MH3510 Project Report

Neo Shun Xian Nicholas

25 October 2020

Background

This data is 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 the number of vehicles that pass through a particular section of a road each day.

Response Variable:
y: Average annual daily traffic

Predictor Variable:
X1: Population of country in which road section is located
X2: Number of lanes in road section
X3: Width of the road section (in feet)
X4: Two-category quality variable indicating whether or not there is control of access to road section (1=access control, 2=no access control)
Can define X4 to be 1 if there is access control and 0 if no control

Numerical predictor variable: X1, X2, X3
Categorical predictor varable: X4

Importing the dataset

```
# read the dataset
data_raw = read.table("aadt.txt",header=FALSE)
# check data by showing the first few entries of data
head(data_raw,5)
```

```
##     V1    V2 V3 V4 V5 V6 V7 V8
## 1 1616 13404  2 52  2  2  5  1
## 2 1329 52314  2 60  2  2  5  1
## 3 3933 30982  2 57  2  4  5  2
## 4 3786 25207  2 64  2  4  5  2
## 5  465 20594  2 40  2  2  5  1
```

Preprocess raw data

```
# select only the columns with the predictor and response variable
data = data.frame(y=data_raw$V1,
                  x1=data_raw$V2,
                  x2=data_raw$V3,
                  x3=data_raw$V4,
                  x4=data_raw$V5)

#defining X4 to be 1 if there is access control and 0 if there isn't
```

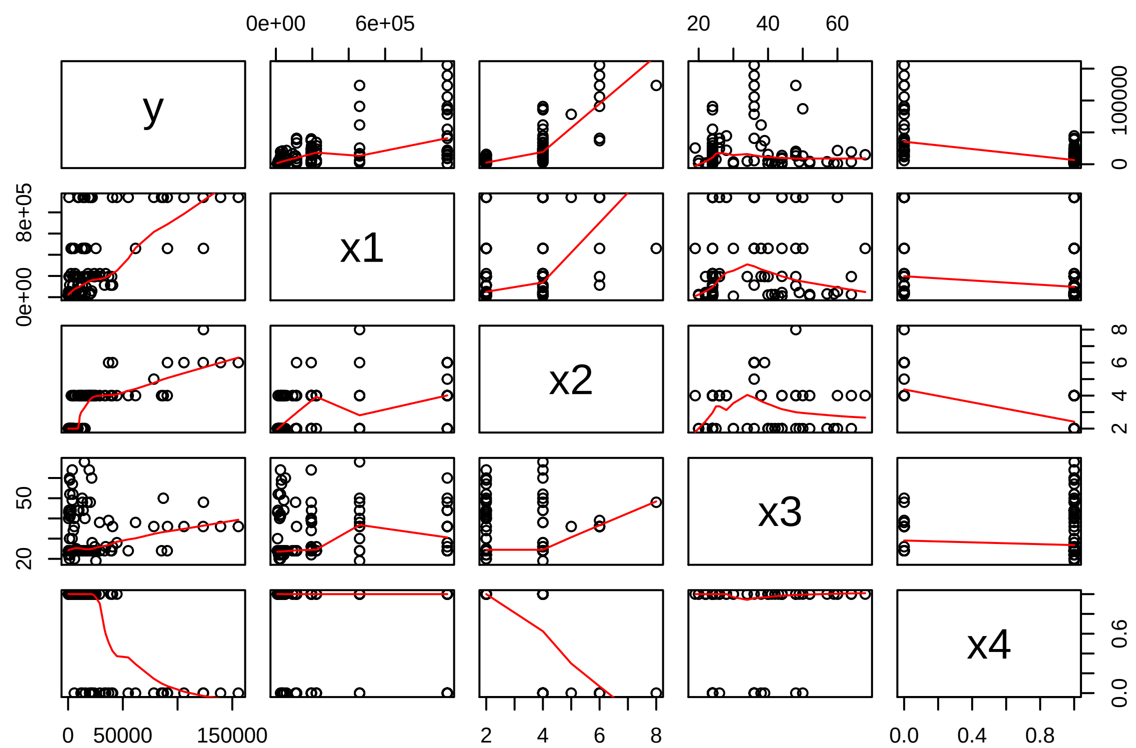
```r
data$x4[data$x4==1] <- 0
data$x4[data$x4==2] <- 1

# take a look at the modification to the data
head(data,5)
```

```
##      y    x1 x2 x3 x4
## 1 1616 13404  2 52  1
## 2 1329 52314  2 60  1
## 3 3933 30982  2 57  1
## 4 3786 25207  2 64  1
## 5  465 20594  2 40  1
```

Scatter Plot Matrix

```r
# plot the scatter plot matrix
pairs(data,panel=panel.smooth)
```



From the plot above, there exist relations between each predictor variables (X1 to X4) and the response variable (y).

