Final Project

STAT 4620/5620 Winter 2025

Final Project

STAT 4620/5620 Winter 2025

Title:

Understanding Commuting in Halifax: A Study on Work Duration and Mode Choice

Group Members:

Member 1: Md. Rifat Hossain Bhuiyan

Member 2: Niaz Mahmud

Member 3: Azam Khan

GitHub Repository:

https://github.com/Rifat1633/STAT-Final.git

Abstract

This research investigates how sociodemographic, built-environment, and activity-time use factors influence daily work duration and commute mode choice. Using data from the 2022 Halifax Travel Activity (HaliTRAC) Survey, various statistical models, including Ordinary Least Squares (OLS) regression, Generalized Linear Models (GLM), Iteratively Reweighted Least Squares (IRLS) regression, and Multinomial Logistic Regression (MNL), were applied to explore the effects of factors such as age, gender, income, vehicle ownership, and commute distance. The findings reveal that higher income, vehicle ownership, and age are associated with longer work durations and increased use of personal vehicles, while lower income and lack of a driving license lead to greater reliance on public transit. Additionally, non-work activities like chores and recreation are negatively correlated with work hours. The built environment, including proximity to central business districts and other essential services, also influences transportation choices. These results highlight the importance of considering sociodemographic and environmental factors in transportation planning and policy, suggesting that promoting flexible work hours, improving public transit access, and enhancing urban infrastructure can lead to more efficient and equitable transportation systems.

Keywords: Work duration, mode choice, regression model, socio-economic factors.

1. Introduction

Understanding the factors influencing daily work duration and transportation mode choice is crucial in the context of urban transportation planning and policy formulation. As cities grow and become more complex, the demand for efficient transportation systems that accommodate diverse travel behaviors and work patterns becomes even more pressing. Work trips are central to urban transportation studies, as they often determine the characteristics of peak and off-peak traffic flows, influencing the overall design and efficiency of urban transport networks. By analyzing the role of sociodemographic factors, commute distances, and built environment characteristics, this research seeks to unravel the intricate relationships that shape work duration and transportation mode choice.

Sociodemographic factors, including age, gender, income, and employment status, play a significant role in determining how individuals allocate their time to work and other activities. Age, for instance, can influence an individual's work habits, with younger individuals potentially engaging in part-time work or varying job patterns, while older individuals might have more stable, full-time employment. Gender disparities in work hours have also been studied, with research suggesting that women often have different work-time patterns due to household responsibilities (Hochschild & Machung, 2012). Income level further impacts work duration, as higher-income individuals often have more flexible working hours, whereas lower-income individuals might work longer hours due to economic necessity (Wachter, 2020). Employment status, such as whether one is employed full-time, part-time, or self-employed, naturally dictates the number of hours spent on work-related activities, influencing the overall work duration (Fan et al., 2015).

Understanding these factors is essential for transportation planners as they directly affect demand for travel during peak periods. For example, those with longer working hours or specific schedules (e.g., shift workers) may contribute to traffic congestion during non-traditional work hours, necessitating different transportation infrastructure to address these needs (Graham & Glaister, 2004).

Transportation mode choice on the other hand is influenced by a variety of factors, including sociodemographic characteristics, the distance to work, workplace location, and the surrounding built environment. Commute distance is a well-established determinant of mode choice, with longer distances often leading to a preference for private vehicles, while shorter distances may encourage the use of public transport or active modes like walking and cycling (Givoni & Rietveld, 2014). The built environment, including the availability of infrastructure such as public transit, bike lanes, and pedestrian-friendly areas, also plays a pivotal role in shaping transportation behavior. In urban areas where public transportation is easily accessible, individuals may be more inclined to use transit as a mode of commuting. On the other hand, in car-dependent cities, individuals may have limited public transportation options, making personal vehicles the more practical choice (Cervero & Kockelman, 1997).

Sociodemographic factors like income and employment status influence these choices as well. Higher-income individuals are more likely to own private vehicles, while lower-income individuals might rely more on public transit due to financial constraints. Additionally, the workplace's location relative to residential areas can affect mode choice. For example, individuals who work in central business districts (CBDs) with robust public transit systems may prefer transit, whereas those working in suburban or rural areas with limited public transport may opt for private vehicles (Boarnet & Crane, 2001).

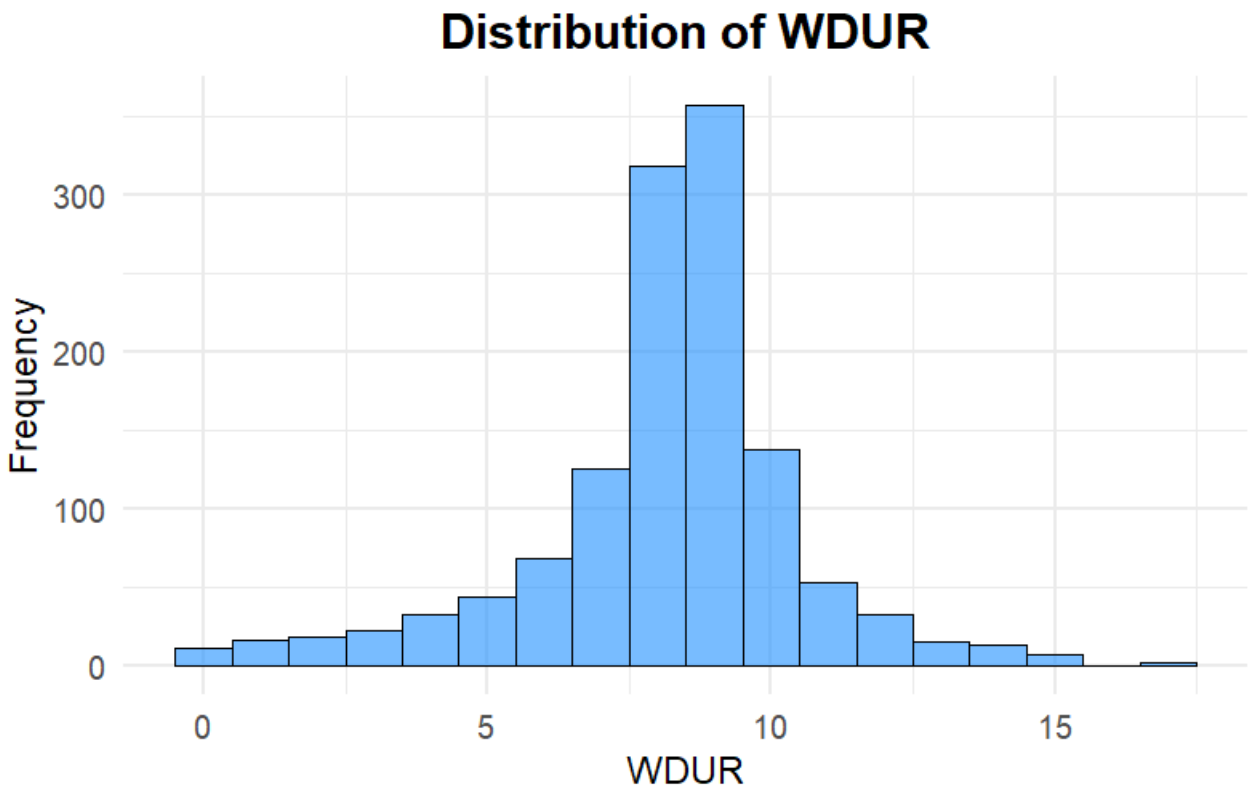
From a policy perspective, understanding the relationships between work duration, mode choice, and sociodemographic factors helps to create more equitable and efficient transportation systems. This research builds on prior literature to explore the relationship between sociodemographic, activity-related, and built environment factors and daily work duration. Additionally, the study examines how factors such as income, commute distance, and the built environment influence commute mode choice. The goal is to answer two key research questions: how sociodemographic factors affect daily work hours, and how various factors impact the choice of commute mode.

2. Data Description

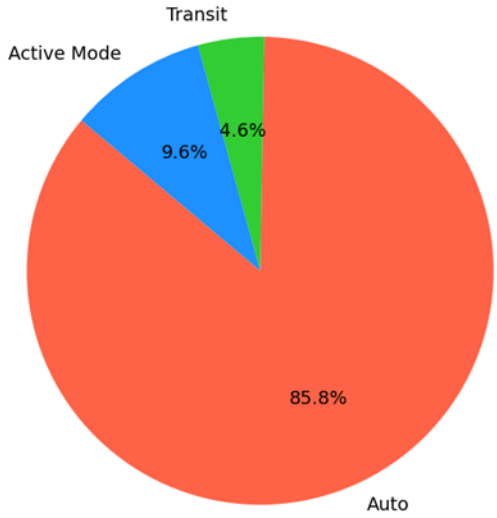
This research uses data from the 2022 Halifax Travel Activity (HaliTRAC) Survey, collected by the Dalhousie Transportation Collaboratory (DalTRAC) in partnership with the Halifax Regional Municipality (HRM). The survey includes detailed information on sociodemographic factors, travel behaviors, employment status, work duration, commuting patterns, and mode choice. undefinedThe dataset comprises 3,731 households, covering 5,095 individuals, with 14,327 recorded activities over a 24-hour period. Key variables include age, gender, income, employment status, work duration, vehicle ownership, commute distance, and work arrangements.

This research primarily utilizes responses from the workers who allocate a significant portion of their time to work-related activities. A total of 1509 independent responses were utilized for the work duration model and 944 responses were utilized for the commute mode choice model. The raw data as well as the data interpretation files have been made available in the following github directory: Data Directory

Figure titled **Distribution of WDUR** demonstrates the frequency of work duration across the responses, which varies between 0 to 15 hours per day.



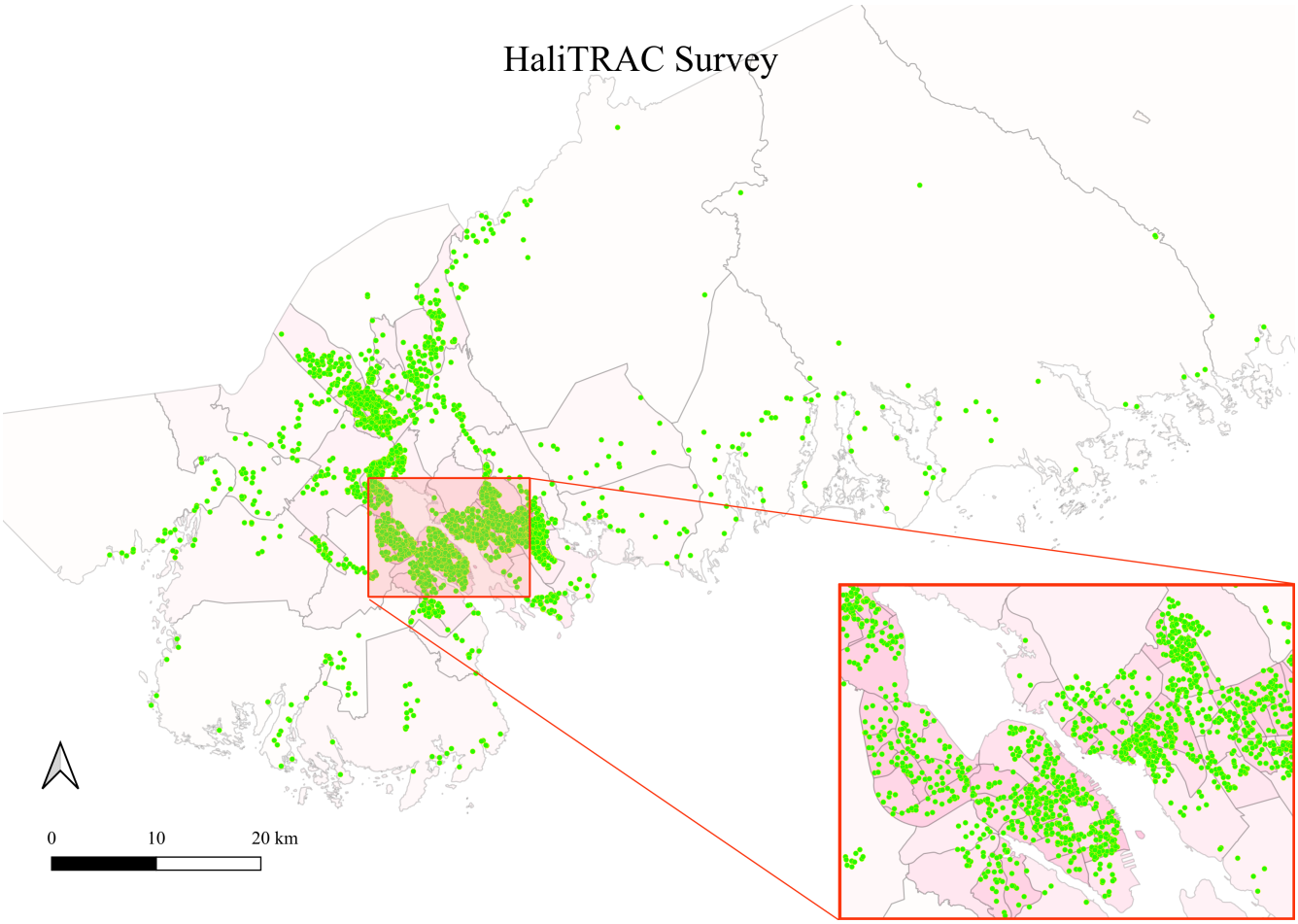
On the other hand, the figure below demonstrates the distribution of **Commute Mode** observed in the dataset.



