MH4518 Project

Done by Keerthana Jayaraman Karthikeyan (U2123613E), Nicholas Yap (U2020269C), Peh Wei Hang (U2022903K) and Kng Yew Chian (U2021777G)

Importing the Dataset

The dataset we are using in this project is the average annual daily traffic dataset, given in a file named `aadt.txt`.

In this project, we aim to predict the average no. of vehicles passing through a particular section of the road each day (Y), using the following predictor variables:

- Population of county in which road section is located (X1)
- No. of lanes in road section (X2)
- Width of road section (X3)
- Whether there is control of access to the section, a 2-category variable (X4)

We first import the dataset below:

```
data <- readLines("aadt.txt")
```

```
## Warning in readLines("aadt.txt"): incomplete final line found on 'aadt.txt'
```

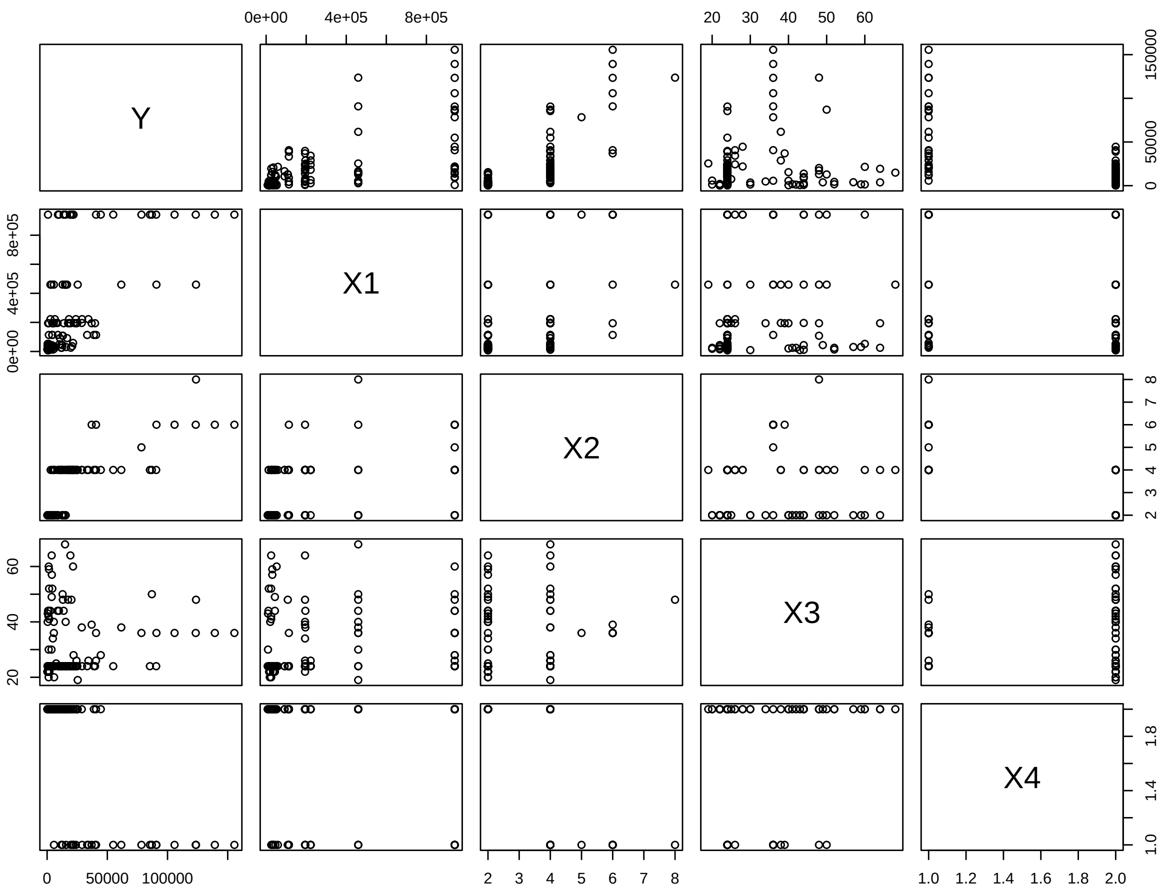
```
data <- gsub("\\s+", " ", data) # replace spaces with a single space
con <- textConnection(data)
aadt <- read.table(con, header=F)
close(con)
colnames(aadt) = c("Y", "X1", "X2", "X3", "X4", "X5", "X6", "X7")

data_raw <- read.table("aadt.txt", header=FALSE)
aadt <- data.frame(y=data_raw$V1,x1=data_raw$V2,x2=data_raw$V3,x3=data_raw$V4,x4=data_raw$V5)
colnames(aadt) = c("Y", "X1", "X2", "X3", "X4")
```

Investigating response/predictor relationships

We first investigate the individual linear relationships between Y and X_1 / X_2 / X_3 / X_4, as they are our primary variables of interest.

```
plot(aadt[c("Y", "X1", "X2", "X3", "X4")])
```



We see a slight indication of a positive linear relationship between X_1 and Y, as well as X_2 and Y, but we are unable to tell whether a linear relationship exists between X_3 / X_4 and Y.

