NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Physical and Mathematical Sciences

MH3510 Regression Analysis (Group Project)

Group members

Name	Matric No.
Chen Gordon Tian Xiao	U2140820A
Wu Wanqi	U2140590G
Huang Jiahua	U2140680F
Min Khant Htoo	U2140545E
Alviento Adrian Nicolas Belleza	U2140615H

1. Graphic Display of Data

The scatter plot matrix for the chosen variables is as shown below (Fig 1). There is a noticeable positive correlation between *y* and *x1*, as well as *y* and *x2*. As *x1* and *x2* increase, so does the response variable, *y*. On the other hand, there is no clear relationship between *y* and *x3*. It is also apparent that the variance of *y* is not consistent for all values of *x*.

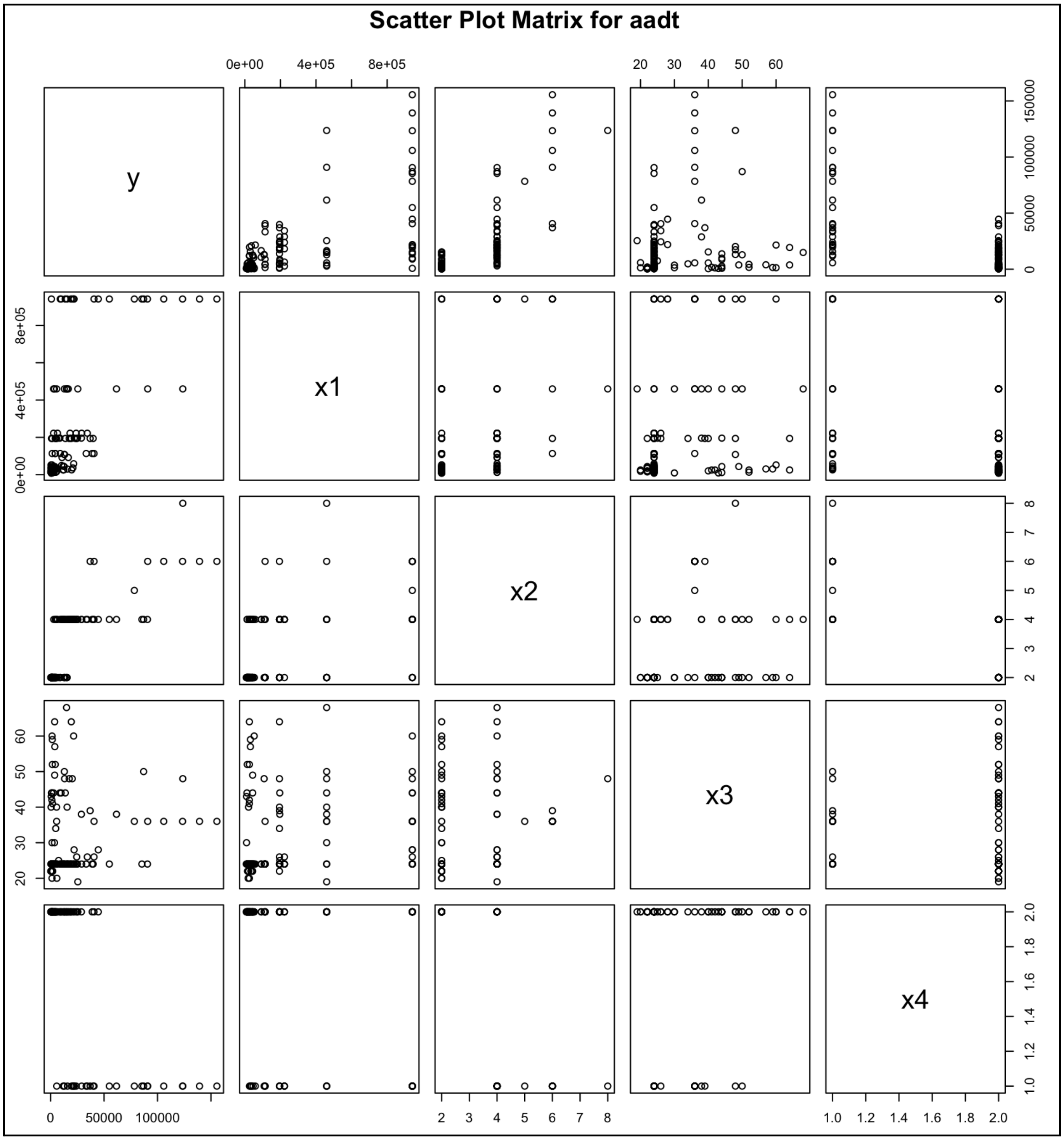

Fig 1

Note: *X4 labels were mapped to 0 and 1 hereafter.*

2. Modeling Multiple Linear Regression

At the outset, the objective was to evaluate the appropriateness of employing a linear regression model for the given task and to ascertain the adequacy of the chosen predictor variables in predicting the response variable. Hence, in the initial stage, the model was created simply from the predictor variables provided. Its summary is shown below (Fig 2).