Model: Additional e^{x_2} term in the Multi Regression Model

From the plot above, there seemed to have an exponential relation between the response variable y and predictor variable x_2. Hence, we fit the data with an additional e^{x_2} term:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 e^{X_2} + \epsilon$$

This will be our suggested model 1.

Fit Model 1

```r
# fit a multiple linear regression model
mlr1 <- lm(y ~ x1+x2+x3+x4+I(exp(x2)), data=data)
summary(mlr1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + I(exp(x2)), data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33494  -7325   2685   4733  68345
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.811e+03  7.781e+03   0.875 0.383208
## x1           3.482e-02  4.480e-03   7.773 3.55e-12 ***
## x2           6.228e+03  1.640e+03   3.798 0.000235 ***
## x3           3.758e+01  1.188e+02   0.316 0.752218
## x4          -2.403e+04  4.279e+03  -5.615 1.39e-07 ***
## I(exp(x2))   2.202e+01  5.780e+00   3.811 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14470 on 115 degrees of freedom
## Multiple R-squared:  0.7805, Adjusted R-squared:  0.7709
## F-statistic: 81.77 on 5 and 115 DF,  p-value: < 2.2e-16
```

Check the adequacy using t-value (standard error) and F-test between 2 models

From the result of model 1, predictor X_3 is not significant due to very small t-value. Hence, it's coefficient may be equal to zero i.e

$$H_0 : \beta_3 = 0$$

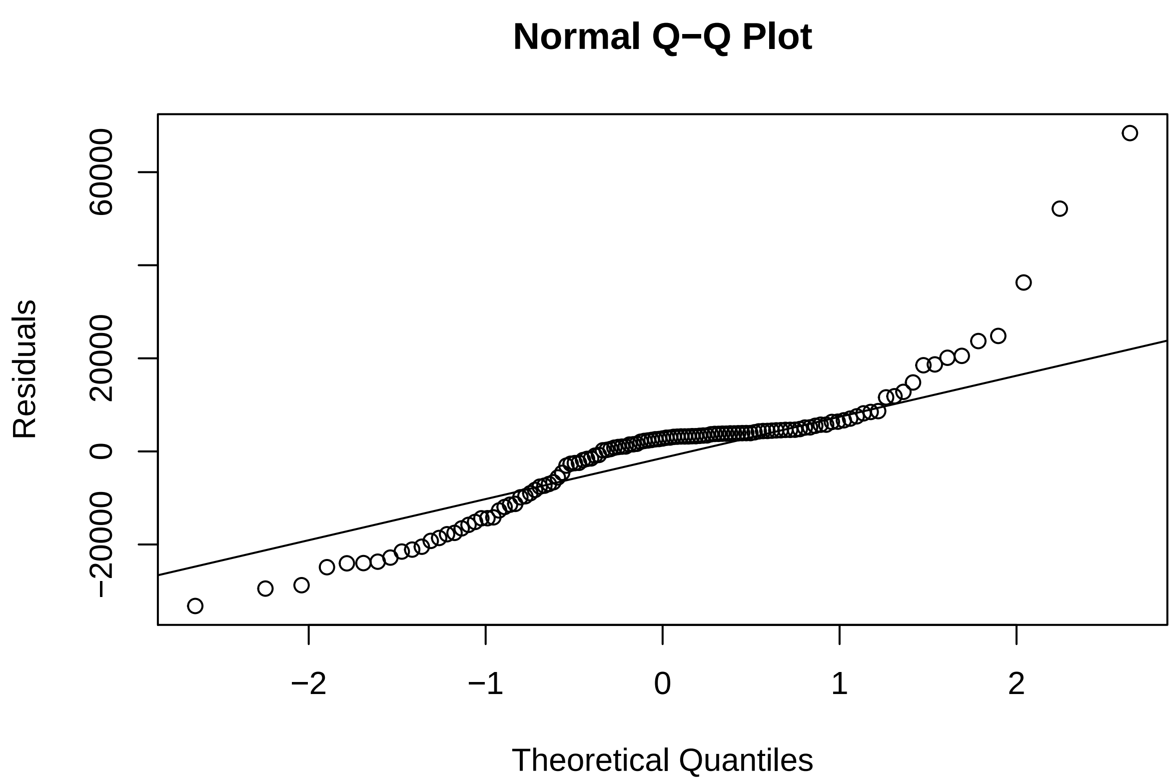
```r
# F-test between 2 models (full and model without predictor variable X3)
mlr1_alt <- lm(y ~ x1+x2+x4+I(exp(x2)), data=data)
# ANOVA
anova(mlr1_alt,mlr1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x4 + I(exp(x2))
## Model 2: y ~ x1 + x2 + x3 + x4 + I(exp(x2))
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    116 2.4108e+10
## 2    115 2.4087e+10  1  20977403 0.1002 0.7522
```