For the predictor variables (X_1 , X_2, X_3, X_4), we also do not see any strong indication of relationships among them.

Fitting SLR models

We attempt to fit SLR models to each of the predictor variables to examine their individual relationships with Y.

```
slr_x1 = lm(Y~X1, data=aadt)
summary(slr_x1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57290  -6597  -3667   5884  97501
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.437e+03  2.776e+03   1.598    0.113
## X1          5.695e-02  6.598e-03   8.631 3.18e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23810 on 119 degrees of freedom
## Multiple R-squared:  0.385,  Adjusted R-squared:  0.3798
## F-statistic:  74.5 on 1 and 119 DF,  p-value: 3.183e-14
```

```r
slr_x2 = lm(Y~X2, data=aadt)
summary(slr_x2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34036 -13493   1332   4417  84534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -35664       4617  -7.724 3.92e-12 ***
## X2             17780       1375  12.934  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19580 on 119 degrees of freedom
## Multiple R-squared:  0.5843, Adjusted R-squared:  0.5808
## F-statistic: 167.3 on 1 and 119 DF,  p-value: < 2.2e-16
```

```r
slr_x3 = lm(Y~X3, data=aadt)
summary(slr_x3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27329 -15936 -11727   3521 134552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9499.2     7912.8   1.200    0.232
## X3             319.3      238.5   1.339    0.183
##
## Residual standard error: 30140 on 119 degrees of freedom
## Multiple R-squared:  0.01484,    Adjusted R-squared:  0.006561
## F-statistic: 1.792 on 1 and 119 DF,  p-value: 0.1832
```

```r
slr_x4 = lm(Y~X4, data=aadt)
summary(slr_x4)
```

```
##
```

```
## Call:
## lm(formula = Y ~ X4, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52254  -7498  -4578   6600  97596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   107526       8810   12.21   <2e-16 ***
## X4            -49575       4827  -10.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22110 on 119 degrees of freedom
## Multiple R-squared:  0.4699, Adjusted R-squared:  0.4654
## F-statistic: 105.5 on 1 and 119 DF,  p-value: < 2.2e-16
```

From the above t-test results, we see that X_1, X_2, and X_4 individually have very significant linear relationships with Y, while X_3 does not.

Fitting a MLR model

Here, we fit a MLR model to regress Y on X_1, X_2, X_3 and X_4.

```r
mlr_model_1 <- lm(Y~X1+X2+X3+X4, data=aadt)
summary(mlr_model_1)
```

```
##
## 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
```

As $F = 88.29 > F_{4,116}^{0.05} = 3.006917$, the correlation between our MLR model and Y is significant. Next, we look at goodness-of-fit of the model. Since $0.6 < R_{adjusted} = 0.7442 < 0.95$, we conclude that the MLR model is a good fit, and is able to explain 74.42% of variance in the response variable.

As the t-test for X_3 is not significant, we try removing it and forming another MLR model with only X_1, X_2 and X_4 as the predictors.

```r
mlr_model_2 <- lm(Y~X1+X2+X4, data=aadt)
summary(mlr_model_2)
```

```
##
## 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
```