```
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) -2.604e+04  5.255e+03  -4.955 2.49e-06 ***
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    0.421    
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
```

Fig 2

All the predictors except for *x3* are significant from the t tests. Furthermore, the F-ratio yielded a p-value less than *0.05*. Thus, the model is significant at *5%* confidence. Multiple and Adjusted R-squared were also relatively high.

Next, we plotted the Normal QQ plot (Fig 3). There seems to be points deviating from the normal line. The distribution is symmetrical but possibly with heavy tails, suggesting that the error terms may not follow a normal distribution.

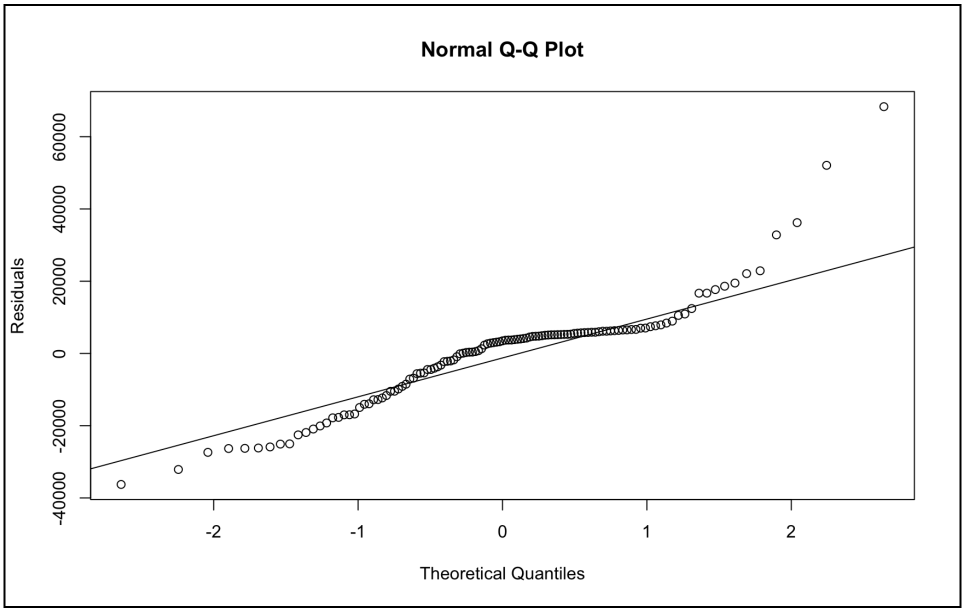

Fig 3

Next, we plotted the residual plots against time, fitted values and each predictor, as shown below (Fig 4). Note, we assume that the dataset is ordered by time of when it was collected.

