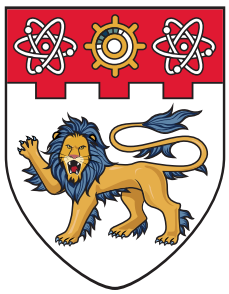

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3510 Regression Analysis

Group Project

Name	Email	Matric Number
Pu Fanyi	FPU001@e.ntu.edu.sg	U2220175K
Shan Yi	SH0005YI@e.ntu.edu.sg	U2222846C
Zhang Kaichen	ZHAN0564@e.ntu.edu.sg	U2123722J
Mu Yichen	M220100@e.ntu.edu.sg	U2240144L
Fu Yilin	FUYI0005@e.ntu.edu.sg	U2240718G

Nanyang Technological University, Singapore

2024/2025 Semester 1

1 Introduction

In this study, we investigate factors affecting Annual Average Daily Traffic (AADT) on road section, using a dataset comprising five variables shown in Table 1. We demonstrate our overall pipeline in Figure 1. After inputting our data, we first observe its distribution. To address the skewness in some of the data, we applied transformations to the data.

After that, we conduct Single Linear Regression (SLR) analysis for x_1 to x_4. We analyze and explain the influence of each variable on the response variable in detail through this approach. In parallel, we also perform Multiple Linear Regression (MLR) analysis to get the more precise model. For variables found to be significant, we proceed to use them for prediction.

The full code can be reached in https://pufanyi.github.io/MH3510-Project/, we also listed some key code snips in Appendix A.

2 Dataset Overview and Representation

The dataset is constituted of 5 columns, with meaning shown in Table 1. And Figure 2 illustrates the distributions of all variables.

In Figure 2, it is clear that y is left-skewed. We consider setting

$$y' = \log^2 y \tag{1}$$

to fix the skewness of the response variable. Equation

In further analysis, we will use y' as the response variable. To predict the value of y, we can easily do inverse transformation $y = \exp\sqrt{y'}$ to get the result.

Similarly, to fix the skewness of x_1, transformations are performed to x_1 with

$$x_1' = \sqrt{x_1} \tag{2}$$

Figure 3 shows the distribution of y' and x_1, it can be seen that the distributions of y' and x_1' are relatively uniform compared with the original y and x_1.

Although x_2 and x_3 also have some skewness, considering that they are integers[†] with a small range in the dataset, applying a transformation may not be useful. We have decided not to transform them for now.

3 Single Variable Regression Analysis

3.1 Population v.s. AADT

According to common sense, cities with larger populations should have more congested traffic. We attempt to use data to verify this hypothesis.

3.1.1 Model Overview

By transforming x_1 to $\sqrt{x_1}$ and y to $\log^2 y$, we ensured that potential non-linear relationships were better captured in the model.

We build a single linear regression model:

$$y' = \beta_0^{(1)} + \beta_1^{(1)} x_1' + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_1^2) \tag{3}$$

The model is fitted with the data in figure , with

$$\begin{cases} \beta_0^{(1)} \approx 53.49 \\ \beta_1^{(1)} \approx 0.064 \end{cases} \tag{4}$$

[*] $x_4 = 1$ means access control; $x_4 = 2$ means no access control

[†] Although x_3 is defined as a continuous variable, it consists entirely of integers in the dataset.

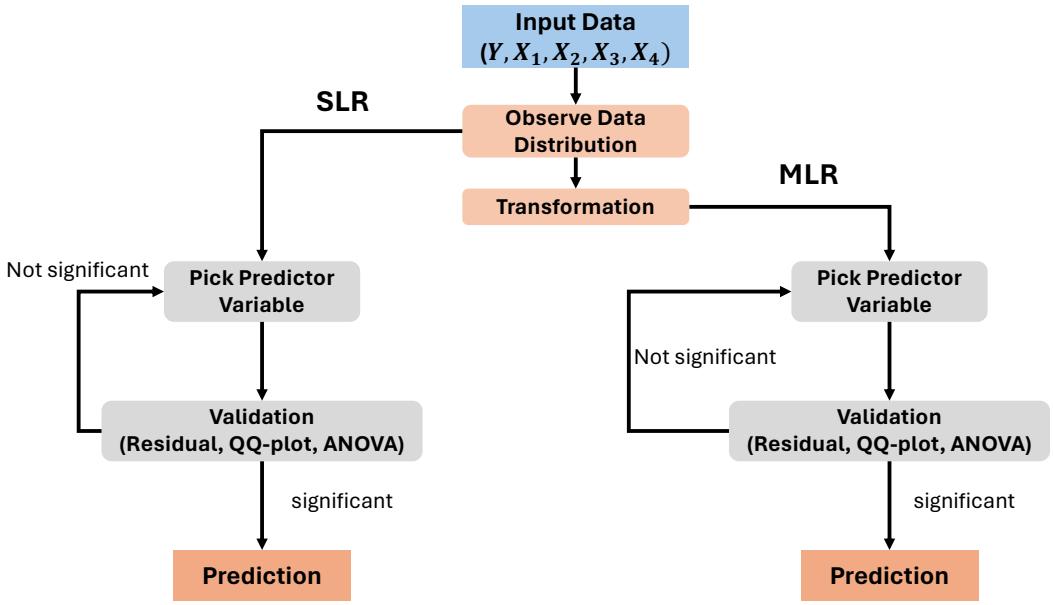

Figure 1: An overview of our pipeline.