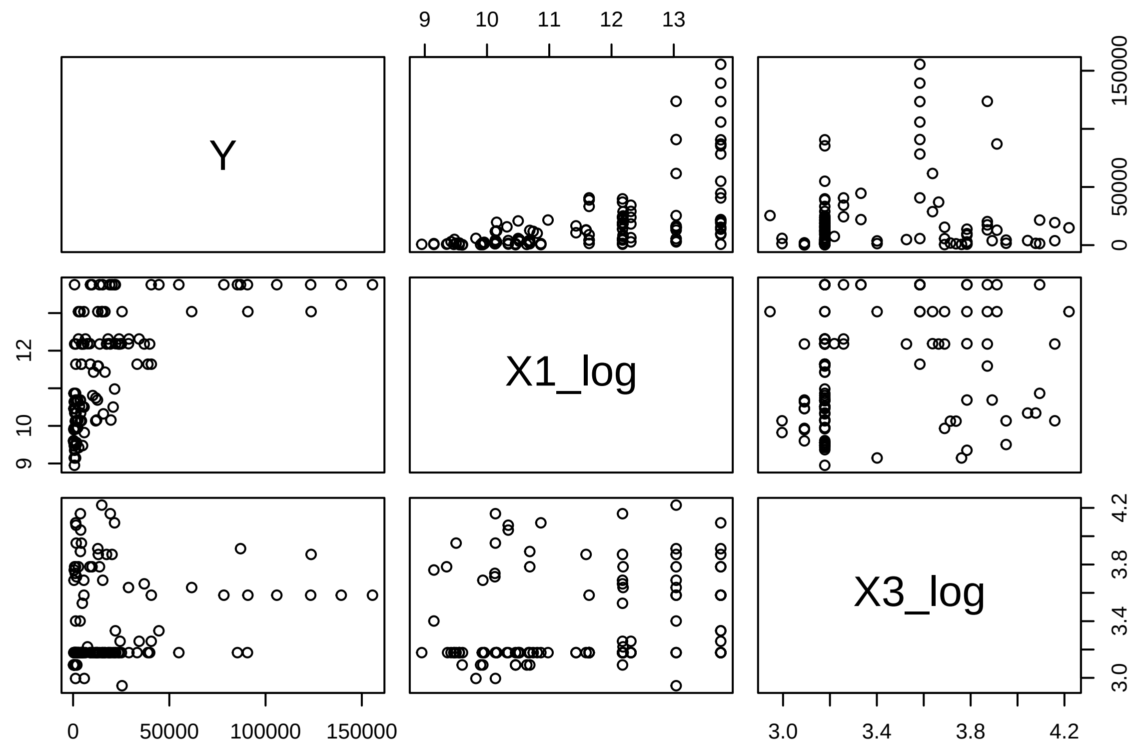
As $F = 117.8 > F^{0.05}_{3,117} = 2.682132$, the correlation between our MLR model and Y is also significant. As $R^2_{adjusted} = 0.745 > 0.7442$, this model is a slightly better fit than the earlier model.

Transformations of Variables

We next perform transformations for X_1 and X_3. We exclude X_2 as it represents the number of lanes with 5 discrete values from 2 to 8, which we treat as a categorical variable. We exclude X_4 as it is a categorical variable as well.

We first perform log-transformations and plot a scatterplot between Y and the log-transformed variables.

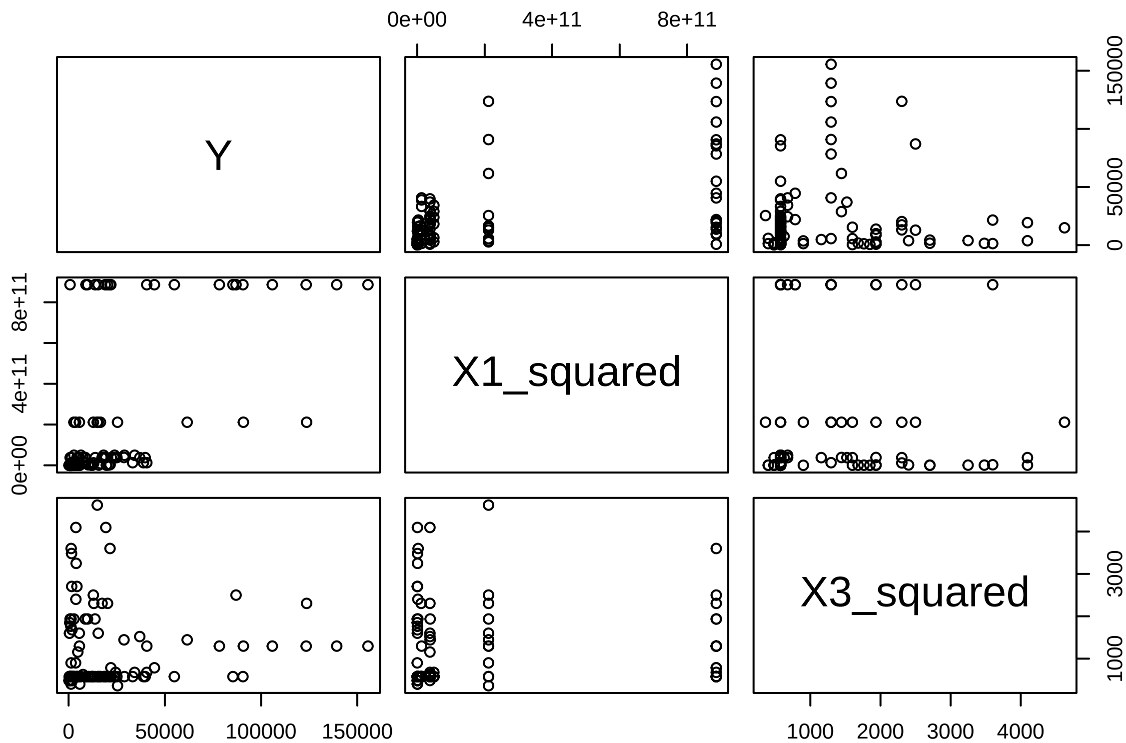
```r
aadt$X1_log <- log(aadt$X1)
aadt$X3_log <- log(aadt$X3)
plot(aadt[c("Y", "X1_log", "X3_log")])
```