The commute mode choice shows very high share of auto mode (personal vehicle), followed by active modes (walking/cycling), and transit.

The survey data used in this research incorporates a comprehensive set of predictor variables spanning **demographic, built environment, and activity time-use factors** to explore work duration and commute mode choice behavior. **Demographic factors** include **Age, Gender, Employment status, Education level, Income level, Driving License, Transit Pass,**

Vehicle ownership, Home ownership, and presence of Children. **Built environment** variables cover accessibility to **Nearest restaurants, malls, schools, CBD, grocery stores, and the Land use index**. **Activity time-use** predictors account for time spent on **In-Home activities, Chores, Personal business, Recreation, Other activities, and the number of Trips made**. These variables collectively provide valuable insights into the factors influencing both work duration and decisions regarding the choice of commute modes.



The figure above demonstrates the locations of the survey respondents. Using **Near Analysis** in ArcGIS, various built environment features such as the nearest restaurants, schools, malls, and grocery stores to the respondents' dwellings were calculated. The data obtained from this GIS analysis was then used to create several built environment variables, which were incorporated into the work duration and mode choice models.

The Table presented below demonstrates the descriptive statistics of the survey data:

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

Table 1: Descriptive Statistics of the Data

Variable	Description	Mean (hr)	%
Work duration	Daily work duration in hours	3.99	
Commute Mode	Auto	806	85.84
	Transit	43	4.58
	Active Mode	90	9.58
Demographic variables	Attributes	Frequency	%
Age group	<25	143	9.5
	25-35	237	15.7
	35-50	433	28.7
	50-65	466	30.9
	>65	230	15.2
Gender	Male	776	51.4
	Female	733	48.6
Employment	Full-time	1087	72
	Part-time	84	5.6
	Retired	240	15.9
	Student	83	5.5
	Unemployed	15	1
Education Level	Low	251	16.6
	Medium	361	23.9
	High	895	59.3
Income Level	<25K	30	2
	25-50K	115	7.6
	50-100K	460	30.5
	>100K	745	49.4
Driving License	Yes	1385	91.8
	No	124	8.2
Transit Pass	Yes	127	8.4
	No	1382	91.6
Vehicle in household	No Vehicle	78	5.2
	One Vehicle	486	32.2
	Two Vehicle	700	46.4
	Three or more vehicle	245	16.2
Home Ownership	Owner	1198	79.4
	Renter	311	20.6
Child present in household	Yes	696	46.1
	No	813	53.9

Variable	Description	Mean (hr)	%
Nearest restaurant	Restaurant closest to the household in kilometers	0.2	0.29
Nearest mall	Mall closest to the household in kilometers	4.75	7.47
Nearest School	School closest to the household in kilometers	0.87	1.16
Nearest CBD	Nearest Central Business District in kilometers	6.1	8.49
Nearest Grocery	Nearest grocery shop in kilometers	1.53	2.29
Land use index	An index between 0 to 1 (high value indicates mixed land use)	0.19	0.11
Activity time-use variables	Description	Mean (hr)	Std. Dev.
Home	In-home activities such as home leisure and sleep	13.92	2.93
Chore	Eating, meal preparation, cleaning, childcare, etc.	0.7	1.58
Personal business	Fitness activities, medical, and banking	0.12	0.57
Recreation	Eating out, meeting with friends, etc.	0.49	1.13
Other	All non-routine, non-traditional activities	0.08	0.54
Trips	Daily total number of trips	3.8	1.96

3. Methods

A series of statistical models were developed to assess the impact of various socioeconomic, demographic, built-environment, and activity-time use factors on daily work duration and commute mode choice. The methodological framework for these models is outlined below.

3.1 Ordinary Least Square Regression Model

The **Ordinary Least Squares (OLS) Regression Model** is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The model assumes that the relationship between the variables is linear. In the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Where,

Y is the dependent variable (e.g., daily work duration).

β_0 is the intercept.

$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients for the predictor variables (e.g., socio-economic, demographic factors).

$X_1, X_2, \ldots, X_p$ are the independent variables (e.g., age, gender, education).

ϵ is the error term or residual.

3.2 Generalized Linear Model (GLM)

The **Generalized Linear Model (GLM)** is an extension of the traditional linear regression model that allows for the dependent variable to follow a distribution other than the normal distribution.

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where,

Y is the dependent variable.

$g(\mu)$ is the link function that relates the mean of the dependent variable to the linear predictors.

$X_1, X_2, \ldots, X_p$ are the independent variables.

β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the independent variables.

GLM allows for flexibility in modeling different types of outcome variables by changing the link function and the distribution of the dependent variable.

3.3 Iteratively Reweighted Least Squares (IRLS)

Iteratively Reweighted Least Squares (IRLS) is an iterative algorithm used for fitting generalized linear models (GLMs), especially when the model includes a non-normal distribution or when using robust regression methods. IRLS is particularly useful for handling cases with heteroscedasticity or when the data contain outliers. The algorithm iteratively reweights the data to minimize the residual sum of squares and improve model fit by adjusting weights at each iteration. The process continues until convergence criteria are met, i.e., when changes in the coefficients are sufficiently small.

$$\beta^{(t+1)} = (\mathbf{X}^T W^{(t)} \mathbf{X})^{-1} \mathbf{X}^T W^{(t)} \mathbf{y}$$

$\beta^{(t+1)}$: The coefficients at the next iteration of IRLS.

$\mathbf{X}$: The matrix of predictor variables (independent variables).

$W^{(t)}$: The diagonal weight matrix at iteration t, where the weights are updated based on the residuals from the previous iteration.

$\mathbf{y}$: The vector of observed values (dependent variable).

$\mathbf{X}^T$: The transpose of the matrix $\mathbf{X}$.

In each iteration, the model is updated with new weights based on the residuals from the previous iteration. The weights are typically computed using a robust function such as Huber's loss or the Tukey's bisquare. The process repeats until the coefficients converge, providing a robust solution to the regression problem.

3.4 Multinomial Logistic Regression (MNL)

Multinomial Logistic Regression (MNL) is an extension of binary logistic regression to handle situations where the dependent variable has more than two categories. MNL models the relationship between a set of independent variables and a categorical dependent variable with more than two levels. It estimates the probabilities of each category relative to a reference category by fitting multiple binary logistic regressions, one for each comparison between the reference category and the other categories.

$$\log\left(\frac{P(Y=j)}{P(Y=0)}\right) = \beta_0^j + \beta_1^j X_1 + \beta_2^j X_2 + \cdots + \beta_p^j X_p \quad \text{for} \quad j = 1, 2, \ldots, J-1$$

$P(Y=j)$: The probability of observing category j of the dependent variable.

$P(Y=0)$: The probability of observing the reference category (usually the first category).

β_0^j: The intercept term for the j-**th** category.

$\beta_1^j, \beta_2^j, \ldots, \beta_p^j$: The coefficients for the independent variables $X_1, X_2, \ldots, X_p$ for the j-**th** category.

$X_1, X_2, \ldots, X_p$: The predictor variables (independent variables).

j: The index for the categories (other than the reference category).

The model compares each category to the reference category (usually the first one) using a set of linear equations. The log-odds of each category relative to the reference category are modeled as a linear function of the predictors.

4. Analysis

This data analysis phase of the research project aims to address the two research questions outlined in the introduction:

1. How do sociodemographic factors (such as age, gender, income, and car ownership) influence daily work hours?
2. How do factors like income, commute distance, and the built environment affect commute mode choice?

To answer these questions, several statistical models have been developed. The model formulation, R code, outputs, and fit statistics are presented in the following sections. For the first research question, we develop three models: an Ordinary Least Squares (OLS) Regression, a Generalized Linear Model (GLM), and an Iteratively Reweighted Least Squares (IRLS) Regression. In these models, daily work duration (in hours) is the response variable, estimated based on a range of sociodemographic and built-environment factors as predictor variables.

To address the second research question, a Multinomial Logistic Regression (MNL) model is employed to explore how demographic and built-environment factors influence the choice of commute mode, including auto, transit, and active modes (e.g., walking or cycling).

4.1 Work Duration (OLS Regression Model)

This model dddddd

```r
# Load necessary libraries
library(readxl)
library(dplyr)
library(car)

# Load the data
data <- read_excel("C:/Onedrive_Sync/OneDrive - Dalhousie University/PhD Dal/Coursework/Da

# Convert categorical variables to factors
data <- data %>%
  mutate(across(c(Gender, AGEGROUP, EDULEVEL, EMP, INCOMELEVEL, LICENSE, TPASS,
                  VEHNUMLVL, BICYCLELVL, HOMEOWNER, LOCYRLVL, HHMEMLVL, HHHASCHLD,
                  HOMELOC, BUS5KM), as.factor))

# Handle missing values
data <- na.omit(data)

# Fit initial model
full_model <- lm(WDURHR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR
                  + Gender + AGEGROUP +
                  EDULEVEL + EMP + INCOMELEVEL + LICENSE + TPASS
                  + VEHNUMLVL + BICYCLELVL +
                  HOMEOWNER + LOCYRLVL + HHMEMLVL + HHHASCHLD
                  + TRIPS + HOMELOC + BUS5KM +
                  NEARMALLKM + NEARSCHKM + NEARGROCKM
                  + NEARCBDKM + NEARRESTAUKM + LANDUSE,
                data = data)

# Perform stepwise regression
stepwise_model <- step(full_model, direction = "both", trace = 0)

# Display final model results
summary(stepwise_model)
```