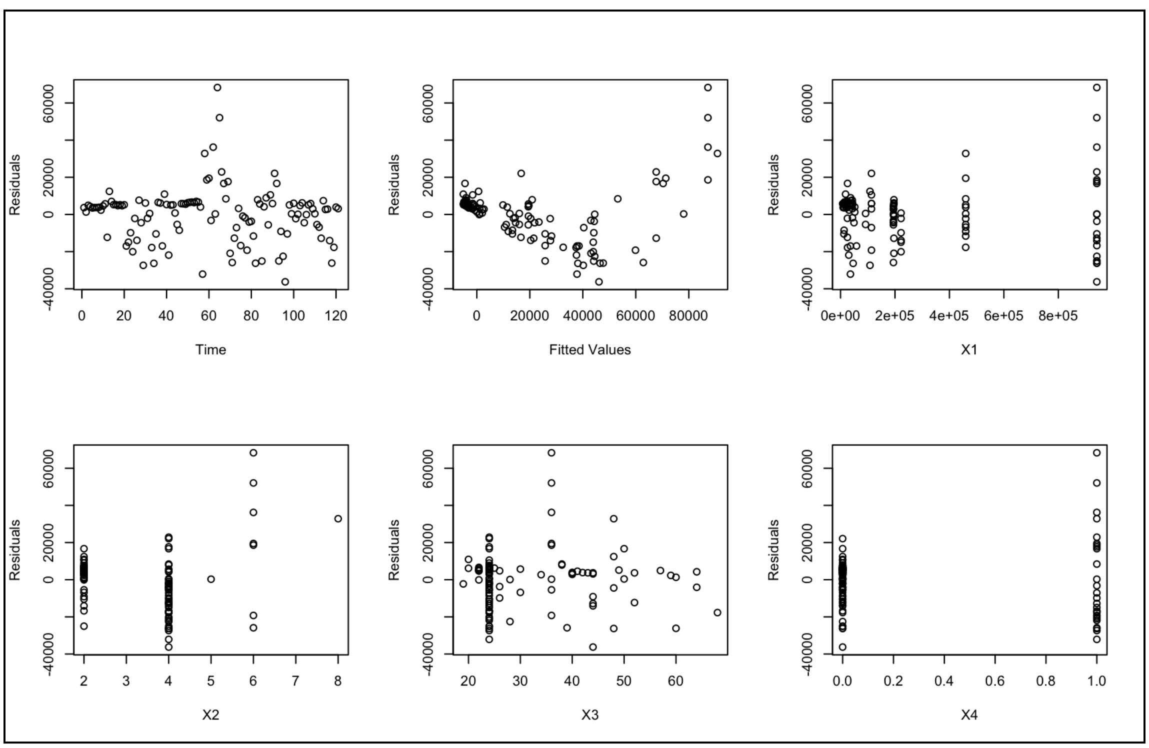

Fig 4

From the residual plotted against fitted values (first plot), it is apparent that a systematic pattern in the shape of a *'U'* can be observed, which suggests non-linearity in the relationship between predictors and response and that linear regression is not appropriate for the task. The U-shaped pattern may suggest that the relationship between the predictors and the response variable is not adequately captured by a linear model. Hence, we considered exploring non-linear relationships or introducing polynomial terms to better fit the data. It also appears that the variance of the error terms generally increases with respect to the fitted values.

Assuming the dataset is ordered sequentially in time, we performed the *Durbin-Watson* test(Fig 5). It is significantly less than 2 (*1.3137*) suggesting positive autocorrelation in the residuals. The very low p-value (*3.101e-05)* indicates that the evidence is strong against the null hypothesis of no autocorrelation. Positive autocorrelation in the residuals of a regression model means that there is a systematic pattern in the residuals that indicates a tendency for consecutive residuals to be positively correlated.

```
        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
```

Fig 5

3. Exploring New Predictors

From the scatter between *y* and *x3* as well as from the results of MLR, the predictor *X3* was not significant in predicting y. Hence we wanted to test the null hypothesis that its coefficient is zero. We performed the *ANOVA* test as shown below (Fig 6).

```
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 6

Since p-value > *0.05*, we do not have sufficient evidence to reject the null hypothesis, which means that *x3* is not a significant predictor. The following (Fig 7) shows the summary of this new model, removing *x3* as a predictor variable. We see an increase in adjusted R-squared from *0.7442* to *0.745*, which slightly improved the initial model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.340e+04  4.110e+03  -5.693 9.40e-08 ***
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
```

Fig 7

Next, we introduce polynomial terms $X1^2$ (Fig 8a) and $X2^2$ (Fig 8b) to add nonlinearity to the model as well as the interaction terms X1*X4 (Fig 8c), X1*X2 (Fig 8d) and X2*X4 (Fig 8e) as each of them likely interact with one another in real life scenarios. We included those terms whose null hypothesis we couldn't reject. In this case we included all terms except $X1^2$.

```
Model 1: Y ~ X1 + I(X1^2) + X2 + X4
Model 2: Y ~ X1 + X2 + X4
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1    116 2.7098e+10
2    117 2.7281e+10 -1 -182832366 0.7827 0.3782
```

Fig 8a

```
Model 1: Y ~ X1 + X2 + I(X2^2) + X4
Model 2: Y ~ X1 + X2 + X4
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    116 2.0041e+10
2    117 2.7281e+10 -1 -7239201969 41.901 2.396e-09 ***
```

Fig 8b

```
Model 1: y ~ x1 + x2 + x4 + I(x1 * x4)
Model 2: y ~ x1 + x2 + x4
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    116 1.1224e+10
2    117 2.7281e+10 -1 -1.6056e+10 165.93 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 8c