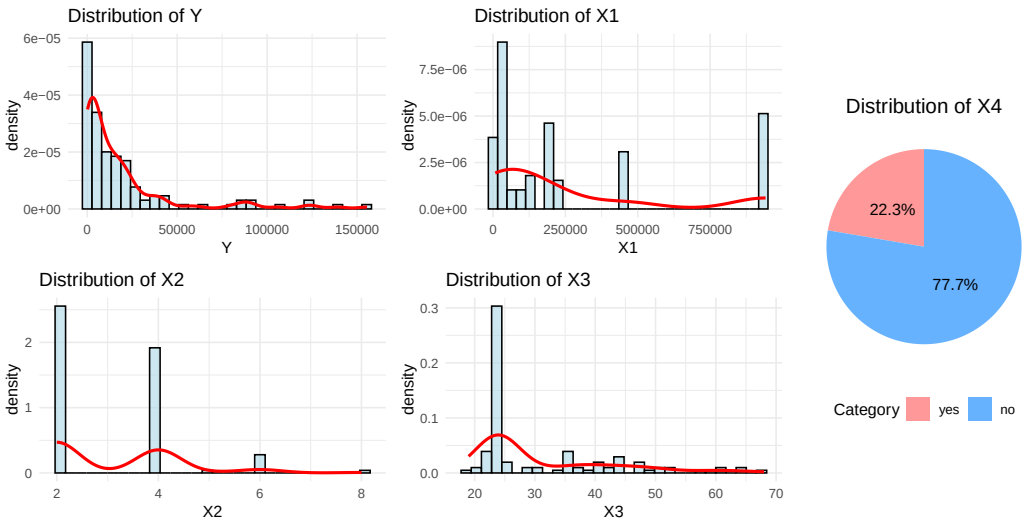

Figure 2: Distribution of the Variables

This finding implies that larger county populations are associated with higher average annual daily traffic, which aligns with real-world expectations where more populated counties tend to have higher traffic volumes.

3.1.2 Residuals Analysis

We draw the residuals plot and QQ-plot in Figure 5 to provide further insights into the model's adequacy.

Residuals Plot The residual plot in Figure 5 shows a generally even distribution of residuals around zero, though there are some signs of non-constant variance. This pattern might indicate that variability in traffic increases with larger predicted values, potentially due to factors such as urban infrastructure or public transport usage that could influence traffic in larger counties.

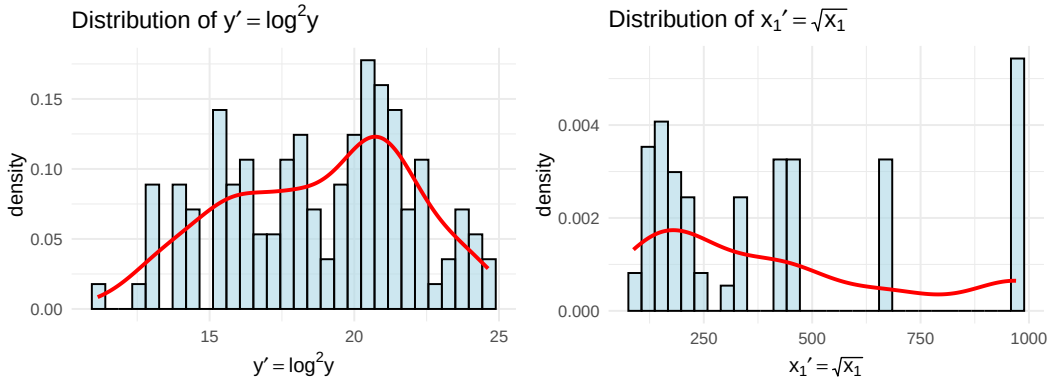

Figure 3: Distribution of variables after transformation

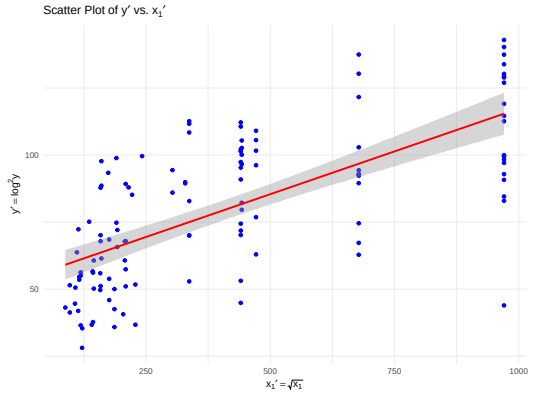

Figure 4: Scatter plot of the model $y' = \beta_0^{(1)} + \beta_1^{(1)} x_1' + \epsilon$