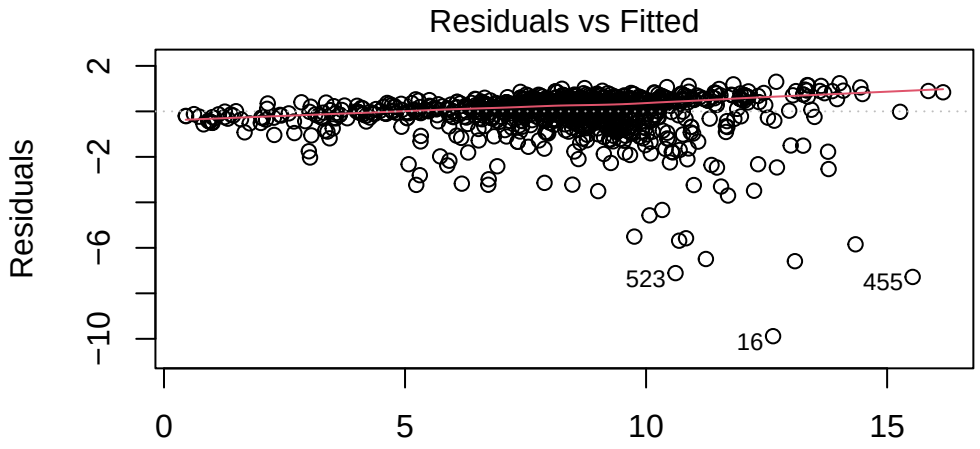
```
Call:
lm(formula = WDURHR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR +
    EDULEVEL + TPASS + HOMEOWNER + TRIPS + NEARMALLKM + NEARCBDKM,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8839 -0.1360  0.2336  0.4119  1.3010
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.516234   0.185114 121.634  < 2e-16 ***
CHOREHR     -0.938361   0.018209 -51.533  < 2e-16 ***
RECHR       -0.885924   0.023887 -37.089  < 2e-16 ***
OTHERHR     -0.905576   0.057417 -15.772  < 2e-16 ***
HOMEHR      -0.913970   0.010606 -86.178  < 2e-16 ***
PBHR        -0.908433   0.041687 -21.792  < 2e-16 ***
EDULEVEL2   -0.120777   0.080866  -1.494 0.135543
EDULEVEL3    0.002501   0.070618   0.035 0.971754
TPASS1      -0.299757   0.086501  -3.465 0.000547 ***
HOMEOWNER2   0.117099   0.062425   1.876 0.060910 .
TRIPS       -0.052959   0.013470  -3.932 8.89e-05 ***
NEARMALLKM  -0.008621   0.006091  -1.415 0.157186
NEARCBDKM    0.008245   0.005519   1.494 0.135469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8812 on 1261 degrees of freedom
Multiple R-squared:  0.863, Adjusted R-squared:  0.8617
F-statistic: 661.9 on 12 and 1261 DF,  p-value: < 2.2e-16
```

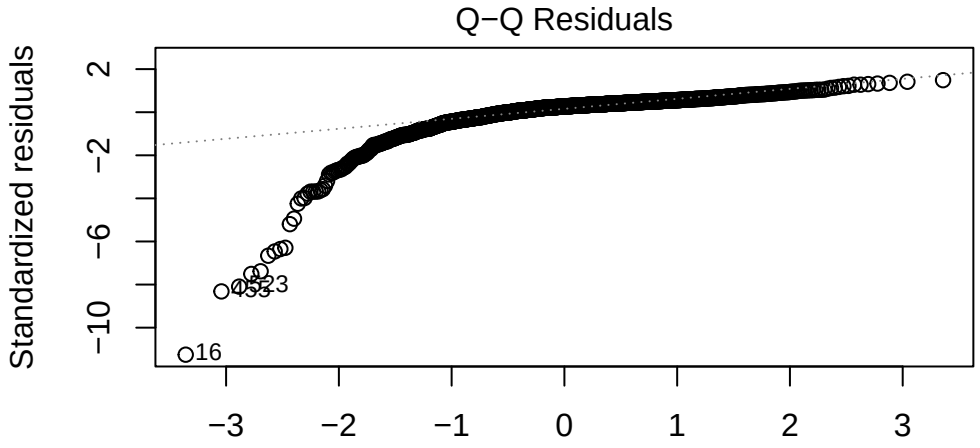
```r
# Check multicollinearity
vif(stepwise_model)
```

```
               GVIF Df GVIF^(1/(2*Df))
CHOREHR    1.338913  1        1.157114
RECHR      1.192254  1        1.091904
OTHERHR    1.031572  1        1.015663
HOMEHR     1.592969  1        1.262129
PBHR       1.018533  1        1.009224
EDULEVEL   1.051277  2        1.012580
TPASS      1.031732  1        1.015742
HOMEOWNER  1.044447  1        1.021982
TRIPS      1.089272  1        1.043682
NEARMALLKM 3.816934  1        1.953698
NEARCBDKM  3.886881  1        1.971517
```

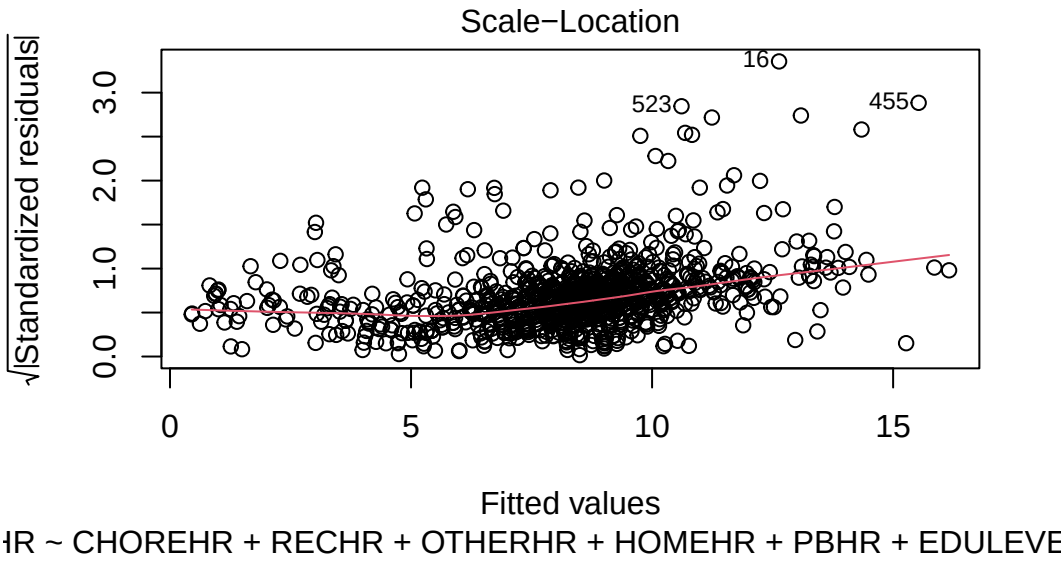
```r
# Diagnostic plots
par(mfrow = c(1, 1))
plot(stepwise_model)
```

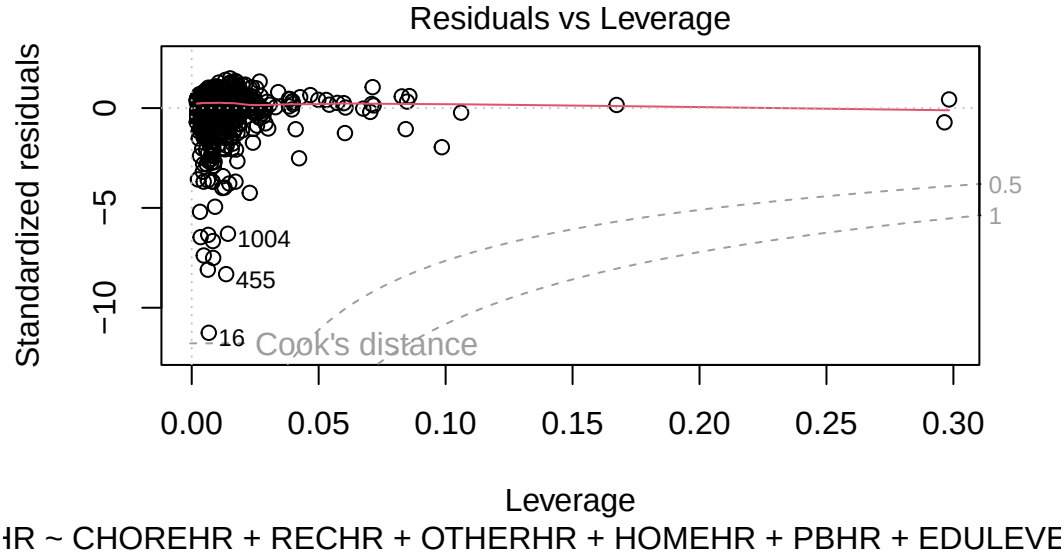


HR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR + EDULEVE



HR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR + EDULEVE






```
par(mfrow = c(1, 1))
```

4.1.1 OLS Model Fit Statistics Interpretation

The model output represents a linear regression analysis of the dependent variable `WDURHR` (work duration in hours) on several predictor variables.

R-Squared and Adjusted R-Squared: The multiple R-squared value of 0.863, and adjusted R-squared value of 0.861 suggest that over 85% of the variance of the dependent variable is explained by the independent variables included in the model.

F-statistic: The F-statistic tests the overall significance of the model. The model shows high F-statistic value (661.9) with a very low p-value, indicating that the model is highly significant and some of the predictors are meaningful in explaining the variability in work duration.

Residual vs. Fitted: This plot shows that while most of the residuals are randomly scattered around the horizontal line at zero, suggesting no major issue with non-linearity or heteroscedasticity. However, there is slight curvature and clustering of residuals at the upper end of the fitted values, hinting minor issue with heteroscedasticity.

Q-Q Plot: This plot tests whether the distribution of the residuals is normal or not. The figure shows slight deviation from normalcy at the lower tail, suggesting some skewness.

Scale-Location Plot: The funnel shape in the scale-location plot indicates mild heteroscedasticity at the higher values of the fitted data.

Residual vs Leverage Plot: From the plot, points 16, 455 etc. seem lie far from the horizontal line and treating them as outliers may increase the overall accuracy of the mode.

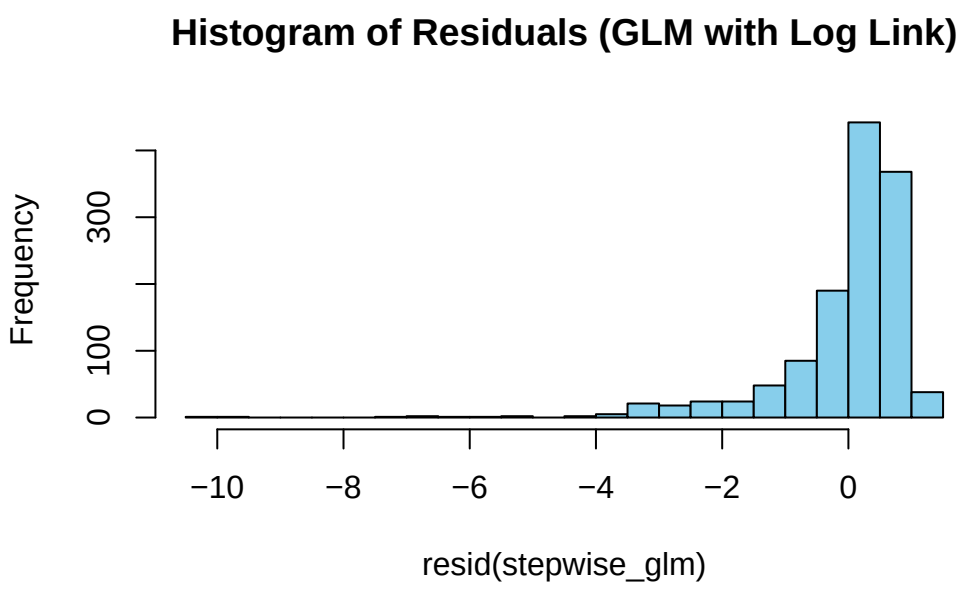
4.2 Work Duration (GLM Model)

Since the OLS regression model diagnosis identified mild heteroscedasticity, and non-normal distribution of the residuals, a Generalized Linear Model (GLM) with log-link was tested. The advantages GLM model has over a LM model are-

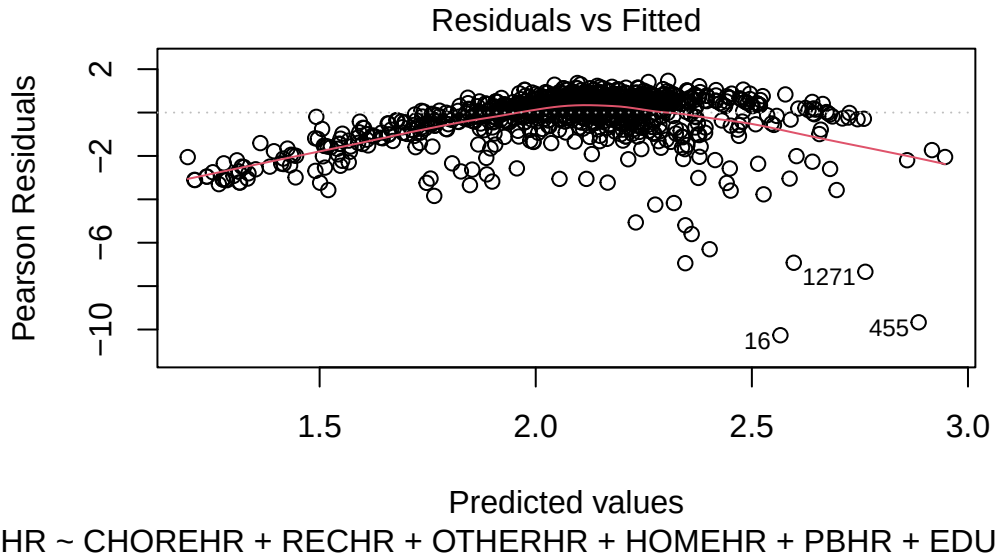
- **Handling Non-Normality**: A GLM with a log link is less sensitive to non-normality of the residuals compared to linear regression.
- **Ensuring Positive Predictions**: Unlike linear regression, which can predict negative values for a strictly positive dependent variable, the GLM with a log link ensures that all predictions are positive, which is applicable to this case.
- **Reducing Heteroscedasticity**: The log link can stabilize the variance in cases where the variance of the dependent variable increases with its magnitude.