```
Model 1: y ~ x1 + x2 + x4 + I(x1 * x2)
Model 2: y ~ x1 + x2 + x4
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    116 1.4651e+10
2    117 2.7281e+10 -1 -1.2629e+10 99.992 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 8d

```
Model 1: y ~ x1 + x2 + x4 + I(x2 * x4)
Model 2: y ~ x1 + x2 + x4
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    116 1.8761e+10
2    117 2.7281e+10 -1 -8519293086 52.674 4.824e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 8e

Moreover, in adding these terms, we notice a significant increase in adjusted R-squared from *0.745* to *0.924*. The model still remained significant at the *5%* level. However, not all predictors were statistically significant (Figure 9). Hence another round of predictor variable reduction was performed.

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

Residuals:
   Min     1Q Median     3Q    Max
-29099  -2575  -1214   3013  32890

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.552e+03  1.140e+04   0.750   0.4549
x1          -8.011e-03  8.475e-03  -0.945   0.3465
x2          -6.478e+03  8.245e+03  -0.786   0.4337
x4          -4.532e+03  2.836e+04  -0.160   0.8733
I(x1 * x2)   6.611e-03  2.658e-03   2.487   0.0143 *
I(x1 * x4)   5.753e-02  7.007e-03   8.209 4.02e-13 ***
I(x2 * x4)   1.301e+03  6.999e+03   0.186   0.8528
I(x2 * x2)   1.826e+03  1.347e+03   1.355   0.1780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8316 on 113 degrees of freedom
Multiple R-squared:  0.9288,    Adjusted R-squared:  0.9244
F-statistic: 210.5 on 7 and 113 DF,  p-value: < 2.2e-16
```

Fig 9

We initially excluded the variable *x2* from the model due to its highest p-value, indicating its limited significance. The results of the ANOVA test (depicted in Fig 10a) provided insufficient evidence to reject the null hypothesis. Consequently, the full model, which includes *x2*, did not demonstrate a significant improvement over the reduced model without *x2*. Therefore, we conclude that *x2* is not a statistically significant predictor, and it was subsequently removed from the model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.552e+03  1.140e+04   0.750   0.4549
x1          -8.011e-03  8.475e-03  -0.945   0.3465
x2          -6.478e+03  8.245e+03  -0.786   0.4337
x4          -4.532e+03  2.836e+04  -0.160   0.8733
I(x1 * x2)   6.611e-03  2.658e-03   2.487   0.0143 *
I(x1 * x4)   5.753e-02  7.007e-03   8.209 4.02e-13 ***
I(x2 * x4)   1.301e+03  6.999e+03   0.186   0.8528
I(x2 * x2)   1.826e+03  1.347e+03   1.355   0.1780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8316 on 113 degrees of freedom
Multiple R-squared:  0.9288,    Adjusted R-squared:  0.9244
F-statistic: 210.5 on 7 and 113 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x4 + I(x1 * x2) + I(x1 * x4) + I(x2 * x4) + I(x2 *
    x2)
Model 2: y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x4) + I(x2 * x2)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    113 7814768512
2    114 7816534262 -1  -1765749 0.0255 0.8733
```

Fig 10a

Subsequently, we proceeded to eliminate the interaction term _x2*x4_ from the model due to its elevated p-value, signifying its limited significance. The ANOVA test results (as illustrated in Fig 10b) failed to provide sufficient evidence to reject the null hypothesis. Following the same rationale as before, we removed the _x2*x4 term from the model, as it did not contribute significantly to the model's performance._

```
Residuals:
   Min     1Q Median     3Q    Max
-29300  -2577  -1215   3027  33098