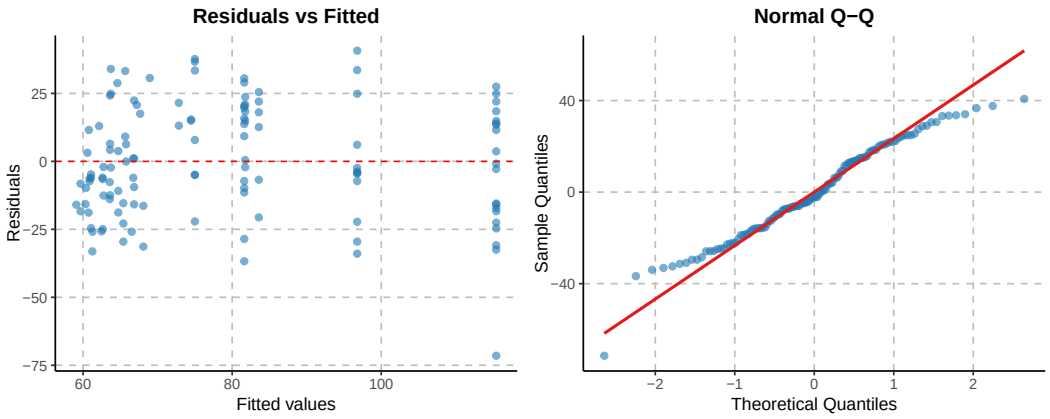

Figure 5: Residuals of the model $\hat{y'} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x_1'$

Columns	Meaning	Type
y	Annual average daily traffic (AADT)	Continuous
x_1	Population of county in which road section in located	Discrete
x_2	Number of lanes in road section	Discrete
x_3	width of road section (in feet)	Continuous
x_4	Whether there is control of access to road section*	Binary

Table 1: Overview of the Variables

QQ-Plot The QQ-plot of residuals shows that the residuals mostly follow a normal distribution, with minor deviations at the tails, suggesting that while the model performs well overall, there might be outliers or specific population segments where the prediction is less accurate.

3.1.3 Residual Statistics

ANOVA Table We calculate the ANOVA table in Table 2.

Source	df	SS	MS	$\mathcal{F}$
Regression	1	$\mathrm{SSR} = 43471$	$\mathrm{MS}_{\mathrm{Reg}} = \mathrm{SSR} = 43471$	$F = \frac{\mathrm{MS}_{\mathrm{Reg}}}{s^2} \approx 102.75$
Residual	$n-2 = 119$	$\mathrm{SSE} = 53916$	$s^2 = \frac{\mathrm{SSE}}{n-2} = 423$	
Total	$n-1 = 120$	$S_{yy} = 97387$		

Table 2: ANOVA table for the population with AADT.

Since F value is in a $\mathcal{F}$ distribution:

$$F = \frac{\mathrm{MS}_{\mathrm{Reg}}}{s^2} \sim \mathcal{F}_{1,n-2} \tag{5}$$

We can build $\mathcal{H}_0$: $\beta_1 = 0$ with $\mathcal{H}_1$: $\beta_1 \neq 0$. And do $\mathcal{F}$-test by calculating the p value:

$$p = \mathbb{P}\left(\mathcal{F}_{1,199} > F\right) < 2 \times 10^{-16} \tag{6}$$

The p value is small enough for us to reject $\mathcal{H}_0$ and conclude that the population actually has a influence to AADT.

$\mathcal{R}^2$ Statistic The $\mathcal{R}^2$ statistic is calculated as

$$\mathcal{R}^2 = \frac{\mathrm{SSR}}{S_{yy}} \approx 0.4634 \tag{7}$$

This means that while there is some relation between population and AADT, there are other effectors that determine the AADT.

3.2 Number of Lanes v.s. AADT

3.2.1 Model Overview

The scatter plot Figure 6 shows how AADT changes with the number of lanes. Each point represents a specific road section, with:

- The x-axis representing the number of lanes.
- The y-axis representing the AADT data after transformation.

In this plot, we also include a trendline to indicate the general pattern in the data.

A positive slope in the trendline has been observed in the plot, which means that as the number of lanes increases, AADT tends to increase as well. This aligns with common sense: adding lanes generally allows a road to support more vehicles, which increases its traffic capacity.

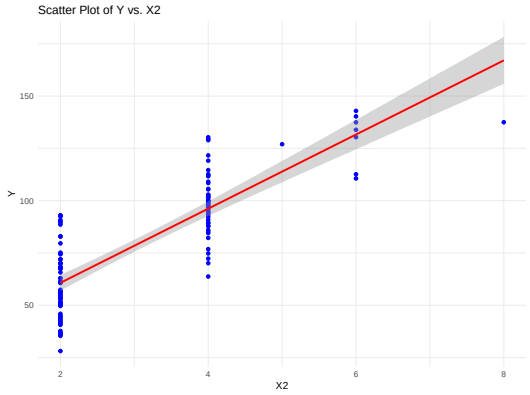

Figure 6: Scatter plot of the model $y' = \beta_0^{(2)} + \beta_1^{(2)} x_2' + \epsilon$