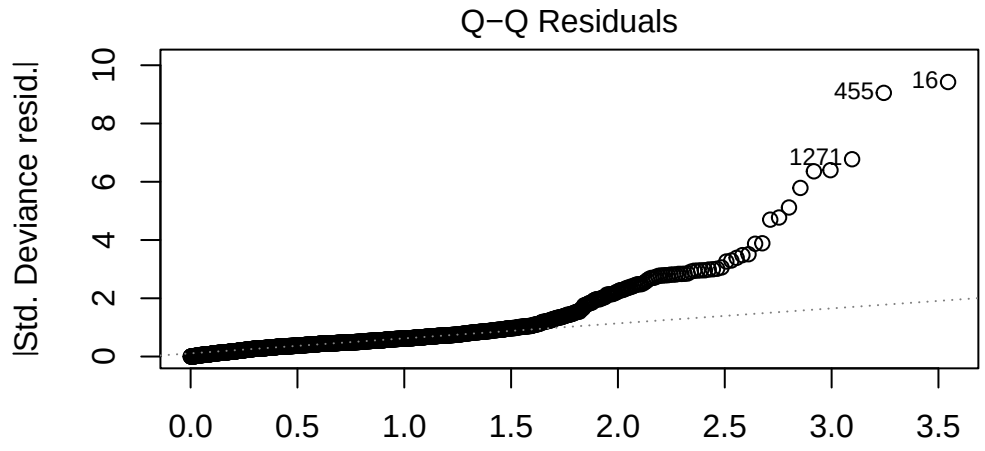
```
=== FINAL STEPWISE-SELECTED GLM MODEL ===


Call:
glm(formula = WDURHR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR +
    EDULEVEL + EMP + TPASS + TRIPS + NEARMALLKM + NEARCBDKM,
    family = gaussian(link = "log"), data = data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.6694747  0.0270732 135.539  < 2e-16 ***
CHOREHR     -0.1053604  0.0030212 -34.873  < 2e-16 ***
RECHR       -0.0931683  0.0036429 -25.575  < 2e-16 ***
OTHERHR     -0.0980618  0.0085897 -11.416  < 2e-16 ***
```

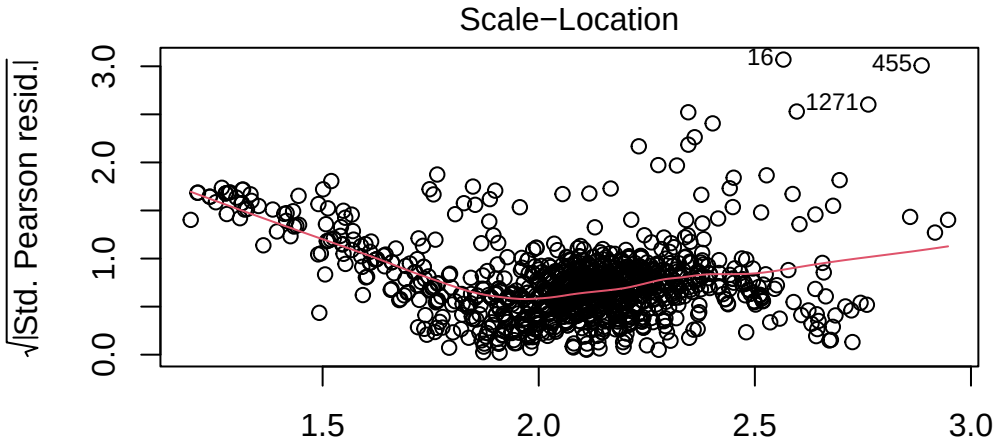
```
HOMEHR      -0.1002438  0.0016115 -62.203  < 2e-16 ***
PBHR        -0.1167287  0.0088460 -13.196  < 2e-16 ***
EDULEVEL2   -0.0300338  0.0125694  -2.389  0.01702 *
EDULEVEL3   -0.0102641  0.0111856  -0.918  0.35900
EMP2        -0.0205657  0.0174538  -1.178  0.23890
EMP3        -0.0216251  0.0103793  -2.083  0.03741 *
EMP4        -0.0225159  0.0173027  -1.301  0.19340
EMP5         0.0528270  0.0328498   1.608  0.10806
TPASS1      -0.0385678  0.0127346  -3.029  0.00251 **
TRIPS       -0.0055910  0.0019273  -2.901  0.00379 **
NEARMALLKM  -0.0014805  0.0008701  -1.702  0.08909 .
NEARCBDKM    0.0014822  0.0007896   1.877  0.06074 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.202405)

    Null deviance: 7146.7  on 1273  degrees of freedom
Residual deviance: 1512.6  on 1258  degrees of freedom
AIC: 3868.2

Number of Fisher Scoring iterations: 6


=== MODEL FIT STATISTICS ===

McFadden's R-squared: 0.3404

Adjusted McFadden's R-squared: 0.3352

AIC: 3868.18

BIC: 3955.73

Residual Deviance: 1512.622

Null Deviance: 7146.674
```

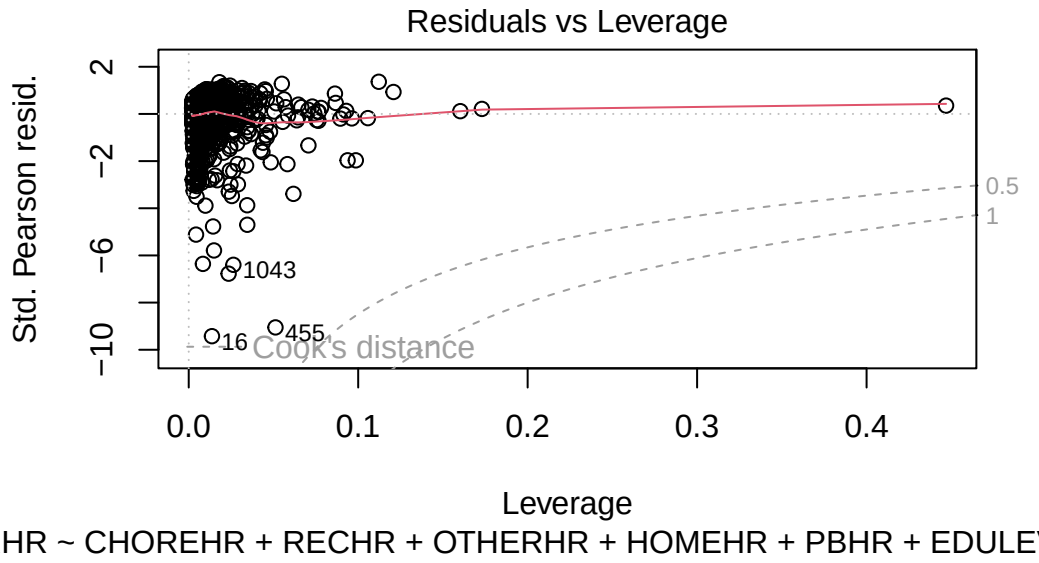



```
=== RESIDUAL DIAGNOSTICS ===
```




Q−Q Residuals
|Std. Deviance resid.|
0
2
4
6
8
10
455
16
1271
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Theoretical Quantiles
HR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR + EDULEV


Scale−Location
√|Std. Pearson resid.|
0.0
1.0
2.0
3.0
16
455
1271
1.5
2.0
2.5
3.0
Predicted values
HR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR + EDULEV



4.2.1 GLM Model Fit Statistics

McFadden's R-squared: McFadden's R-squared is a measure of how well the model fits the data, similar to the R-squared in linear regression. A value of 0.3404 means that the model explains about **34%** of the variability in the dependent variable.

AIC (Akaike Information Criterion): The AIC is a measure of model fit, penalized for the number of predictors in the model. A lower AIC indicates a better model. The value **3868.18** suggests a moderate fit.

BIC (Bayesian Information Criterion): Similar to the AIC, the BIC penalizes the model for complexity, but it applies a stronger penalty for the number of parameters. Like AIC, a lower BIC is preferred. The value **3955.73** is indicative of moderate fit.

Deviance Explained: The deviance explained is calculated as the reduction in deviance due to the predictors in the model. A value of **0.7883** suggests that about **78.8%** of the variability in `work duration` has been explained by the predictors, which is comparatively lower than the linear model.

Model Diagnostics Plots: The model diagnostic plots don't show any improvement over the linear model, as issues such as heteroscedasticity, non-normal residual distribution, and effect of outliers still persists.

4.3 Work Duration (IRLS Model)

Iteratively Reweighted Least Squares (IRLS) is an algorithm used in robust regression that iteratively adjusts weights to reduce the influence of outliers. It powers methods like M-estimation by solving weighted least squares problems where outliers receive progressively smaller weights.

Steps of Iteratively Reweighted Least Squares (IRLS):

1. **Initialize** the model using OLS (or any initial guess for the coefficients).
2. **Compute the residuals** for the current model.
3. **Calculate weights** for each residual. Typically, a **Huber weight function** is used, which applies the **absolute value loss** for large residuals and the **squared error loss** for small residuals.
4. **Re-estimate** the model using weighted least squares, using the weights computed in the previous step.
5. **Repeat** until convergence or until a maximum number of iterations is reached.

