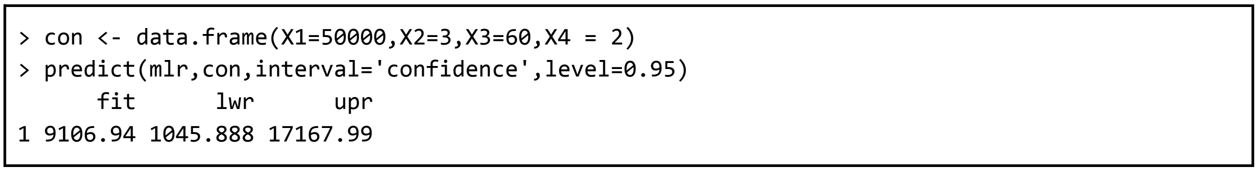
Confidence interval for each model in the newly formed model:



```
> bon_level_R = 0.05/5
> confint(mlr_R, level = 1-bon_level_R)
                    0.5 %        99.5 %
(Intercept) -2.745924e+03  2.136147e+04
X1           1.248658e-01  1.644927e-01
X2          -1.571589e+04 -1.428660e+03
I(X1 * X4)  -7.717846e-02 -5.491726e-02
I(X2^2)      1.441309e+03  3.347407e+03
```

## 6.2 Prediction

- Initial model ($Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4$)



```
> con <- data.frame(X1=50000,X2=3,X3=60,X4 = 2)
> predict(mlr,con,interval='confidence',level=0.95)
      fit      lwr      upr
1 9106.94 1045.888 17167.99
```

```
> predict(mlr,con,interval='prediction',level=0.95)
     fit       lwr      upr
1 9106.94 -22236.34 40450.22
```

- Confidence interval for mean response
  $l = \beta_0 + 50000\beta_1 + 3\beta_2 + 60\beta_3 + 2\beta_4$
  95% confidence interval is [1046, 17168]

- Confidence interval for new observation
  $l = \beta_0 + 50000\beta_1 + 3\beta_2 + 60\beta_3 + 2\beta_4 + \varepsilon$
  95% confidence interval is [-22236, 40450]

- New model ($Y= \beta_0 + \beta_1 X1 + \beta_2 X1X4 + \beta_3 X2 + \beta_4 X2^2$)

```
> con_R <- data.frame(X1=50000, X2=3, X4=2)
> predict(mlr_R,con_R,interval='confidence',level=0.95)
      fit      lwr      upr
1 5769.346 3609.093 7929.599
> predict(mlr_R,con_R,interval='prediction',level=0.95)
      fit       lwr      upr
1 5769.346 -11165.13 22703.82
```

- Confidence interval for mean response
  $l = \beta_0 + 50000\beta_1 + 100000\beta_2 + 3\beta_3 + 9\beta_4$
  95% confidence interval is [3609, 7930]

- Confidence interval for new observation
  $l = \beta_0 + 50000\beta_1 + 100000\beta_2 + 3\beta_3 + 9\beta_4 + \varepsilon$
  95% confidence interval is [-11165, 22704]

The lower bound of confidence interval for the new observation is negative. However, note that the value of $l$ must not be negative since the response variable here is the average annual daily traffic.

The range of predictions in the newly formed model is narrower. This suggests that the new model is better and more precise.

## 7.0 - Conclusion

Firstly, we want to know the relationship between the variables. So, we make a scatter plot matrix and we consider adding a second order term I(X$2^2$) since there is a quadratic relationship between Y and X2. There is no obvious linear relationship between other variables. Next, we create a correlation plot and conclude that there is not an issue of **multicollinearity** between the variables.

After that, we make standard multiple linear regression equation as follows:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4$$

Using this regression model, we find out the estimated coefficients of X1, X2, X3 and X4 are 3.303x$10^{-2}$, 9.158x$10^{3}$, 1.003x$10^{2}$ and -2.361x$10^{4}$ respectively. Now we conduct several tests to further improve the model.