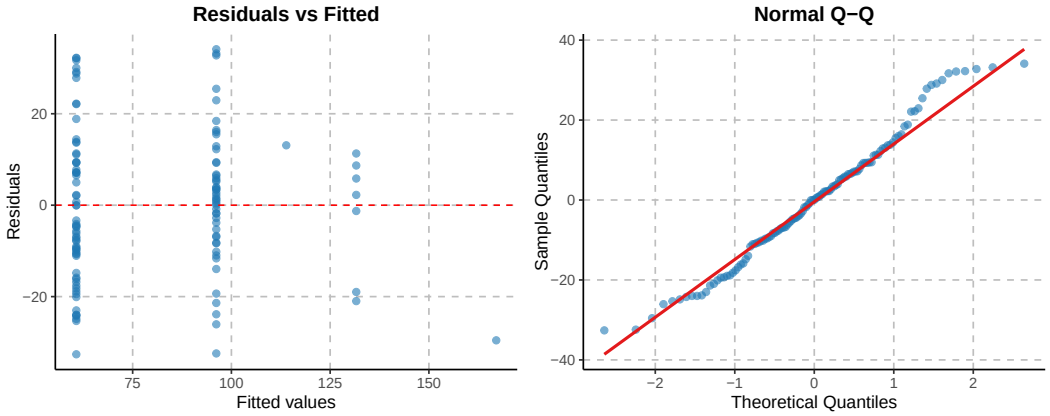

Figure 7: Residuals of the model $\hat{y'} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x_2$

We construct the linear model for the responsible variable y' and predicted variable x_2:

$$y' = \beta_0^{(2)} + \beta_1^{(2)} x_2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_2^2) \tag{8}$$

We fitted the model with the data, shown as the red line in Figure 6:

$$\begin{cases} \hat{\beta}_0^{(2)} \approx 25.304 \\ \hat{\beta}_1^{(2)} \approx 17.718 \end{cases} \tag{9}$$

Intercept When the number of lanes is zero, the model predicts y' of approximately 25.3. While this might not be meaningful practically (since we rarely see roads with zero lanes), it serves as a baseline.

Slope for x_2 For each additional lane, the model predicts an increase of about 17.7 in y'. This means adding lanes has a significant positive effect on traffic capacity.

Significance The very low p-value $\left(< 2 \times 10^{-16}\right)$ for the number of lanes indicates that this relationship is statistically significant, implying that the number of lanes is an important factor in determining AADT.

3.2.2 Residual Analysis

Residuals vs Fitted Plot Figure 7 examines how well the linear model predicts AADT based on the number of lanes. If the residuals (the differences between actual and predicted AADT) are scattered

randomly around the zero line, it indicates that the number of lanes effectively explains variations in AADT. In practical terms, this would mean that road sections with different lane counts generally show predictable changes in traffic volume, and our model captures this relationship well. However, if a pattern appears in the residuals, such as a consistent underestimation or overestimation of AADT for certain lane counts, it might suggest that factors beyond lane count (e.g., location, road type) are also influencing traffic volume, indicating that our model may need additional variables to improve accuracy.

Normal Q-Q Plot Refer to Figure 7 again, it assesses whether the residuals are normally distributed, which is an assumption for linear regression. A normal distribution of residuals suggests that the relationship between lane count and AADT is generally well-captured by the linear model. If the points fall along a straight line in the Q-Q plot, it indicates that the model's errors are random and unbiased, meaning our predictions are reasonably reliable across different lane counts. However, significant deviations from this line might indicate that the relationship between lane count and AADT isn't fully linear or that other factors are affecting traffic volume in ways the model doesn't capture, possibly warranting a more complex or adjusted model.

3.2.3 Residual Statistics

ANOVA Table We calculate the ANOVA table in Table 3.

Source	**df**	**SS**	**MS**	$\mathcal{F}$
Regression	1	$\mathrm{SSR} = 43471$	$\mathrm{MS}_{\mathrm{Reg}} = \mathrm{SSR} = 43471$	$F = \frac{\mathrm{MS}_{\mathrm{Reg}}}{s^2} \approx 102.75$
Residual	$n - 2 = 119$	$\mathrm{SSE} = 53916$	$s^2 = \frac{\mathrm{SSE}}{n-2} = 423$	
Total	$n - 1 = 120$	$S_{yy} = 97387$		

Table 3: ANOVA table for the population with AADT.

According to Equation 5, we can build $\mathcal{H}_0$: $\beta_1 = 0$ with $\mathcal{H}_1$: $\beta_1 \neq 0$. And do $\mathcal{F}$-test by calculating the p value:

$$p = \mathbb{P}\left(\mathcal{F}_{1,199} > F\right) < 2 \times 10^{-16} \tag{10}$$

The p value is small enough for us to reject $\mathcal{H}_0$ and conclude that the population actually has a influence to AADT.

$\mathcal{R}^2$ Statistic The $\mathcal{R}^2$ statistic is calculated as

$$\mathcal{R}^2 = \frac{\mathrm{SSR}}{S_{yy}} \approx 0.4634 \tag{11}$$

This means that while there is some relation between population and AADT, there are other effectors that determine the AADT.

3.2.4 Summary

This analysis shows that the number of lanes has a strong, positive effect on AADT. This makes intuitive sense, as wider roads with more lanes are better suited to handle larger volumes of traffic. The linear model and diagnostic plots suggest that the relationship is well captured by our model, with residuals generally behaving as expected.

In summary, as we add more lanes to a road, we can expect an increase in daily traffic capacity, which matches our common-sense understanding of road infrastructure and traffic flow.