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.022e+04  4.509e+03   2.268  0.02522 *
x1          -8.327e-03  8.206e-03  -1.015  0.31243
x2          -7.723e+03  2.690e+03  -2.871  0.00488 **
I(x1 * x2)   6.716e-03  2.564e-03   2.619  0.01001 *
I(x1 * x4)   5.756e-02  6.974e-03   8.254 3.03e-13 ***
I(x2 * x4)   1.911e+02  8.401e+02   0.227  0.82047
I(x2 * x2)   2.031e+03  4.073e+02   4.987 2.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8280 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
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x4) + I(x2 * x2)
Model 2: y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x2)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    114 7816534262
2    115 7820081848 -1  -3547587 0.0517 0.8205
```

Fig 10b

In the final stage of our model refinement process (depicted in Fig 10c), we excluded the predictor *x1* due to its associated p-value. Subsequent to this removal, the predictors in the resulting reduced model (as shown in Fig 10d) all exhibit statistical significance. This step concludes our model selection, ensuring that the retained predictors are all deemed statistically significant.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.019e+04  4.487e+03   2.270  0.02505 *
x1          -7.859e-03  7.912e-03  -0.993  0.32265
x2          -7.778e+03  2.668e+03  -2.915  0.00427 **
I(x1 * x2)   6.483e-03  2.340e-03   2.770  0.00653 **
I(x1 * x4)   5.868e-02  4.914e-03  11.943  < 2e-16 ***
I(x2 * x2)   2.068e+03  3.730e+02   5.543 1.91e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8246 on 115 degrees of freedom
Multiple R-squared:  0.9287,	Adjusted R-squared:  0.9256
F-statistic: 299.7 on 5 and 115 DF,  p-value: < 2.2e-16
```

```
Model 1: y ~ x1 + x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x2)
Model 2: y ~ x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x2)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    115 7820081848
2    116 7887174273 -1 -67092425 0.9866 0.3226
```

Fig 10c

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.978e+03  4.482e+03   2.226  0.02792 *
x2          -8.196e+03  2.634e+03  -3.111  0.00235 **
I(x1 * x2)   4.315e-03  8.443e-04   5.111 1.28e-06 ***
I(x1 * x4)   6.044e-02  4.583e-03  13.190  < 2e-16 ***
I(x2 * x2)   2.193e+03  3.509e+02   6.251 6.99e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8246 on 116 degrees of freedom
Multiple R-squared:  0.9281,	Adjusted R-squared:  0.9256
F-statistic: 374.4 on 4 and 116 DF,  p-value: < 2.2e-16
```

Fig 10d

4. Handling non-normal residuals

With these nonlinear terms, we have a model that significantly performs better than the first model with an adjusted R-squared of 0.9256. However, it still suffers from a non-normally distributed error term as shown below (Fig 11).

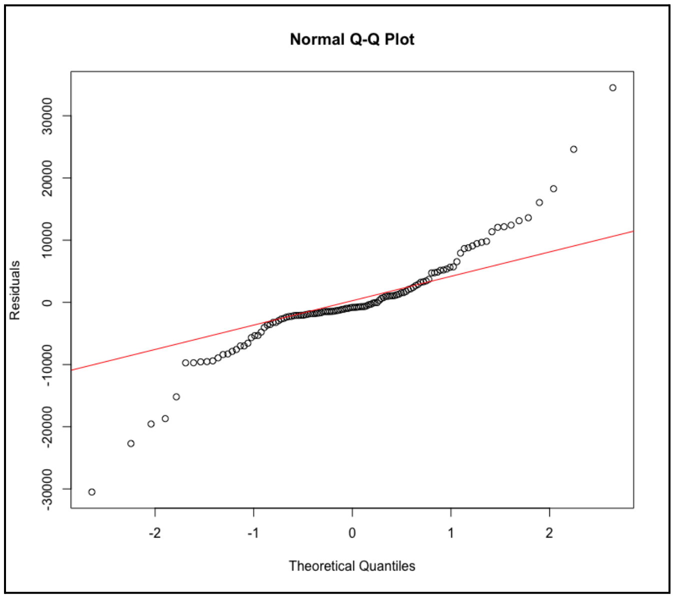

Fig 11

Thus, we applied the square root to the response variable *y*. The resulting plot (Fig 12) implies a more normally distributed error term as shown below.

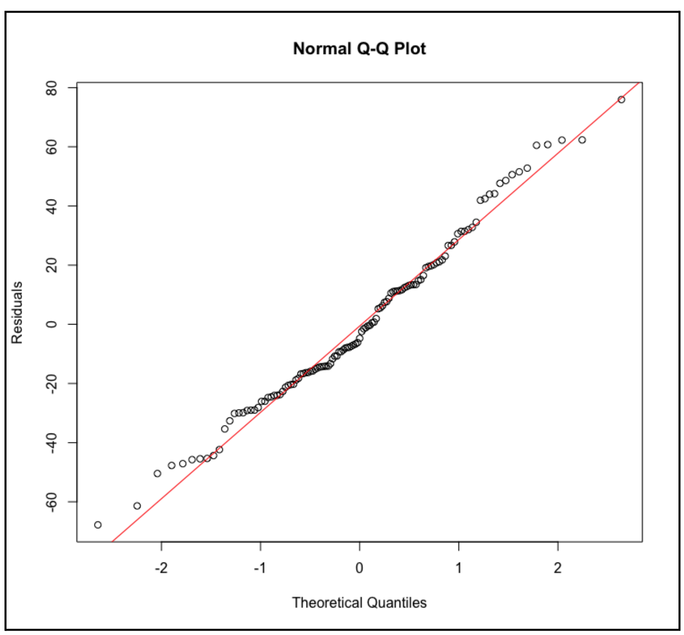

Fig 12

Furthermore, the new model is still significant at the *5%* level (Fig 13). However, some predictors (*X2^2*) are no longer significant.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.053e+01  1.602e+01  -1.282 0.202474
x2           3.582e+01  9.415e+00   3.805 0.000228 ***
I(x1 * x2)   1.385e-05  3.017e-06   4.590 1.13e-05 ***
I(x1 * x4)   1.199e-04  1.638e-05   7.319 3.53e-11 ***
I(x2 * x2)  -4.941e-01  1.254e+00  -0.394 0.694260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.47 on 116 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8818
F-statistic: 224.7 on 4 and 116 DF,  p-value: < 2.2e-16
```