From the viewpoint of the fitted model, we first conduct a t-test to test the significance of each predictor variable and observe that predictor variables X1, X2, and X4 are significant, while X3 is not significant and thus we consider removing X3. Next, by using F-test, we conclude that the MLR model is said to be highly statistically significant. Finally, we also see a relatively strong linear relationship between Y and the four X variables with an adjusted R-squared value of 0.74.

Next, from the residuals view, we see that the MLR model does not fully explain the behaviour of the system as residuals of the model do not follow a normal distribution. The residual plot of X1 might show there is some linear relationship between X1 and the other predictors. The plot of residuals against time suggests that they are independent of time. Using the Durbin-Watson test, we also conclude that the successive residuals are positively serially correlated.

We further conduct hypothesis testing to check the behaviour of model coefficients and the following observations were made:

1. The coefficients of X1, X2 and X4 were found not to be not zero.
2. The coefficient of X1 is a factor of the coefficient of X2
3. The coefficients for X1, X2 and X3 are most likely constants and the coefficient of X4 is not a constant.

Moreover, we also conduct experiments to see whether any scaling to the data will improve the model. However, from the experiments we observe that the normalization did not provide the model with any improvements and the log transformation caused the performance of the model to worsen. Thus, we do not use this scaling.

We also check the significance of adding interaction terms and observe that adding X1X2, X1X4 and X2^2 would increase the adjusted $R^2$ of the model significantly. This showed. This showed that the addition of the interaction terms are not only significant, but also helped the model fit better to the data.

Thus, our **final proposed model is**

$$Y = \beta0 + \beta1X1 + \beta2X1X4 + \beta3X2 + \beta4X2^2$$

Finally, we create the confidence interval for the proposed and full model. The 95% confidence interval for mean response is [1046,17168] in the full model and [3609,7930] in our proposed model. For the new observation, the 95% confidence interval is [-22236,40450] for the full model and [-11165,22704] for the proposed model. From this we see that the confidence interval range is narrowed for the proposed model as compared to the original model. This suggests that the proposed model is better and more precise.