3.3 Road Width v.s. AADT

3.3.1 Model Overview

We build the linear regression model

$$y' = \beta_0^{(3)} + \beta_1^{(3)} x_3 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_3^2) \tag{12}$$

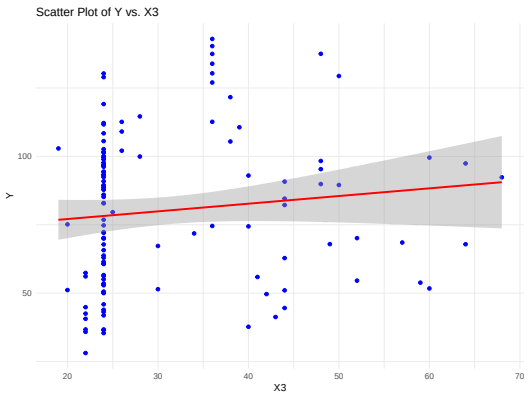

Figure 8: Residuals of the model $\hat{y'} = \hat{\beta}_0^{(3)} + \hat{\beta}_1^{(3)} x_3$

The scatter plot in Figure 8 shows the relationship between Annual Average Daily Traffic (AADT) and the width of road sections (in feet). Each point represents a specific road section, with: The x-axis represents the width of the road section in feet x_3. The y-axis represents the AADT.

In this plot, a trendline has been added to indicate the general pattern in the data. Although the trendline suggests a positive slope, indicating that as the road width increases, AADT tends to increase, this relationship is not statistically significant, as shown by the linear model.

- **Intercept**: 71.49, which suggests that at a hypothetical width of zero, the model would predict an AADT of approximately 71.5, though this does not have practical meaning.
- **Slope for** x_3: 0.28, indicating a predicted increase in AADT of approximately 0.28 for each additional foot in road width. However, this effect is not statistically significant (p-value = 0.207), suggesting that road width does not significantly impact AADT in this model.

3.3.2 Residual Statistics

ANOVA Table We calculate the ANOVA table in Table 4

Source	**df**	**SS**	**MS**	$\mathcal{F}$
Regression	1	$\text{SSR} = 1254$	$\text{MS}_{\text{Reg}} = \text{SSR} = 1254.46$	$F = \frac{\text{MS}_{\text{Reg}}}{s^2} \approx 1.6127$
Residual	$n - 2 = 119$	$\text{SSE} = 92564$	$s^2 = \frac{\text{SSE}}{n-2} = 777.85$	
Total	$n - 1 = 120$	$S_{yy} = 93818$		

Table 4: ANOVA table for the population with AADT.

$\mathcal{R}^2$ Statistic The $\mathcal{R}^2$ statistic is calculated as

$$\mathcal{R}^2 = \frac{\text{SSR}}{S_{yy}} \approx 0.013 \tag{13}$$

implies that road width explains only about 1.3% of the variation in AADT, suggesting that other factors are more influential.

3.3.3 Summary

This analysis suggests that the width of a road section has a weak and statistically insignificant relationship with AADT. While wider roads generally allow for higher traffic volumes, this model does not capture that effect effectively.

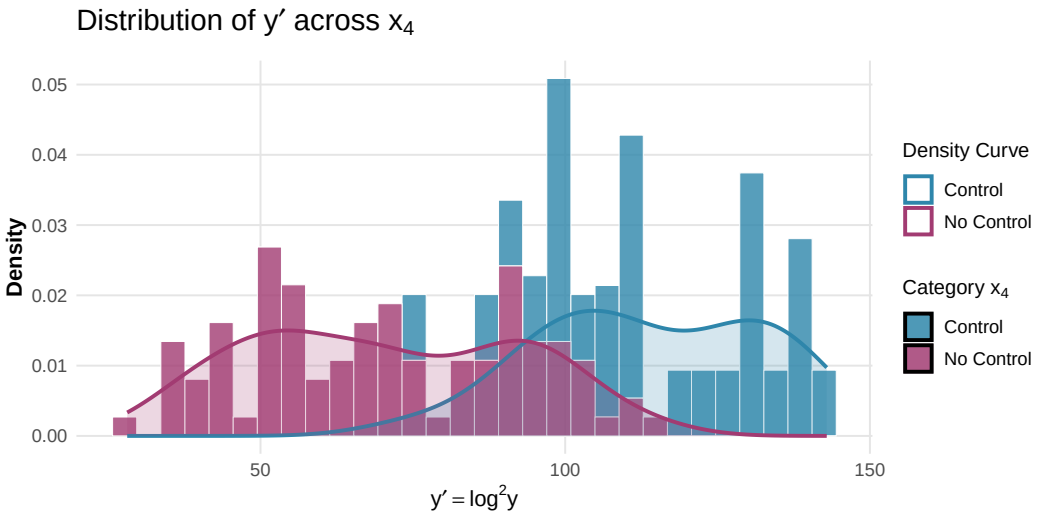

Figure 9: Distribution of AADT over whether to control access.

3.4 Whether to Control Access v.s. AADT

Controlling access to the road section is one of the important methods for improving traffic flow. In this chapter, we explore whether this approach can effectively increase the annual average daily traffic.

We divide the data into two categories: the first category consists of data where control measures are implemented, and the second category consists of data without control measures. The distribution of these 2 categories is shown in Figure 9.

We then establish a linear model:

$$y'_{ij} = \theta_i + \epsilon_{ij}, \quad i \in \{1, 2\}, j \in \{1, \cdots, n_i\}, \quad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_4^2\right) \tag{14}$$

We can fit the model by

$$\begin{cases} \hat{\theta}_1 = \overline{y_{1\cdot}} \approx 114.10 \\ \hat{\theta}_2 = \overline{y_{2\cdot}} \approx 70.48 \end{cases} \tag{15}$$

We are interested in whether these two categories have a significant difference. So we build $\mathcal{H}_0$: $\theta_1 = \theta_2$ versus $\mathcal{H}_1$: $\theta_1 \neq \theta_2$, and we can build the ANOVA table, shown in Table 5.