Fig 13

Thus, we perform another round of model reduction, removing *X2^2*. The results from the *ANOVA* table (Fig 14) shows that there is not significant evidence to reject the null hypothesis that one performs better than the other. Thus we removed *X2^2*.

```
Analysis of Variance Table

Model 1: y ~ x2 + I(x1 * x2) + I(x1 * x4) + I(x2 * x2)
Model 2: y ~ x2 + I(x1 * x2) + I(x1 * x4)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    116 100741
2    117 100876 -1   -134.86 0.1553 0.6943
```

Fig 14

Note: We set *y* to be *sqrt(y)* from this point onwards (by aadt$y <- sqrt(aadt$y))

The summary of this new model shows that it is still statistically significant at the 5% level, with an adjusted R-squared of *0.8826* (Fig 15). In this model, all the predictors are now statistically significant.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.500e+01  7.688e+00  -1.951   0.0534 .
x2           3.227e+01  2.702e+00  11.945  < 2e-16 ***
I(x1 * x2)   1.385e-05  3.006e-06   4.607 1.05e-05 ***
I(x1 * x4)   1.189e-04  1.613e-05   7.372 2.61e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.36 on 117 degrees of freedom
Multiple R-squared:  0.8856,    Adjusted R-squared:  0.8826
F-statistic: 301.8 on 3 and 117 DF,  p-value: < 2.2e-16
```

Fig 15

With this final model, we plot the residuals against the fitted values (Fig 17) and there is a clear reduction in the pattern observed in Fig 4, suggesting the final model indeed fits the data better. Moreover, residuals plotted against original predictors indicates a residual distribution centered around 0 with relatively constant variance, suggesting a good model fit.

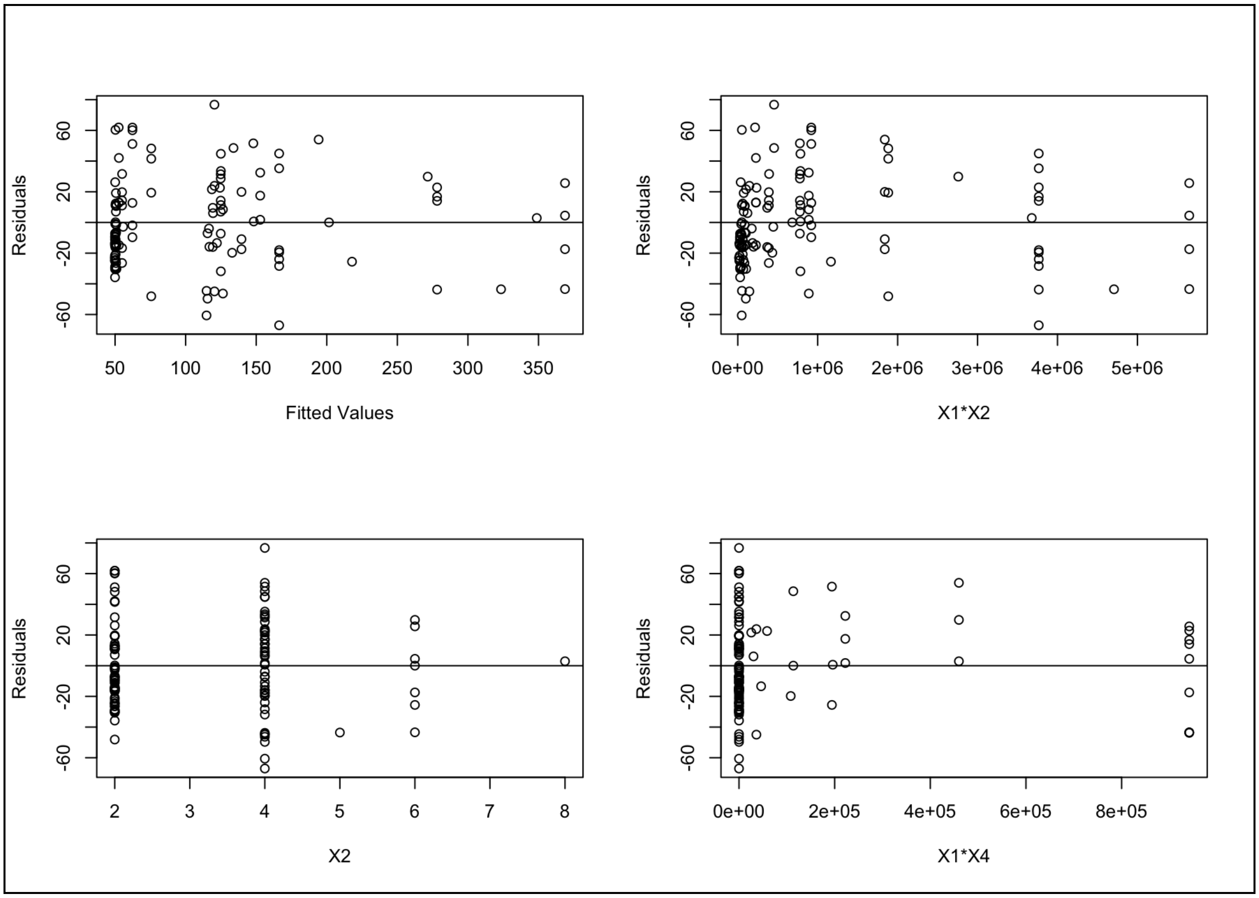

Fig 17

5. Prediction