## 8.0 - Complete R Code

```
# import the required libraries
library(MASS)
library(dplyr)
library(ggplot2)
library(lmtest)
library(ellipse)
library(corrplot)

# analyse dataset
aadt_main=read.table("data/aadt.txt", header = FALSE)
aadt <- data.frame(Y=aadt_main$V1, X1=aadt_main$V2, X2=aadt_main$V3,
X3=aadt_main$V4, X4=aadt_main$V5)
plot(aadt, panel = panel.smooth)
cor(aadt)
corrplot(cor(aadt), type="upper", method="color", addCoef.col="black",
number.cex=0.6)

# Model 0: standard MLR with only predictors
mlr <- lm(Y~ X1+X2+X3+X4, data = aadt)
summary(mlr)
anova(mlr)

# Model 0: Normality checking.
qqnorm(residuals(mlr),ylab='Residuals')
qqline(residuals(mlr))

# Model 0: Draw some plots of residuals.
par(mfrow=c(1,3))
plot(residuals(mlr),ylab='Residuals',xlab='Time')
plot(residuals(mlr),fitted(mlr),ylab='Residuals',xlab='Fitted values')
plot(residuals(mlr),aadt_main$V1,ylab='Residuals',xlab='Response variable')
par(mfrow=c(1,1))
par(mfrow=c(1,4))
plot(residuals(mlr),aadt_main$V2,ylab='Residuals',xlab='X1')
plot(residuals(mlr),aadt_main$V3,ylab='Residuals',xlab='X2')
plot(residuals(mlr),aadt_main$V4,ylab='Residuals',xlab='X3')
plot(residuals(mlr),aadt_main$V5,ylab='Residuals',xlab='X4')
par(mfrow=c(1,1))

#Durbin-Watson test
dwtest(Y ~ X1+X2+X3+X4, data = aadt)

# manually remove X3 since t-test p-value from summary(mlr) is high
aadt2 <- data.frame(aadt)
aadt2$X3 <- NULL
mlr2 <- lm(Y~ ., data=aadt2)
summary(mlr2)

# use stepAIC package to select features to remove instead (still removes just X3)
mlr3 <- stepAIC(mlr, direction = 'both')
summary(mlr3)

# func to scale one column
scaler <- function(in.df, x, func, metrics){
 in.df <- data.frame(in.df) # copy df
 in.df[, x] <- func(in.df[, x]) # scale
 mlr <- lm(Y~ ., data=in.df)
 # sigma_ is estimated standard deviation of gaussian sigma
 metrics <- list(sigma_=c(metrics$sigma_, summary(mlr)$sigma),
adj.r2=c(metrics$adj.r2, summary(mlr)$adj.r.squared))
 output <- list(summary_=summary(mlr), metrics=metrics)
 return(output)
}
metrics <- list(sigma_=summary(mlr2)$sigma, adj.r2=summary(mlr2)$adj.r.squared)

# scaling data without X3 in an attempt to get better model fitting
# make X1 normal distribution
result <- scaler(aadt2, 'X1', scale, metrics)
result$metrics # no diff in scaling X1 (when X3 is removed)
```

```
# repeat but use log instead
result <- scaler(aadt2, 'X1', log, result$metrics)
result$metrics # log(X1) reduces performance

stat_table <- data.frame(estimated.sigma.sd=result$metrics$sigma_,
adj.r2=result$metrics$adj.r2)
rownames(stat_table) <- c('Without X3', 'without X3, Normalize X1', 'without X3, Log X1'
)
stat_table

#Test for more complicated relationship
#Test for H0: 300000*beta1 = beta2
mlr4 <- lm(Y~ I(X1 + 300000*X2) + X3 + X4, data = aadt)
summary(mlr4)
anova(mlr4, mlr)

#Test whether coefficients are constants
mlr_constant_x1 <- lm(Y~ offset(0.033 * X1) + X2 + X3 + X4, data = aadt)
summary(mlr_constant_x1)
anova(mlr_constant_x1, mlr)

mlr_constant_x2 <- lm(Y~ offset(9158 * X2) + X1 + X3 + X4, data = aadt)
summary(mlr_constant_x2)
anova(mlr_constant_x2, mlr)

mlr_constant_x3 <- lm(Y~ offset(100 * X3) + X1 + X2 + X4, data = aadt)
summary(mlr_constant_x3)
anova(mlr_constant_x3, mlr)

mlr_constant_x4 <- lm(Y~ offset(23610 * X4) + X1 + X2 + X3, data = aadt)
summary(mlr_constant_x4)
anova(mlr_constant_x4, mlr)

#Test for interaction terms
```

```
# test 1
mlr_IT <- lm(Y~ X1 + X2 + X4 + I(X1*X2), data = aadt)
summary(mlr_IT)

# test 2
mlr_IT <- lm(Y~ X1 + X2 + X4 + I(X1*X3), data = aadt)
summary(mlr_IT)

# test 3
mlr_IT <- lm(Y~ X1 + X2 + X4 + I(X1*X4), data = aadt)
summary(mlr_IT)

# test 4
mlr_SO <- lm(Y~ X1 + X2 + X4 + I(X2^2), data = aadt)
summary(mlr_SO)

#final proposed model

mlr_P <- lm(Y~ X1 + X2 + X4 + I(X1*X2) + I(X1*X4) + I(X2^2), data = aadt)
summary(mlr_P)

mlr_R <- lm(Y~ X1 + X2 + I(X1*X4) + I(X2^2), data = aadt)
summary(mlr_R)

anova(mlr_P, mlr_R)

#Confidence Interval
plot(ellipse(mlr_R,c(2,3),level = 0.95),type = 'l', col = 4)
points(coef(mlr_R)[2],coef(mlr_R)[3])

#Bonferroni limit
bon_level = 0.05/5
confint(mlr, level = 1-bon_level)
bon_level_R = 0.05/5
```

```
confint(mlr_R, level = 1-bon_level_R)

# Prediction using the full model
con <- data.frame(X1=50000,X2=3,X3=60,X4 = 2)
predict(mlr,con,interval='confidence',level=0.95)
predict(mlr,con,interval='prediction',level=0.95)

# Prediction using new model
con_R <- data.frame(X1=50000, X2=3, X4=2)
predict(mlr_R,con_R,interval='confidence',level=0.95)
predict(mlr_R,con_R,interval='prediction',level=0.95)
```