Source	**df**	**SS**	**MS**	$\mathcal{F}$
Between groups	$k-1=1$	$\mathrm{SST}=39903$	$\mathrm{MST}=\frac{\mathrm{SST}}{k-1}=39903$	$F=\frac{\mathrm{MST}}{\mathrm{MSE}}\approx 88.07$
Within groups	$n-k=119$	$\mathrm{SSE}=53916$	$\mathrm{MSE}=\frac{\mathrm{SSE}}{n-k}=453$	
Total	$n-1=120$	$S_{yy}=93819$		

Table 5: ANOVA table for whether to control access.

As

$$F = \frac{\mathrm{MST}}{\mathrm{MSE}} \sim \mathcal{F}_{1,119} \tag{16}$$

We can calculate p-value by:

$$p = \mathbb{P}\left(\mathcal{F}_{1,199} > F\right) \approx 5.37 \times 10^{-16} \tag{17}$$

This is a quite small number, so we can conclude that there is a significant difference by controlling the access to the road section.

4 Multiple Linear Regression

4.1 Multiple Linear Regression with Full Dataset

As x_4 is a classification variable, we create a *dummy variable*

$$x_4' = 2 - x_4 \tag{18}$$

So that $x_4 = 1$ means we select the first class (control) and $x_4 = 0$ means we select the second class.

So we can represent data by:

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ x_1' \\ x_2 \\ x_3 \\ x_4' \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{x_1} \\ x_2 \\ x_3 \\ 2 - x_4 \end{bmatrix}, \quad X = \begin{bmatrix} \boldsymbol{x_1}^\top \\ \boldsymbol{x_2}^\top \\ \vdots \\ \boldsymbol{x_n}^\top \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1}' & \dots & x_{1,4}' \\ 1 & x_{2,1}' & \dots & x_{2,4}' \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1}' & \dots & x_{n,4}' \end{bmatrix} \tag{19}$$

And build the linear regression model for the full dataset:

$$\log^2 y = \beta_0 + \beta_1 \sqrt{x_1} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (2 - x_4) + \epsilon = \boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon \tag{20}$$

We fit the model by

$$\hat{\beta} = (X^\top X)^{-1} X y' \approx \begin{bmatrix} 27.26 \\ 0.036 \\ 11.07 \\ 0.019 \\ -13.24 \end{bmatrix} \tag{21}$$

4.2 Adjusted Model

Figure 10 shows the model prediction and residuals. We can find that the σ of residuals are dependent with y', so we need to adjust the transformation to make σ a constant.

We succeed in the goal by doing the transformation

$$y' = \log^4 y \tag{22}$$

So the model becomes

$$\log^4 y = \beta_0 + \beta_1 \sqrt{x_1} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (2 - x_4) + \epsilon \tag{23}$$

Figure 11 shows residuals after adjustment. We can see that the σ is mostly the same for each y.

4.3 Adequacy Checking

To check the adequacy of the model, we build an ANOVA / ANCOVA table for each variable. Shown in Table 6. We can find that x_1', x_2, x_4' have information with y, while x_3 can be removed as it have a great p-value.

Source	**DF**	**RSS**	$\mathcal{F}$	p-value
Reduce x_1	117	679542372	99.815	$< 2.2 \times 10^{-16}$
Reduce x_2	117	680342805	100.07	$< 2.2 \times 10^{-16}$
Reduce x_3	117	365429973	0.0562	0.8129
Reduce x_4	117	459976649	30.083	2.457×10^{-7}
Full Model	116	365252858		

Table 6: ANOVA / ANCOVA table for adequacy checking.

So we decide to reduce x_3 from our model.