From the above analysis, since the F-value is small, we cannot reject the null hypothesis at 0.1 level of significance, therefore, we eliminate the predictor variable X_3 from further analysis

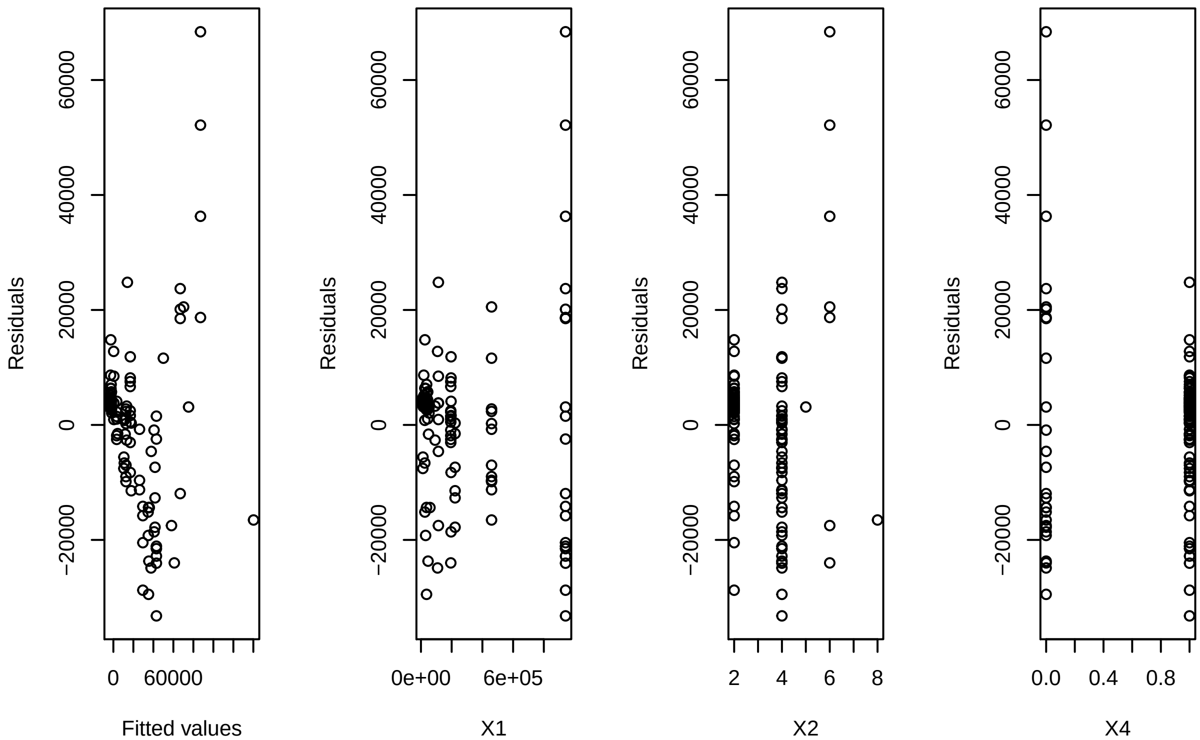
Check normality of residuals

```
# normality checking
qqnorm(residuals(mlr1_alt), ylab='Residuals')
qqline(residuals(mlr1_alt))
```



From the QQ plot above, we can see that the residuals are not normally distributed. We will now draw some plots of the residuals against each predictor variable, X_1, X_2, X_4

```
par(mfrow=c(1,4))
plot(fitted(mlr1_alt), residuals(mlr1_alt), ylab='Residuals', xlab='Fitted values')
plot(data$x1, residuals(mlr1_alt), ylab='Residuals', xlab='X1')
plot(data$x2, residuals(mlr1_alt), ylab='Residuals', xlab='X2')
plot(data$x4, residuals(mlr1_alt), ylab='Residuals', xlab='X4')
```



From the residuals against fitted values plot, we can observe that the variances of residuals have increased as the fitted values increases. From the residuals against X_1 plot, there seemed to have a linear relationship between them. There isn't any obvious pattern seen from the plot of residuals against X_2 and residuals against X_4 plot. Therefore we can try to propose a model that includes the interaction between X_1 and X_2 as well as X_1 and X_4.