Here, **IRLS (Iteratively Reweighted Least Squares)** was applied to improve on **LM (Linear Regression)** and **GLS (Generalized Least Squares)** due to the presence of **outliers** in the data that could distort model estimates. IRLS is robust to outliers, providing more reliable coefficient estimates by down-weighting extreme residuals. Additionally, it helps address **mild heteroscedasticity** by using weights that reduce the influence of data points with larger residuals, improving model stability and accuracy. The following sections contain the code and the model outputs.

```r
# Load necessary libraries
library(readxl)  # For reading Excel files
library(dplyr)   # For data manipulation
library(MASS)    # For robust regression (rlm) and other helpers

# Load the data
data <- read_excel("C:/Onedrive_Sync/OneDrive - Dalhousie University/PhD Dal/Coursework/Da

# Check the structure of the data
str(data)
```

```
tibble [1,509 x 47] (S3: tbl_df/tbl/data.frame)
 $ INDID       : num [1:1509] 4432 4435 4438 4453 4457 ...
 $ HHID        : num [1:1509] 2629 2630 2630 2641 2644 ...
 $ WORKDURTOTAL: num [1:1509] 450 540 480 600 495 600 510 450 570 15 ...
 $ WDURHR      : num [1:1509] 7.5 9 8 10 8.25 10 8.5 7.5 9.5 0.25 ...
 $ LNWDUR      : num [1:1509] 6.11 6.29 6.17 6.4 6.2 ...
 $ CHORE       : num [1:1509] 150 105 0 0 60 360 0 0 0 0 ...
 $ REC         : num [1:1509] 105 0 45 90 0 0 0 270 30 0 ...
 $ OTHER       : num [1:1509] 0 0 90 0 0 0 0 0 0 0 ...
 $ HOME        : num [1:1509] 735 795 825 750 615 ...
 $ PERSONAL    : num [1:1509] 0 0 0 0 0 0 0 0 0 0 ...
 $ CHOREHR     : num [1:1509] 2.5 1.75 0 0 1 6 0 0 0 0 ...
 $ RECHR       : num [1:1509] 1.75 0 0.75 1.5 0 0 0 4.5 0.5 0 ...
 $ OTHERHR     : num [1:1509] 0 0 1.5 0 0 0 0 0 0 0 ...
```

```
 $ HOMEHR      : num [1:1509] 12.2 13.2 13.8 12.5 10.2 ...
 $ PBHR        : num [1:1509] 0 0 0 0 0 0 0 0 0 0 ...
 $ SHOPRECDUR  : num [1:1509] 1.75 0 0.75 1.5 0 0 0 4.5 0.5 0 ...
 $ SHOPRECVARHR: num [1:1509] 0.9502 0.7995 0.0499 0.7016 0.7982 ...
 $ Gender      : num [1:1509] 2 2 1 2 1 1 2 1 2 2 ...
 $ AGE         : num [1:1509] 22 54 66 25 45 42 22 41 29 27 ...
 $ AGEGROUP    : num [1:1509] 1 4 5 2 3 3 1 3 2 2 ...
 $ EDU         : num [1:1509] 6 6 6 6 4 2 4 6 6 6 ...
 $ EDULEVEL    : num [1:1509] 3 3 3 3 2 1 2 3 3 3 ...
 $ EMP         : num [1:1509] 1 1 2 1 1 1 1 1 1 1 ...
 $ INCOME      : num [1:1509] NA 7 7 6 5 5 5 6 4 7 ...
 $ INCOMELEVEL : num [1:1509] NA 4 4 3 3 3 3 3 2 4 ...
 $ LICENSE     : num [1:1509] 1 1 1 1 1 1 0 1 1 1 ...
 $ TPASS       : num [1:1509] 0 1 0 0 1 1 1 0 0 0 ...
 $ VEHNUM      : num [1:1509] 3 2 2 1 0 0 0 0 1 1 ...
 $ VEHNUMLVL   : num [1:1509] 3 2 2 1 0 0 0 0 1 1 ...
 $ BICYCLE     : num [1:1509] 2 1 1 1 0 0 2 0 0 0 ...
 $ BICYCLELVL  : num [1:1509] 2 1 1 1 0 0 2 0 0 0 ...
 $ HOMEOWNER   : num [1:1509] 1 1 1 2 1 1 2 2 1 2 ...
 $ LOCATIONYEAR: num [1:1509] 22 16 16 0 20 20 1 2 3 6 ...
 $ LOCYRLVL    : num [1:1509] 4 4 4 1 4 4 1 1 2 3 ...
 $ HHMEM       : num [1:1509] 3 4 4 1 3 3 3 1 1 2 ...
 $ HHMEMLVL    : num [1:1509] 2 2 2 1 2 2 2 1 1 2 ...
 $ HHHASCHLD   : num [1:1509] 1 1 1 0 1 1 1 0 0 0 ...
 $ TELEWORK    : num [1:1509] 0 1 1 1 1 1 1 1 1 1 ...
 $ TRIPS       : num [1:1509] 4 6 7 5 4 3 5 8 4 2 ...
 $ HOMELOC     : num [1:1509] 2 3 3 1 1 1 2 1 2 1 ...
 $ BUS5KM      : num [1:1509] 1 0 0 1 1 1 1 1 1 1 ...
 $ NEARMALLKM  : num [1:1509] 2.75 24.72 24.72 2.03 2.53 ...
 $ NEARSCHKM   : num [1:1509] 0.605 6.743 6.743 0.384 0.876 ...
 $ NEARGROCKM  : num [1:1509] 0.304 7.323 7.323 0.494 0.276 ...
 $ NEARCBDKM   : num [1:1509] 0.908 22.706 22.706 0 0 ...
 $ NEARRESTAUKM: num [1:1509] 0.42 7.53 7.53 0.102 0.132 ...
 $ LANDUSE     : num [1:1509] 0.183 0.183 0.183 0.334 0.183 ...
```

```r
# Convert categorical variables to factors
categorical_vars <- c("Gender", "AGEGROUP", "EDULEVEL", "EMP", "INCOMELEVEL", "LICENSE", "
                      "VEHNUMLVL", "BICYCLELVL", "HOMEOWNER", "LOCYRLVL", "HHMEMLVL", "HHH
                      "HOMELOC", "BUS5KM")

data <- data %>%
  mutate(across(all_of(categorical_vars), as.factor))

# Check for missing values
colSums(is.na(data))
```

```
      INDID         HHID WORKDURTOTAL       WDURHR       LNWDUR        CHORE
```

```
          0           0           0           0           0           0
        REC       OTHER        HOME    PERSONAL     CHOREHR       RECHR
          0           0           0           0           0           0
    OTHERHR      HOMEHR        PBHR SHOPRECDUR SHOPRECVARHR      Gender
          0           0           0           0           0           0
        AGE    AGEGROUP         EDU    EDULEVEL         EMP      INCOME
         21           0           1           2           0         159
INCOMELEVEL     LICENSE       TPASS      VEHNUM  VEHNUMLVL     BICYCLE
        159           0           0           0           0           0
 BICYCLELVL  HOMEOWNER LOCATIONYEAR    LOCYRLVL      HHMEM    HHMEMLVL
          0           0           0           0           0           0
  HHHASCHLD    TELEWORK       TRIPS     HOMELOC     BUS5KM  NEARMALLKM
          0           0          67          67           0           0
  NEARSCHKM  NEARGROCKM   NEARCBDKM NEARRESTAUKM     LANDUSE
          0           0           0           0           0
```

```r
# Handle missing values (listwise deletion)
data <- na.omit(data)  # Remove rows with any missing values

# Prepare the model formula
model_formula <- WDURHR ~ CHOREHR + RECHR + OTHERHR + HOMEHR + PBHR + Gender + AGEGROUP +
  EDULEVEL + EMP + INCOMELEVEL + LICENSE + TPASS + VEHNUMLVL + BICYCLELVL +
  HOMEOWNER + LOCYRLVL + HHMEMLVL + HHHASCHLD + TRIPS + HOMELOC + BUS5KM +
  NEARMALLKM + NEARSCHKM + NEARGROCKM + NEARCBDKM + NEARRESTAUKM + LANDUSE

# Initial model using ordinary least squares (OLS)
ols_model <- lm(model_formula, data = data)

# Get the initial coefficients and residuals
coefficients_ols <- coef(ols_model)
residuals_ols <- residuals(ols_model)

# Function for Huber weighting (robust weight function)
huber_weight <- function(residuals, k = 1.5) {
  # Huber weight function
  abs_resid <- abs(residuals)
  weights <- ifelse(abs_resid <= k, 1, k / abs_resid)
  return(weights)
}

# Maximum number of iterations for IRLS
max_iter <- 100
tolerance <- 1e-6

# Initialize residuals and weights
residuals_current <- residuals_ols
weights_current <- huber_weight(residuals_current)
```

```r
# Start IRLS algorithm
for (i in 1:max_iter) {
  # Weighted least squares regression with current weights
  weighted_model <- lm(model_formula, data = data, weights = weights_current)

  # Update coefficients and residuals
  coefficients_current <- coef(weighted_model)
  residuals_current <- residuals(weighted_model)

  # Calculate new weights using the Huber function
  weights_current <- huber_weight(residuals_current)

  # Check for convergence: if the coefficients change very little, stop the iteration
  if (sum((coefficients_current - coefficients_ols)^2) < tolerance) {
    cat("Convergence reached at iteration", i, "\n")
    break
  }

  # Update the coefficients for the next iteration
  coefficients_ols <- coefficients_current
}
```

```
Convergence reached at iteration 4
```

```r
# Final model summary after IRLS
cat("\n=== Final Robust Regression (IRLS) Model Summary ===\n")
```

```

=== Final Robust Regression (IRLS) Model Summary ===
```

```r
summary(weighted_model)
```

```

Call:
lm(formula = model_formula, data = data, weights = weights_current)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.8999 -0.1448  0.1654  0.3256  0.9321

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 23.006904   0.248381   92.627  < 2e-16 ***
CHOREHR     -0.968719   0.013282  -72.936  < 2e-16 ***
```

```
RECHR         -0.931705   0.017455  -53.378  < 2e-16 ***
OTHERHR       -0.955543   0.041990  -22.757  < 2e-16 ***
HOMEHR        -0.950972   0.007811 -121.756  < 2e-16 ***
PBHR          -0.923950   0.030283  -30.511  < 2e-16 ***
Gender2       -0.019958   0.036479   -0.547 0.584410
AGEGROUP2      0.051622   0.089797    0.575 0.565480
AGEGROUP3      0.011889   0.084785    0.140 0.888504
AGEGROUP4     -0.019933   0.086305   -0.231 0.817390
AGEGROUP5      0.056958   0.099198    0.574 0.565944
EDULEVEL2     -0.088464   0.064160   -1.379 0.168202
EDULEVEL3     -0.025390   0.059001   -0.430 0.667034
EMP2          -0.092459   0.086686   -1.067 0.286365
EMP3          -0.023982   0.060123   -0.399 0.690054
EMP4          -0.054179   0.105406   -0.514 0.607345
EMP5           0.259201   0.177372    1.461 0.144176
INCOMELEVEL2   0.225111   0.143482    1.569 0.116925
INCOMELEVEL3   0.168298   0.139093    1.210 0.226523
INCOMELEVEL4   0.264673   0.142803    1.853 0.064064 .
LICENSE1       0.075052   0.078616    0.955 0.339940
TPASS1        -0.230130   0.065264   -3.526 0.000437 ***
VEHNUMLVL1    -0.101295   0.096375   -1.051 0.293441
VEHNUMLVL2    -0.060296   0.103315   -0.584 0.559589
VEHNUMLVL3     0.011533   0.111638    0.103 0.917737
BICYCLELVL1    0.001767   0.056178    0.031 0.974911
BICYCLELVL2   -0.028479   0.054844   -0.519 0.603656
BICYCLELVL3   -0.041290   0.057591   -0.717 0.473546
HOMEOWNER2     0.075314   0.058760    1.282 0.200176
LOCYRLVL2     -0.010851   0.066482   -0.163 0.870376
LOCYRLVL3      0.027508   0.072896    0.377 0.705974
LOCYRLVL4      0.001895   0.062401    0.030 0.975778
HHMEMLVL2     -0.053280   0.068146   -0.782 0.434455
HHHASCHLD1    -0.016934   0.041396   -0.409 0.682561
TRIPS         -0.058067   0.009881   -5.877 5.39e-09 ***
HOMELOC2      -0.001876   0.049708   -0.038 0.969906
HOMELOC3      -0.024477   0.110113   -0.222 0.824124
BUS5KM1        0.034246   0.080240    0.427 0.669602
NEARMALLKM    -0.005097   0.004920   -1.036 0.300416
NEARSCHKM     -0.003360   0.022304   -0.151 0.880269
NEARGROCKM     0.011647   0.010530    1.106 0.268912
NEARCBDKM      0.003208   0.005273    0.608 0.542993
NEARRESTAUKM   0.072099   0.058575    1.231 0.218604
LANDUSE       -0.056305   0.161650   -0.348 0.727663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6304 on 1230 degrees of freedom
Multiple R-squared:  0.9291,    Adjusted R-squared:  0.9266
```

```
F-statistic: 374.9 on 43 and 1230 DF,  p-value: < 2.2e-16
```

```r
# Model fit statistics
cat("\n=== MODEL FIT STATISTICS ===\n")
```

```
=== MODEL FIT STATISTICS ===
```

```r
# AIC and BIC for the IRLS model
cat("AIC:", round(AIC(weighted_model), 2), "\n")
```

```
AIC: 2520.07
```

```r
cat("BIC:", round(BIC(weighted_model), 2), "\n")
```

```
BIC: 2751.82
```

```r
# RMSE (Root Mean Squared Error) for the IRLS model
rmse_irls <- sqrt(mean(residuals(weighted_model)^2))
cat("RMSE:", round(rmse_irls, 4), "\n")
```

```
RMSE: 0.8792
```

```r
# MAE (Mean Absolute Error) for the IRLS model
mae_irls <- mean(abs(residuals(weighted_model)))
cat("MAE:", round(mae_irls, 4), "\n")
```

```
MAE: 0.4604
```