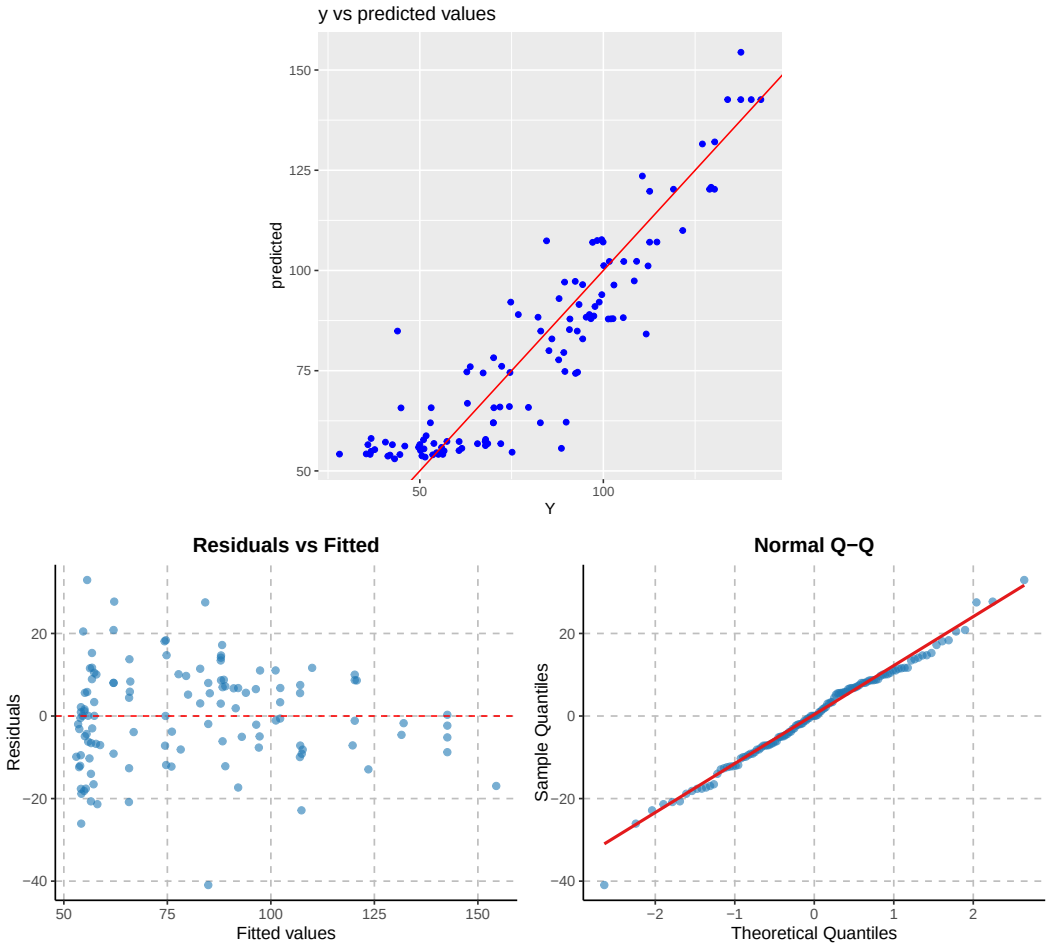

Figure 10: Regression model and residuals for the full model

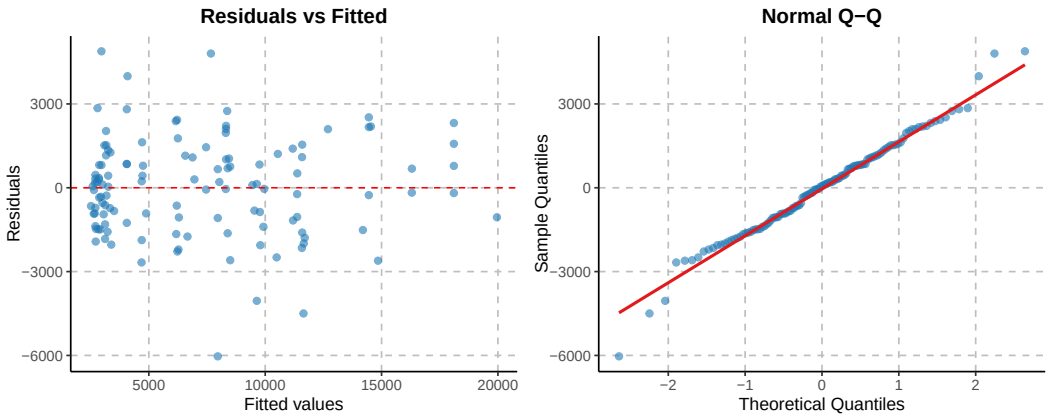

Figure 11: Regression model and residuals for the full model

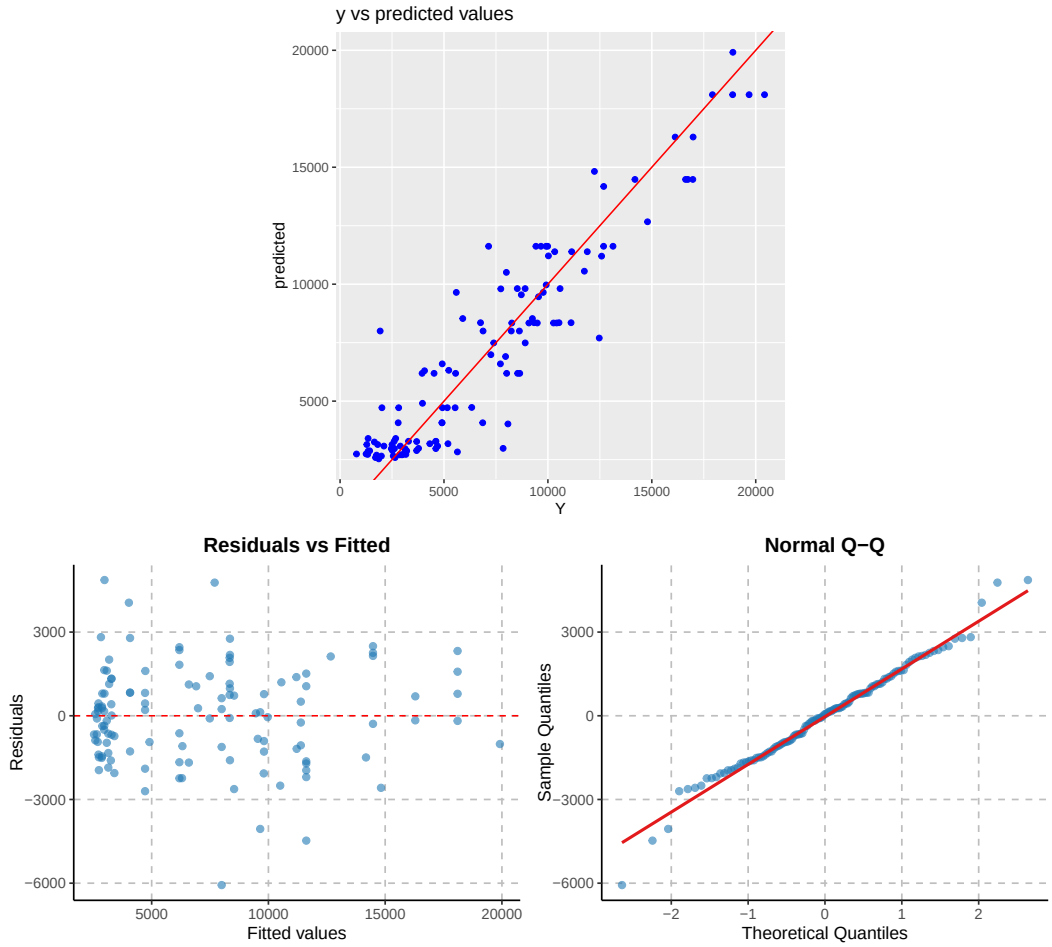

Figure 12: Predictions and residuals for the finalised model

4.4 Model Finalization

So our final multiple linear regression model is

$$\log^4 y = \beta_0 + \beta_1\sqrt{x_1} + \beta_2 x_2 + \beta_4(2 - x_4) + \epsilon \tag{24}$$

Figure 12 shows the predictions and residuals for the finalised model. We can observed that most of the residuals lie in the range $(-3000, 3000)$.

5 Prediction

5.1 Point Estimation

From Equation 24, we can get:

$$\hat{y} = \exp\sqrt[4]{\hat{\beta}_0 + \hat{\beta}_1\sqrt{x_1} + \hat{\beta}_2 x_2 + \hat{\beta}_4(2 - x_4)} \tag{25}$$

Substituting by

$$\begin{cases} x1 = 50000 \\ x2 = 3 \\ x3 = 60 \\ x4 = 2 \end{cases} \tag{26}$$

We can get

$$\hat{y} \approx 27379.73 \tag{27}$$

5.2 Interval Estimation

We can get the 90% confidence interval of $\hat{y}'$:

$$(l', r') = (9329.20, 12468.89) \tag{28}$$

So after doing the inverse function of Equation 22:

$$y = \exp \sqrt[4]{y'} \tag{29}$$

We can get the 90% confidence interval:

$$(l, r) = (18544.14, 38836.96) \tag{30}$$

6 Conclusion

Our single variable regression analysis showed that population size and the number of lanes significantly predict AADT, consistent with expectations that larger populations and additional lanes increase traffic capacity. However, road width was found to have a weak and statistically insignificant relationship with AADT. Access control demonstrated a substantial impact on traffic, as shown through a two-group comparison where controlled access increased AADT.