Using the provided values of *x1=50000, x2=3 , x3=60 and x4=2*, our final model (model_three) gave a prediction of 7037.661 average annual daily traffic. In addition, we predict the 2 confidence intervals for the original model (model) and our final model (model_three).

	CI for mean	CI for new observations
model	[1045.888, 17167.99]	[-22236.34, 40450.22]
model_three	[6000.65, 8157.282]	[644.3826, 20276.86]

Model_three predicts *sqrt(y)*, therefore, we calculate the predictions using [lower^2, *upper^2*], which may explain why the 2 CIs for model_three are not centered around the same value.

For reference, the original, unsquared 2 CIs for *model_three* are:

- **CI for mean**: [77.46386, 90.31767]
- **CI for new observations**: [25.38469, 142.3968]

6. R Code

 # Display Scatter Plot Matrix

```
aadt_raw <- read.table('aadt.txt')
View(aadt_raw)
colnames(aadt_raw)[1]='y'
colnames(aadt_raw)[2]='x1'
colnames(aadt_raw)[3]='x2'
colnames(aadt_raw)[4]='x3'
colnames(aadt_raw)[5]='x4'
aadt_raw$x4[aadt_raw$x4==2]<-0
aadt <- aadt_raw[c(1,2,3,4,5)]

pairs(aadt[, c(1:5)], main = "Scatter Plot Matrix for aadt")
```

 # Multiple Linear Regression

```
model<-lm(y~x1+x2+x3+x4,data=aadt)
summary(model)
names(model)
modelS<-summary(model)
names(modelS)
```

 # QQ Plot (1)

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

```
# Residual Plot
par(mfrow=c(1,6))
plot(residuals(model),ylab='Residuals',xlab='Time')
plot(fitted(model),residuals(model),ylab='Residuals',xlab='Fitted values')
plot(aadt$x1,residuals(model),ylab='Residuals',xlab='x1')
plot(aadt$x2,residuals(model),ylab='Residuals',xlab='x2')
plot(aadt$x3,residuals(model),ylab='Residuals',xlab='x3')
plot(aadt$x4,residuals(model),ylab='Residuals',xlab='x4')

# Sequential Dependence (Durbin-Watson test)
# install.packages("lmtest")
library(lmtest)
dwtest(y ~ x1+x2+x3+x4, data=aadt)