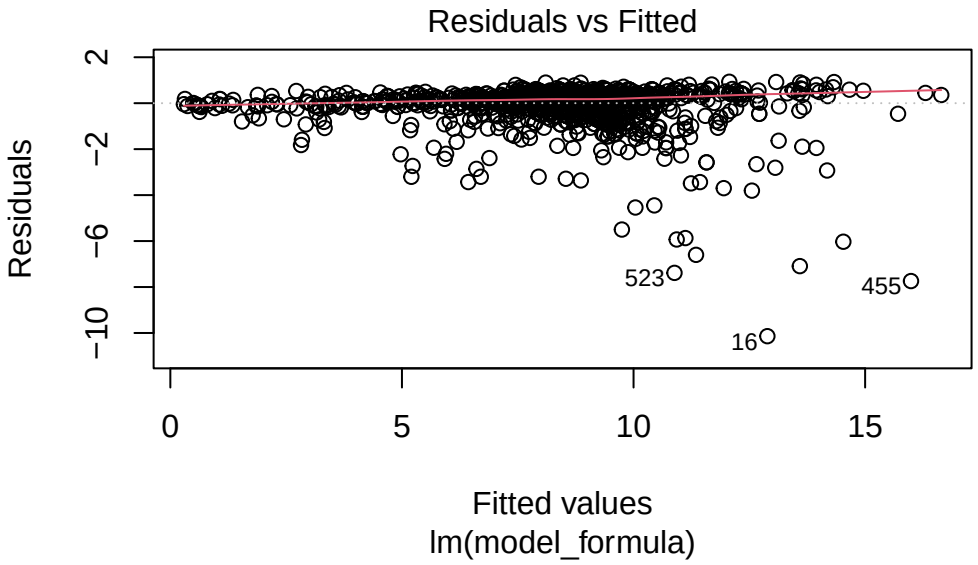
```r
# Check for multicollinearity using Variance Inflation Factor (VIF)
cat("\n=== MULTICOLLINEARITY CHECK ===\n")
```

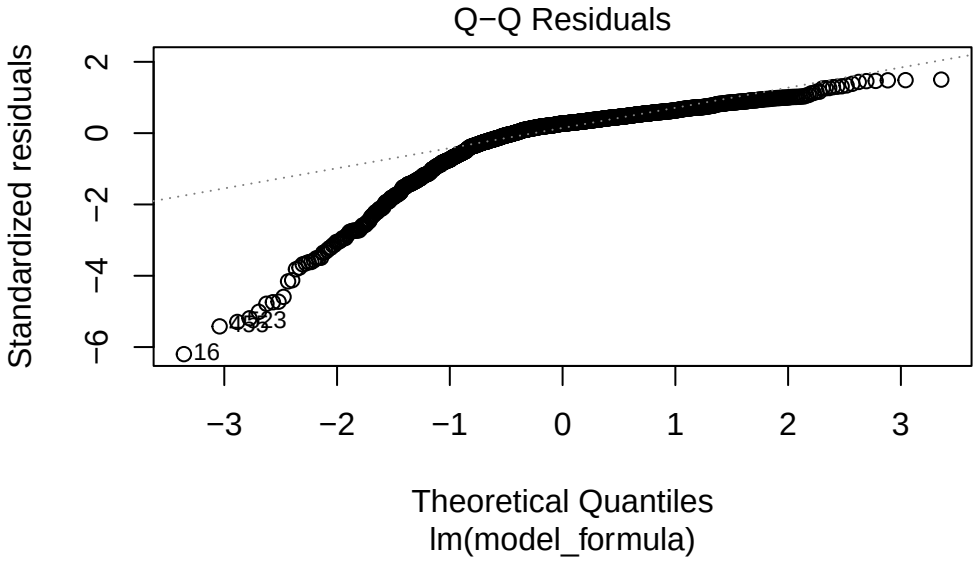
```
=== MULTICOLLINEARITY CHECK ===
```

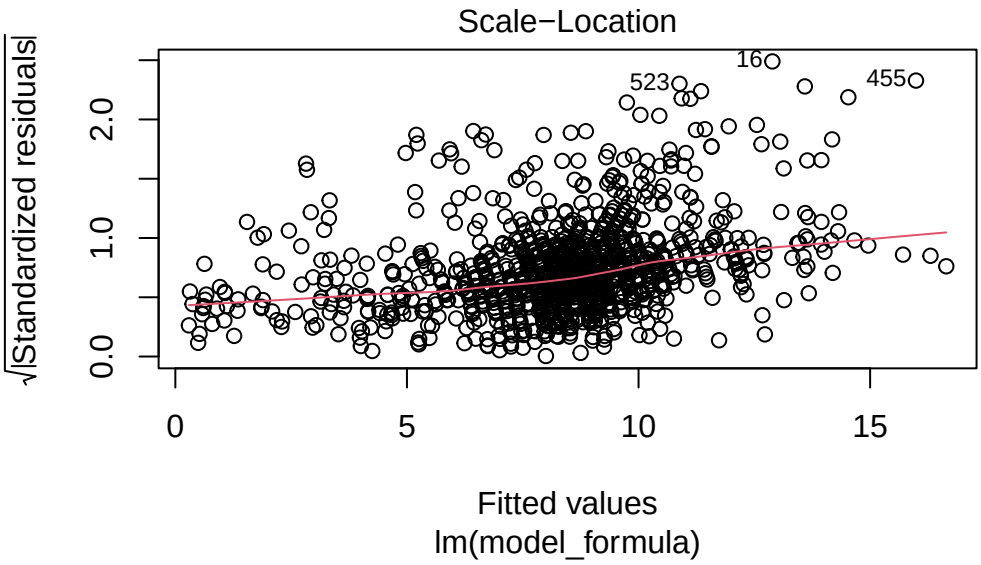
```r
vif_results <- car::vif(weighted_model)
print(vif_results)
```

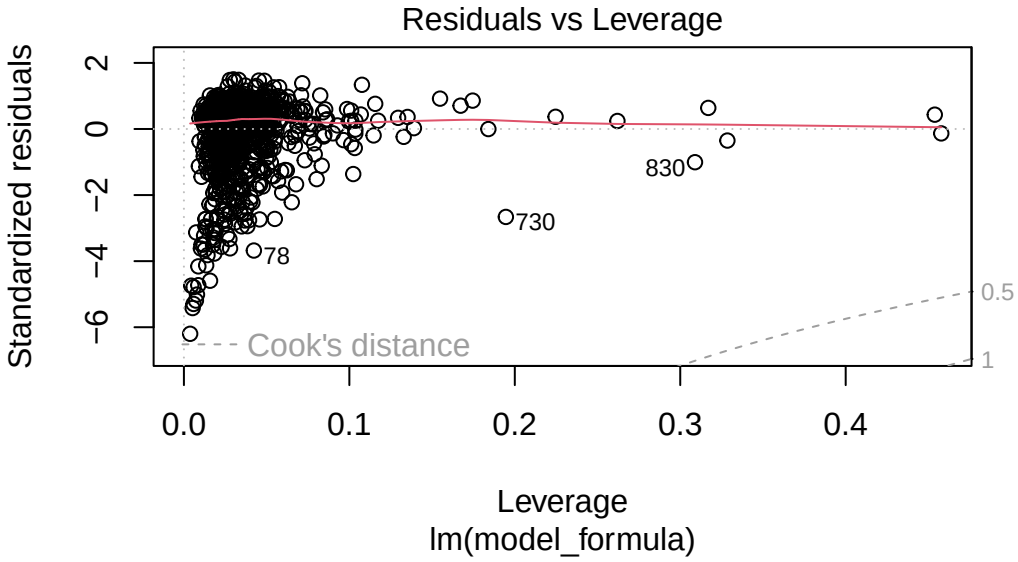
```
                GVIF Df GVIF^(1/(2*Df))
CHOREHR     1.379167  1        1.174379
RECHR       1.235950  1        1.111733
OTHERHR     1.065388  1        1.032176
HOMEHR      1.650068  1        1.284550
PBHR        1.047009  1        1.023235
Gender      1.046667  1        1.023067
AGEGROUP    2.830691  4        1.138903
EDULEVEL    1.485680  2        1.104031
EMP         3.117993  4        1.152748
INCOMELEVEL 1.819792  3        1.104936
LICENSE     1.491326  1        1.221199
TPASS       1.110914  1        1.053999
VEHNUMLVL   2.110058  3        1.132528
BICYCLELVL  1.331891  3        1.048926
HOMEOWNER   1.782963  1        1.335276
LOCYRLVL    1.688072  3        1.091185
HHMEMLVL    1.389831  1        1.178911
HHHASCHLD   1.345382  1        1.159906
TRIPS       1.125961  1        1.061113
HOMELOC     4.904798  2        1.488179
BUS5KM      3.328296  1        1.824362
NEARMALLKM  4.816544  1        2.194663
NEARSCHKM   2.190697  1        1.480100
NEARGROCKM  1.977097  1        1.406093
NEARCBDKM   6.857281  1        2.618641
NEARRESTAUKM 1.105563  1        1.051458
LANDUSE     1.072672  1        1.035699
```

```r
# Diagnostic plots for the final IRLS model
par(mfrow = c(1, 1))
plot(weighted_model)
```


Residuals vs Fitted
Residuals
2
−2
−6
−10
0
5
10
15
523
455
16
Fitted values
lm(model_formula)


Q−Q Residuals
Standardized residuals
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3
16
Theoretical Quantiles
lm(model_formula)





4.3.1 IRLS Model Fit Statistics

The model fit and performance statistics suggest that the **IRLS (Iteratively Reweighted Least Squares)** model provides a robust and well-fitting solution, especially considering the presence of outliers and mild heteroscedasticity.

- **AIC (2520.07)** and **BIC (2751.82)**: These values are reasonable and suggest a good fit, considering the complexity of the model. The lower these values are, the better the model is at balancing goodness of fit with complexity.

- **RMSE (0.8792)**: The **Root Mean Squared Error** indicates that the average deviation of the predicted values from the actual values is **0.8792**. This suggests that the model's predictions are relatively accurate, with small errors on average.
- **MAE (0.4604)**: The **Mean Absolute Error** represents the average magnitude of the errors without considering their direction. A value of **0.4604** suggests that, on average, the model's predictions are off by **0.46** units, which is acceptable for most practical applications.
- **Residual Standard Error (0.6304)**: This is another measure of how well the model fits the data. The **lower** the value, the better the model fits the data. The **residual standard error** is quite low, indicating that the model fits the data well.
- **R-squared (0.9291)** and **Adjusted R-squared (0.9266)**: These values indicate that approximately **93%** of the variance in the dependent variable (`WDURHR`) is explained by the model. This is a very strong fit, suggesting that the predictors included in the model are highly informative.
- **F-statistic (374.9)**: The high F-statistic with a **p-value $<$ 2.2e-16** indicates that the model is highly significant and that at least some of the predictors are contributing meaningfully to explaining the variability in

4.4 Commute Mode Choice (MNL Model)

Multinomial Logistic Regression (MNL) is a statistical model used for predicting categorical outcomes with more than two possible categories. In the context of **commute mode choice** (Auto, Bus, and Active Mode), the MNL model estimates the probability of an individual choosing one of these modes based on predictor variables.