Hence, we fit the model:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 e^{X_2} + \beta_6 X_1 X_2 + \beta_7 X_1 X_4 + \epsilon$$

Fit Model 2

```r
# fit a multiple linear regression model
mlr2 <- lm(y ~ x1+x2+x4+I(exp(x2))+I(x1*x2)+I(x1*x4), data=data)
summary(mlr2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x4 + I(exp(x2)) + I(x1 * x2) + I(x1 *
##     x4), data = data)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -32955  -2475  -1046   2947  36566
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.123e+03  5.823e+03  -0.880 0.380866
## x1           4.328e-02  1.328e-02   3.258 0.001477 **
## x2           4.498e+03  1.240e+03   3.627 0.000430 ***
## x4          -1.158e+03  3.540e+03  -0.327 0.744148
## I(exp(x2))   1.538e+01  3.475e+00   4.427 2.20e-05 ***
## I(x1 * x2)   8.882e-03  2.619e-03   3.391 0.000957 ***
## I(x1 * x4)  -5.902e-02  7.212e-03  -8.184 4.38e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8610 on 114 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9189
## F-statistic: 227.6 on 6 and 114 DF,  p-value: < 2.2e-16
```

Check the adequacy using t-value (standard error) and F-test between 2 models

From the result of model 2, the term X_4 is not significant due to very small t-value. Hence, it's coefficient may be equal to zero **after including interaction terms $\mathbf{X_1X_2}$ and $\mathbf{X_1X_4}$** i.e

$$H_0 : \beta_4 = 0$$

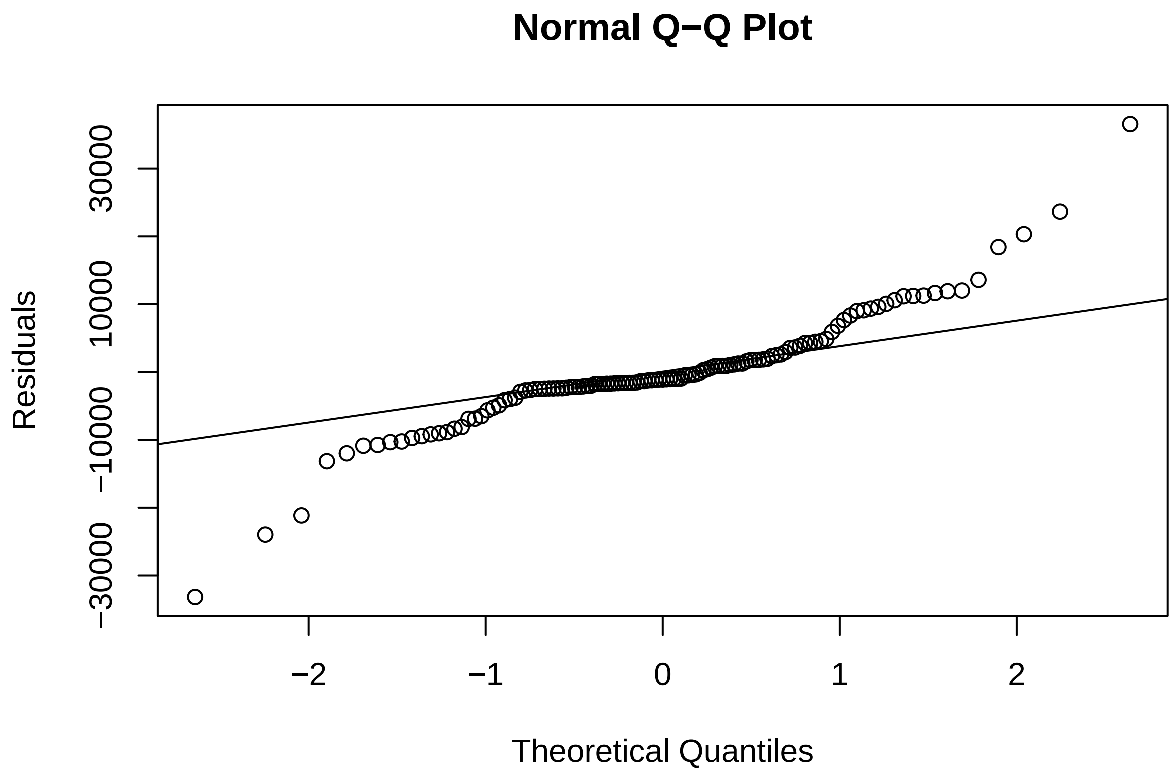
```
# F-test between 2 models (full and model without predictor variable X4)
mlr2_alt <- lm(y ~ x1+x2+I(exp(x2))+I(x1*x2)+I(x1*x4), data=data)
# ANOVA
anova(mlr2_alt,mlr2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + I(exp(x2)) + I(x1 * x2) + I(x1 * x4)
## Model 2: y ~ x1 + x2 + x4 + I(exp(x2)) + I(x1 * x2) + I(x1 * x4)
##   Res.Df        RSS Df Sum of Sq     F Pr(>F)
## 1    115 8459619834
## 2    114 8451684672  1   7935162 0.107 0.7441
```