# Test for whether some coefficients are zeros (1)
model_nox3  <- lm(y~x1+x2+x4,data=aadt)
anova(model_nox3,model)
summary(model_nox3)

# Adding non-linear predictor variables
model_x1x1  <- lm(y~x1+x2+x4+I(x1^2),data=aadt)
anova (model_x1x1,model_nox3)

model_x2x2  <- lm(y~x1+x2+x4+I(x2^2),data=aadt)
anova (model_x2x2,model_nox3)

model_x1x2  <- lm(y~x1+x2+x4+I(x1*x2),data=aadt)
anova (model_x1x2,model_nox3)

model_x2x4  <- lm(y~x1+x2+x4+I(x2*x4),data=aadt)
anova (model_x2x4,model_nox3)

model_x1x4  <- lm(y~x1+x2+x4+I(x1*x4),data=aadt)
anova (model_x1x4,model_nox3)
```

```
model_seven  <- lm(y~x1+x2+x4+I(x2*x2)+ I(x1*x4) +I(x1*x2) + I(x2*x4),data=aadt)
summary(model_seven)
```

 # Test for whether some coefficients are zeros (2)

```
model_six   <- lm(y~x1+x2+I(x2*x2)+ I(x1*x4) +I(x1*x2) + I(x2*x4),data=aadt)
anova(model_seven,model_six)
summary(model_six)

model_five   <- lm(y~x1+x2+I(x2*x2)+I(x1*x2) + I(x1*x4),data=aadt)
anova(model_six, model_five)
summary(model_five)

model_four    <- lm(y~x2+I(x2*x2) +I(x1*x2) + I(x1*x4),data=aadt)
anova(model_five, model_four)
summary(model_four)
```

 # QQ Plot (2) - Normality Checking

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

 # Square Root Transformation

```
aadt$y <- sqrt(aadt$y)
model_four    <- lm(y~x2+I(x2*x2) +I(x1*x2) + I(x1*x4),data=aadt)
summary(model_four)
```

 # Test for whether some coefficients are zeros (3)

```
model_three <- lm(y~x2 +I(x1*x2) + I(x1*x4),data=aadt)
anova(model_four, model_three)
summary(model_three)
```

 # Residual Plot

```
par(mfrow=c(2,2))
plot(fitted(model_three),residuals(model_three),ylab='Residuals',xlab='Fitted values')
abline(h=0)
plot(aadt$x1*aadt$x2,residuals(model_three),ylab='Residuals',xlab='x1*x2')
abline(h=0)
plot(aadt$x2,residuals(model_three),ylab='Residuals',xlab='x2')
abline(h=0)
plot(aadt$x1*aadt$x4,residuals(model_three),ylab='Residuals',xlab='x1*x4')
abline(h=0)
```

 # Prediction

```
new_data <- data.frame(x1 = 50000, x2 = 3, x3 = 60, x4 =0)
predicted_Y <- predict(model_three, newdata = new_data)
predicted_Y^2
```

 # Confidence Interval: model

```
modelS = summary(model)
con <- c(1, 50000,3,60,0)
lhat <- sum(con*coef(model))
lhat
t05 <- qt(0.975, 116)
bm <- t05*modelS$sigma*sqrt(con%*%modelS$cov.unscaled%*%con)
c(lhat-bm,lhat+bm)
c3 <- 1
bm <- t05*modelS$sigma*sqrt(con%*%modelS$cov.unscaled%*%con+c3)
c(lhat-bm,lhat+bm)
con <- data.frame(x1=50000,x2=3,x3=60,x4=0)
predict(model,con,interval='confidence',level=0.95)
predict(model,con,interval='prediction',level=0.95)
```

 # Confidence Interval: model_three

```
model_threeS = summary(model_three)
con <- c(1,3,50000*3,0)
lhat <- sum(con*coef(model_three))
lhat
t05 <- qt(0.975, 117)
bm <- t05*model_threeS$sigma*sqrt(con%*%model_threeS$cov.unscaled%*%con)
c(lhat-bm,lhat+bm)
c3 <- 1
bm <- t05*model_threeS$sigma*sqrt(con%*%model_threeS$cov.unscaled%*%con+c3)
c(lhat-bm,lhat+bm)
con <- data.frame(x1=50000,x2=3,x3=60,x4=0)
predict(model_three,con,interval='confidence',level=0.95)^2
predict(model_three,con,interval='prediction',level=0.95) ^2
```