From the above scatterplot, we see that Y and X_1 seem to have a linear relationship, but we are unable to conclude that for Y and X_3.

We next perform square transformation of Y, X_1^2, and X_3^2, and plot a scatterplot between them.

```
aadt$X1_squared <- aadt$X1^2
aadt$X3_squared <- aadt$X3^2
plot(aadt[c("Y", "X1_squared", "X3_squared")])
```



We see that majority of the values for both X_1^2 and X_3^2 are clustered towards the left, which does not look desirable.

We next try to form a model using $log(X_1)$, X_2 and X_4 as predictors for Y.

```
mlr_model_3 <- lm(Y~X1_log+X2+X4, data=aadt)
summary(mlr_model_3)
```

```
##
## Call:
## lm(formula = Y ~ X1_log + X2 + X4, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34732  -8949    741   7529  77362
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -34238      16936  -2.022   0.0455 *
## X1_log          5714       1185   4.821 4.34e-06 ***
## X2              9580       1716   5.583 1.55e-07 ***
## X4            -23648       4898  -4.828 4.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16760 on 117 degrees of freedom
## Multiple R-squared:  0.7004, Adjusted R-squared:  0.6927
```

```
## F-statistic: 91.16 on 3 and 117 DF,  p-value: < 2.2e-16
```

This model does not perform as well, as its $R^2_{adjusted} = 0.693$ which is significantly lower than the first two models.

We then try to fit a model using X_1, X_2, $log(X_3)$ and X_4 next.

```r
mlr_model_4 <- lm(Y~X1+X2+X3_log+X4, data=aadt)
summary(mlr_model_4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3_log + X4, data = aadt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36603  -8272   3038   6286  68130
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.920e+03  1.756e+04   0.508    0.612
## X1           3.277e-02  4.716e-03   6.948 2.30e-10 ***
## X2           9.110e+03  1.529e+03   5.960 2.78e-08 ***
## X3_log       4.653e+03  4.493e+03   1.036    0.302
## X4          -2.368e+04  4.501e+03  -5.263 6.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15270 on 116 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.7451
## F-statistic: 88.71 on 4 and 116 DF,  p-value: < 2.2e-16
```

This model has $R^2_{adjusted} = 0.7451$, which is the highest among all the models.

Comparing Models with ANOVA

From above, we have 4 MLR models using the following predictor variables:

1. X_1, X_2, X_3, X_4
2. X_1, X_2, X_4
3. $log(X_1)$, X_2, X_4
4. X_1, X_2, $log(X_3)$, X_4

As model 2 is a reduced model of model 1 and model 4, we run an ANOVA between model 2 and 1, and between model 2 and 4.

```r
anova(mlr_model_1, mlr_model_2)
```

```
## 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
```

As $Pr(> F) = 0.4213$ which is non-significant, we conclude that model 1 does not lead to a significant improvement over model 2.

```r
anova(mlr_model_4, mlr_model_2)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X3_log + X4
## Model 2: Y ~ X1 + X2 + X4
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    116 2.7031e+10
## 2    117 2.7281e+10 -1 -249959550 1.0727 0.3025
```

Similarly, as $Pr(> F) = 0.3025$ which is non-significant, we conclude that model 4 does not lead to a significant improvement over model 2.

Prediction

Lastly, we try to predict Y given $X_1 = 50000, X_2 = 3, X_3 = 60, X_4 = 2$ on all the models.

```r
new_data = data.frame(X1=50000, X1_log=log(50000), X2=3, X3=60, X3_log=log(60), X4=2)
predict(mlr_model_1, new_data)
```

```
##       1
## 9106.94
```

```r
predict(mlr_model_2, new_data)
```

```
##        1
## 6207.477
```

```r
predict(mlr_model_3, new_data)
```

```
##        1
## 9026.108
```

```r
predict(mlr_model_4, new_data)
```

```
##        1
## 9570.147
```