In the multiple linear regression model, we incorporated these significant factors, refining our understanding of how they collectively influence traffic patterns. This comprehensive approach provides a clearer picture of the primary determinants of traffic volume and informs infrastructure planning decisions.

And finally, we provide our prediction in Equation 27 and 30.

A Code Snips

The full code can be accessed in https://pufanyi.github.io/MH3510-Project/, here we listed some of the code snips.

A.1 Data Loading

```r
file <- "../assets/aadt.txt"
data_raw <- read.table(file, col.names = columns)
data_ori <- data_raw[, c("Y", "X1", "X2", "X3", "X4")]
```

A.2 Data Transformation

```r
y_prime <- log(data$Y)^2
x1_prime <- sqrt(data$X1)
# ... Analysis code here
data$Y <- log(data$Y)^2
```

A.3 Single Linear Regression and Analysis

We take x_1 as an example

```r
slr_X1 <- lm(y_prime ~ x1_prime, data = data)
summary(slr_X1)
anova(slr_X1)
```

A.4 Multiple Linear Regression

```r
data$Y <- data$Y^2 # Further transformation
data$X4 <- 2 - data$X4 # Dummy variable

mlr <- lm(Y ~ X1 + X2 + X3 + X4, data = data)
summary(mlr)

# check adequecy
eliminate_x1_mlr <- lm(Y ~ X2 + X3 + X4, data = data)
anova(eliminate_x1_mlr, mlr)
# ... check other variables

# reduce model
mlr <- eliminate_x3_mlr
```

A.5 Estimation

```r
x1 <- 50000
x2 <- 3
x3 <- 60
x4 <- 2

# transformation
x1 <- sqrt(x1)
x4 <- 2 - x4

# prediction
input <- data.frame(X1 = x1, X2 = x2, X3 = x3, X4 = x4)
```

```r
y_hat <- predict(mlr, input)

# transform back
y_hat <- exp(y_hat^(1 / 4))

# interval estimation
alpha <- 0.1
interval <- predict(mlr, input, interval = "confidence",
                    level = 1 - alpha)

# transform back
l <- exp(interval[2]^(1 / 4))
r <- exp(interval[3]^(1 / 4))
print(paste("(", l, ", ", r, ")"))
```