From the above analysis, since the F-value is small, we cannot reject the null hypothesis at 0.1 level of significance, therefore, we eliminate the predictor variable X_4 from further analysis.

Check normality of residuals

```
# normality checking
qqnorm(residuals(mlr2_alt), ylab='Residuals')
qqline(residuals(mlr2_alt))
```



From the QQ plot above, we can see that the residuals are not normally distributed, but it seems to be closer to normality as compared to the alternative model, model 1. Hence we will declare the alternative model of model 2 as model 3, with X_4 term removed, to do further analysis. i.e.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 e^{X_2} + \beta_6 X_1 X_2 + \beta_7 X_1 X_4 + \epsilon$$

Fit Model 3

```
# fit a multiple linear regression model
mlr3 <- lm(y ~ x1+x2+I(exp(x2))+I(x1*x2)+I(x1*x4), data=data)
summary(mlr3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + I(exp(x2)) + I(x1 * x2) + I(x1 * x4),
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33167  -2469  -1104   2596  36556
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.784e+03  2.840e+03  -2.389 0.018542 *
```

```
## x1            4.569e-02  1.101e-02   4.149 6.41e-05 ***
## x2            4.746e+03  9.768e+02   4.859 3.76e-06 ***
## I(exp(x2))    1.537e+01  3.461e+00   4.440 2.08e-05 ***
## I(x1 * x2)    8.514e-03  2.355e-03   3.614 0.000448 ***
## I(x1 * x4)   -6.067e-02  5.124e-03 -11.840  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8577 on 115 degrees of freedom
## Multiple R-squared:  0.9229, Adjusted R-squared:  0.9195
## F-statistic: 275.3 on 5 and 115 DF,  p-value: < 2.2e-16
```

Check the adequacy using t-value (standard error) and F-test between 2 models

From the result of model 3, it seems like the model fits quite well with reasonably large t-values and also the fact that the R-squared and adjusted R-squared value 0.9229 and 0.9195 respectively, which are reasonably high values. We will still try to remove one term with the lowest t-value among the terms, X_1X_2 to see if the model is a better fit to the data. i.e we test our hypothesis:

$$H_0 : \beta_6 = 0$$

```r
# F-test between 3 models (full and model without predictor variable X1X2)
mlr3_alt <- lm(y ~ x1+x2+I(exp(x2))+I(x1*x4), data=data)
# ANOVA
anova(mlr3_alt,mlr3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + I(exp(x2)) + I(x1 * x4)
## Model 2: y ~ x1 + x2 + I(exp(x2)) + I(x1 * x2) + I(x1 * x4)
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    116 9420645384
## 2    115 8459619834  1 961025550 13.064 0.000448 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above analysis, since the F-value is large, we reject the null hypothesis at 0.1 level of significance, therefore, we **do not** eliminate the predictor variable X_1X_2 from further analysis.

Final Model

After looking at the F-test, R-squared & adjusted R-squared as well as the significance of t-test, and also the comparison of the full and reduced model for model simplification, we proposed the following model to fit the data:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 e^{X_2} + \beta_6 X_1 X_2 + \beta_7 X_1 X_4 + \epsilon$$

Prediction

Using x1=50000, x2=3, x3=60 and x4=2 to fit our final model.

```r
pred_data = data.frame(x1=50000, x2=3, x4=2)
# confidence interval of mean response
predict(mlr3,pred_data,interval='confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 5258.023 3257.955 7258.091
```

```r
# confidence interval of the predicted y value
predict(mlr3,pred_data,interval='prediction', level=0.95)
```

```
##        fit       lwr      upr
## 1 5258.023 -11848.34 22364.39
```