The model works by comparing each mode against a reference category (active mode in this project) and estimating the log-odds of choosing one mode over the reference, given the values of the predictors. These predictors are classified into two categories:

1. **Demographic Factors**: Variables such as **Age**, **Gender**, **Employment Status**, **Education**, **Income Level**, **Driving License**, **Transit Pass**, **Vehicle Ownership**, **Home Ownership**, and **Children** influence an individual's commute choice by capturing socio-economic characteristics and lifestyle factors that may affect preferences for certain modes.
2. **Built Environment Factors**: Variables such as **Nearest Restaurant**, **Nearest Mall**, **Nearest School**, **Nearest CBD (Central Business District)**, **Nearest Grocery Store**, and **Land Use Index** reflect how proximity to key amenities and the overall urban environment affect travel mode choice. People are more likely to choose modes that align with their proximity to services, public transport options, and the convenience of travel.

By modeling these variables, MNL helps in understanding how demographic and environmental factors influence decisions among multiple commute modes, providing insights for urban planning and transportation policy.

4.4.1 Mode Choice Model (No Treatment for Unbalanced Data)

```
# Load necessary libraries
library(readxl)
library(dplyr)
library(nnet)
```

```
Warning: package 'nnet' was built under R version 4.4.3
```

```
library(broom)       # For tidy model output
```

```
Warning: package 'broom' was built under R version 4.4.3
```

```
library(kableExtra)  # For nice table formatting
```

```
Warning: package 'kableExtra' was built under R version 4.4.3
```

```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(MASS)        # For stepAIC function

# Explicitly specify dplyr's select function
select <- dplyr::select

# Load and prepare data
data <- read_excel("C:/Onedrive_Sync/OneDrive - Dalhousie University/PhD Dal/Coursework/Da

# Convert categorical variables to factors
categorical_vars <- c("Gender", "AGEGROUP", "EDULEVEL", "EMP", "INCOMELEVEL", "LICENSE", "
                      "VEHNUMLVL", "BICYCLELVL", "HOMEOWNER", "LOCYRLVL", "MODE", "HHMEMLV
                      "HOMELOC", "BUS5KM")

data <- data %>%
  mutate(across(all_of(categorical_vars), as.factor)) %>%
  na.omit()

# Set reference level for MODE (response variable)
data$MODE <- relevel(data$MODE, ref = "3")  # Setting level 3 as reference

# Fit initial null model (intercept only)
null_model <- multinom(MODE ~ 1, data = data)
```

```
# weights:  6 (2 variable)
initial  value 864.607871
iter  10 value 407.460811
iter  10 value 407.460811
iter  10 value 407.460811
final  value 407.460811
converged
```

```r
# Fit full model with all predictors
full_model <- multinom(MODE ~ Gender + AGEGROUP +
                         EDULEVEL + EMP + INCOMELEVEL + LICENSE + TPASS + VEHNUMLVL + BICY
                         HOMEOWNER + LOCYRLVL + HHMEMLVL + HHHASCHLD + TRIPS + HOMELOC + BU
                         NEARMALLKM + NEARSCHKM + NEARGROCKM + NEARCBDKM + NEARRESTAUKM + N
                       data = data,
                       maxit = 1000)
```

```
# weights:  120 (78 variable)
initial  value 864.607871
iter  10 value 403.989001
iter  20 value 312.023310
iter  30 value 269.212427
iter  40 value 260.783349
iter  50 value 260.296882
iter  60 value 260.237474
iter  70 value 260.235896
final  value 260.235843
converged
```

```r
# Perform stepwise AIC selection (forward only) - suppress intermediate output
invisible(capture.output(
  step_model <- stepAIC(null_model,
                        scope = list(lower = null_model, upper = full_model),
                        direction = "forward",
                        trace = FALSE)  # Set trace=FALSE to hide selection process
))

# Create a tidy table of results from the final stepwise model
results_table <- tidy(step_model, conf.int = TRUE) %>%
  mutate(
    t.stat = estimate / std.error,
    p.value = 2 * (1 - pnorm(abs(t.stat))),
    significance = case_when(
      p.value < 0.001 ~ "***",
      p.value < 0.01 ~ "**",
      p.value < 0.05 ~ "*",
      p.value < 0.1 ~ ".",
```

```
      TRUE ~ ""
    )
  ) %>%
  select(
    y.level, term,
    Coefficient = estimate,
    `Std. Error` = std.error,
    `t-stat` = t.stat,
    `p-value` = p.value,
    significance
  )

# Print the formatted table of final model results
results_table %>%
  kable(digits = 3, align = c("l", "l", "r", "r", "r", "r", "c")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = FALSE) %>%
  pack_rows(index = table(results_table$y.level)) %>%
  add_header_above(c(" " = 2, "Final Stepwise Multinomial Logit Model Results" = 5)) %>%
  footnote(
    general = "Reference category for MODE is level 3",
    symbol = c("*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1")
  )
```

		Final Stepwise Multinomial Logit Model Results				
y.level	term	Coefficient	Std. Error	t-stat	p-value	significance
1						
1	(Intercept)	-2.183	0.601	-3.634	0.000	***
1	NEARCBDKM	0.380	0.097	3.934	0.000	***
1	TPASS1	0.266	0.556	0.480	0.632	
1	NEARGROCKM	1.251	0.461	2.713	0.007	**
1	VEHNUMLVL1	1.587	0.486	3.265	0.001	**
1	VEHNUMLVL2	1.782	0.510	3.497	0.000	***
1	VEHNUMLVL3	1.805	0.623	2.895	0.004	**
1	EMP2	1.387	0.844	1.642	0.100	
1	EMP3	1.844	0.759	2.430	0.015	*
1	EMP4	-0.051	1.100	-0.046	0.963	
1	EMP5	-3.540	1.605	-2.206	0.027	*
1	BICYCLELVL1	-0.467	0.392	-1.190	0.234	
1	BICYCLELVL2	-1.211	0.408	-2.967	0.003	**
1	BICYCLELVL3	-0.864	0.460	-1.880	0.060	.
1	NEARSCHKM	1.719	0.670	2.566	0.010	*
1	LOCYRLVL2	1.080	0.467	2.312	0.021	*
1	LOCYRLVL3	0.625	0.495	1.261	0.207	
1	LOCYRLVL4	0.583	0.377	1.548	0.122	

Reference category for MODE is level 3

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

2						
2	(Intercept)	-4.307	1.016	-4.239	0.000	***
2	NEARCBDKM	0.340	0.102	3.329	0.001	***
2	TPASS1	2.034	0.638	3.188	0.001	**
2	NEARGROCKM	0.602	0.595	1.010	0.312	
2	VEHNUMLVL1	0.174	0.652	0.267	0.789	
2	VEHNUMLVL2	0.466	0.666	0.699	0.484	
2	VEHNUMLVL3	0.067	0.933	0.072	0.943	
2	EMP2	0.636	1.132	0.562	0.574	
2	EMP3	1.628	0.907	1.794	0.073	.
2	EMP4	0.345	1.513	0.228	0.820	
2	EMP5	-0.376	1.459	-0.258	0.796	
2	BICYCLELVL1	-0.070	0.576	-0.121	0.904	
2	BICYCLELVL2	-2.531	1.099	-2.303	0.021	*
2	BICYCLELVL3	-1.780	0.893	-1.993	0.046	*
2	NEARSCHKM	1.298	0.850	1.528	0.126	
2	LOCYRLVL2	2.782	0.912	3.049	0.002	**
2	LOCYRLVL3	2.561	0.941	2.722	0.006	**
2	LOCYRLVL4	1.838	0.852	2.156	0.031	*

Note:
Reference category for MODE is level 3
* *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

```
# Model fit statistics for final model
cat("\nFinal Model Fit Statistics:\n")
```

```
Final Model Fit Statistics:
```

```
cat("AIC:", AIC(step_model), "\n")
```

```
AIC: 635.7908
```

```
cat("BIC:", BIC(step_model), "\n")
```

```
BIC: 803.847
```

```
cat("Log-Likelihood:", logLik(step_model), "\n")
```

```
Log-Likelihood: -281.8954
```

```r
# McFadden's pseudo R-squared for final model
mcfadden_r2 <- 1 - (logLik(step_model)/logLik(null_model))
cat("McFadden's R-squared:", mcfadden_r2, "\n")
```

```
McFadden's R-squared: 0.3081657
```

```r
# Confusion matrix for final model
predicted_classes <- predict(step_model)
confusion_matrix <- table(Actual = data$MODE, Predicted = predicted_classes)
cat("\nConfusion Matrix for Final Model:\n")
```

```

Confusion Matrix for Final Model:
```

```r
print(confusion_matrix)
```

```
      Predicted
Actual   3   1   2
     3  22  57   0
     1  11 657   1
     2   2  31   6
```

```r
# Calculate accuracy for final model
accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
cat("\nClassification Accuracy:", round(accuracy, 3), "\n")
```

```

Classification Accuracy: 0.87
```

4.4.2 Mode Choice Model (Treatment for Unbalanced Data)

In this project, **Random Replicaton** was used to address the issue of **class imbalance** in the mode choice data, where **85.8%** of the responses were for **Auto**, **4.6%** for **Transit**, and **9.8%** for **Active Mode**. This imbalance caused the model to be biased towards the majority class (Auto), making it difficult to accurately predict the minority classes (Transit and Active Mode). By applying **Random Replication**, synthetic samples were generated for the minority classes, helping to balance the dataset and improve the model's ability to distinguish between the different modes of transportation. This approach enhanced the model's performance by ensuring more accurate predictions for all three commute modes, leading to better generalization and fairness in the model. The code, model output, and the model fit statistics have been presented below:

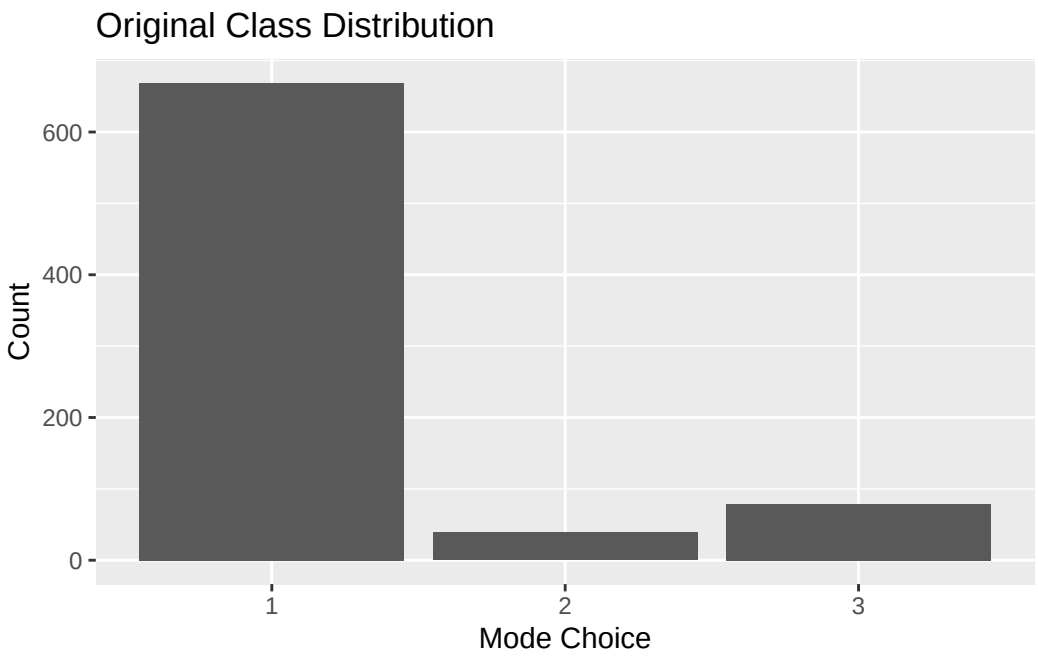
```r
# Load necessary libraries
library(readxl)        # For reading Excel files
library(dplyr)         # For data manipulation
library(nnet)          # For multinomial logistic regression
library(broom)         # For tidy model output
library(kableExtra)    # For nice table formatting
library(MASS)          # For stepAIC function
library(ggplot2)       # For visualizations

# Load and prepare data
data <- read_excel("C:/Onedrive_Sync/OneDrive - Dalhousie University/PhD Dal/Coursework/Da

# Convert categorical variables to factors
categorical_vars <- c("Gender", "AGEGROUP", "EDULEVEL", "EMP", "INCOMELEVEL", "LICENSE", "
                      "VEHNUMLVL", "BICYCLELVL", "HOMEOWNER", "LOCYRLVL", "MODE", "HHMEMLV
                      "HOMELOC", "BUS5KM")

data <- data %>%
  mutate(across(all_of(categorical_vars), as.factor)) %>%
  na.omit()

# Check initial class distribution
cat("Original class distribution:\n")
```

```
Original class distribution:
```

```r
print(table(data$MODE))
```

```
  1   2   3
669  39  79
```

```r
# Plot original class distribution
ggplot(data, aes(x = MODE)) +
  geom_bar() +
  labs(title = "Original Class Distribution", x = "Mode Choice", y = "Count")
```




```
# 1. Simple Oversampling (Random Replication) ----------------------------

# Set target size for each class (desired number of samples, e.g., 333)
target_size <- 333

# Initialize balanced dataset
balanced_data <- data.frame()

# Apply oversampling for each class
set.seed(123)
for (class in levels(data$MODE)) {
  current_class <- data %>% filter(MODE == class)
  current_count <- nrow(current_class)

  if (current_count < target_size) {
    # Calculate how many samples are needed to reach target_size
    needed_samples <- target_size - current_count
    oversampled_class <- current_class[sample(1:current_count, size = needed_samples, repla

    # Combine with original class
    balanced_class <- rbind(current_class, oversampled_class)
  } else {
    # For majority class, just take a sample of size target_size
    balanced_class <- current_class %>%
      sample_n(size = target_size, replace = FALSE)
  }

  balanced_data <- rbind(balanced_data, balanced_class)
```

```
}

# Verify balanced class distribution
cat("\nClass distribution after oversampling:\n")
```

```
Class distribution after oversampling:
```

```
print(table(balanced_data$MODE))
```

```
  1   2   3
333 333 333
```

```
# Plot balanced class distribution
ggplot(balanced_data, aes(x = MODE)) +
  geom_bar() +
  labs(title = "Balanced Class Distribution After Oversampling", x = "Mode Choice", y = "C
```


```
# 2. Model Development ----------------------------------------------------
balanced_data$MODE <- relevel(balanced_data$MODE, ref = "3")

# Fit initial null model
null_model <- multinom(MODE ~ 1, data = balanced_data, trace = FALSE)

# Fit full model with carefully selected predictors
```

```r
full_model <- multinom(MODE ~ Gender + AGEGROUP + EDULEVEL + EMP +
                         INCOMELEVEL + LICENSE + VEHNUMLVL + HOMELOC +
                         TRIPS + NEARCBDKM,
                       data = balanced_data,
                       maxit = 1000,
                       trace = FALSE)

# Verify convergence
if(full_model$convergence != 0) {
  full_model <- multinom(formula(full_model), data = balanced_data,
                         maxit = 5000, trace = FALSE)
}

# 3. Stepwise Selection ------------------------------------------------
safe_scope <- list(
  lower = formula(null_model),
  upper = formula(full_model)
)

step_model <- tryCatch({
  stepAIC(null_model,
          scope = safe_scope,
          direction = "forward",
          trace = FALSE)
}, error = function(e) {
  return(full_model)
})

# 4. Model Evaluation --------------------------------------------------
results_table <- broom::tidy(step_model, conf.int = TRUE) %>%
  mutate(
    t.stat = estimate / std.error,
    p.value = 2 * (1 - pnorm(abs(t.stat))),
    significance = case_when(
      p.value < 0.001 ~ "***",
      p.value < 0.01 ~ "**",
      p.value < 0.05 ~ "*",
      p.value < 0.1 ~ ".",
      TRUE ~ ""
    ),
    odds.ratio = exp(estimate)
  ) %>%
  select(
    y.level, term,
    Coefficient = estimate,
    `Std. Error` = std.error,
```

```
    `Odds Ratio` = odds.ratio,
    `t-stat` = t.stat,
    `p-value` = p.value,
    significance
  )

# Print results
results_table %>%
  kable(digits = 3, align = c("l", "l", "r", "r", "r", "r", "r", "c")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = FALSE) %>%
  pack_rows(index = table(results_table$y.level)) %>%
  add_header_above(c(" " = 2, "Stepwise Multinomial Logit Results" = 6)) %>%
  footnote(
    general = "Reference category for MODE is level 3",
    symbol = c("*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1")
  )
```

		Stepwise Multinomial Logit Results					
y.level	term	Coefficient	Std. Error	Odds Ratio	t-stat	p-value	significance
1							
1	(Intercept)	-3.181	1.174	0.042	-2.710	0.007	**
1	HOMELOC2	1.147	0.288	3.148	3.979	0.000	***
1	HOMELOC3	13.923	0.000	1113423.652	13016996.915	0.000	***
1	LICENSE1	0.188	0.471	1.206	0.398	0.690	
1	AGEGROUP2	-0.812	0.469	0.444	-1.731	0.083	.
1	AGEGROUP3	-0.187	0.450	0.829	-0.417	0.677	
1	AGEGROUP4	0.347	0.469	1.415	0.740	0.460	
1	AGEGROUP5	-0.081	0.523	0.922	-0.155	0.877	
1	NEARCBDKM	0.267	0.060	1.306	4.478	0.000	***
1	EMP2	1.403	0.533	4.069	2.632	0.008	**
1	EMP3	1.776	0.433	5.905	4.098	0.000	***
1	EMP4	-0.450	1.029	0.638	-0.437	0.662	
1	EMP5	-2.414	1.159	0.089	-2.083	0.037	*
1	VEHNUMLVL1	1.402	0.435	4.065	3.227	0.001	**
1	VEHNUMLVL2	1.964	0.449	7.126	4.376	0.000	***
1	VEHNUMLVL3	2.137	0.528	8.472	4.049	0.000	***
1	INCOMELEVEL2	1.473	0.999	4.362	1.474	0.141	
1	INCOMELEVEL3	0.662	0.992	1.939	0.667	0.505	
1	INCOMELEVEL4	0.475	1.007	1.609	0.472	0.637	
1	Gender2	-0.392	0.213	0.675	-1.840	0.066	.
1	TRIPS	-0.064	0.056	0.938	-1.134	0.257	
2							
2	(Intercept)	2.213	0.769	9.145	2.879	0.004	**

2	HOMELOC2	0.190	0.287	1.210	0.664	0.507	
2	HOMELOC3	-4.150	0.000	0.016	-5747135.245	0.000	***
2	LICENSE1	-1.456	0.315	0.233	-4.617	0.000	***
2	AGEGROUP2	-0.180	0.390	0.835	-0.461	0.645	
2	AGEGROUP3	-1.243	0.414	0.288	-3.006	0.003	**
2	AGEGROUP4	0.030	0.419	1.030	0.071	0.943	
2	AGEGROUP5	-0.593	0.482	0.552	-1.231	0.218	
2	NEARCBDKM	0.299	0.061	1.348	4.904	0.000	***
2	EMP2	-0.044	0.579	0.957	-0.077	0.939	
2	EMP3	1.573	0.438	4.822	3.595	0.000	***
2	EMP4	-1.022	0.881	0.360	-1.160	0.246	
2	EMP5	-0.374	0.563	0.688	-0.665	0.506	
2	VEHNUMLVL1	0.416	0.322	1.516	1.292	0.196	
2	VEHNUMLVL2	1.036	0.342	2.819	3.030	0.002	**
2	VEHNUMLVL3	0.947	0.437	2.578	2.166	0.030	*
2	INCOMELEVEL2	0.115	0.600	1.122	0.192	0.848	
2	INCOMELEVEL3	-0.783	0.629	0.457	-1.246	0.213	
2	INCOMELEVEL4	-1.514	0.646	0.220	-2.343	0.019	*
2	Gender2	-0.630	0.205	0.533	-3.075	0.002	**
2	TRIPS	-0.137	0.052	0.872	-2.618	0.009	**

Note:
Reference category for MODE is level 3
* *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

```
# Model statistics
cat("\nFinal Model Fit Statistics:\n")
```

```
Final Model Fit Statistics:
```

```
cat("AIC:", AIC(step_model), "\n")
```

```
AIC: 1645.832
```

```
cat("BIC:", BIC(step_model), "\n")
```

```
BIC: 1851.916
```

```
cat("Log-Likelihood:", logLik(step_model), "\n")
```

```
Log-Likelihood: -780.916
```

```r
mcfadden_r2 <- 1 - (logLik(step_model)/logLik(null_model))
cat("McFadden's R-squared:", mcfadden_r2, "\n")
```

```
McFadden's R-squared: 0.2884681
```

```r
# Confusion matrix
predicted_classes <- predict(step_model)
confusion_matrix <- table(Actual = balanced_data$MODE, Predicted = predicted_classes)
cat("\nConfusion Matrix:\n")
```

```
Confusion Matrix:
```

```r
print(confusion_matrix)
```

```
      Predicted
Actual   3   1   2
     3 248  37  48
     1  55 223  55
     2  84  83 166
```

```r
cat("\nOverall Accuracy:", mean(predicted_classes == balanced_data$MODE), "\n")
```

```
Overall Accuracy: 0.6376376
```

```r
# Save results
saveRDS(list(
  data = balanced_data,
  models = list(null = null_model, full = full_model, final = step_model),
  results = results_table,
  confusion = confusion_matrix
), "mode_choice_results.rds")
```

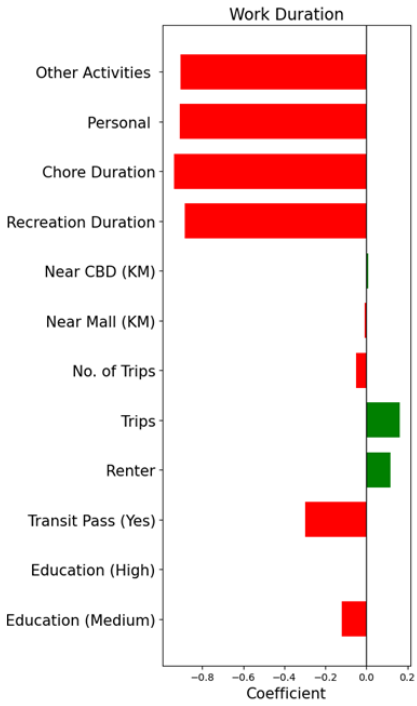
Comparison between the Balanced and Unbalanced Model:

While the overall accuracy of the balanced model is lower, the prediction accuracy of the underrepresented transportation modes increase significantly- 53.7% from 7.1% for Class 2 (transit), and 64.1% from 62.9% for active mode.

5. Results

5.1 Work Duration Model Results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.516234   0.185114 121.634  < 2e-16 ***
CHOREHR     -0.938361   0.018209 -51.533  < 2e-16 ***
RECHR       -0.885924   0.023887 -37.089  < 2e-16 ***
OTHERHR     -0.905576   0.057417 -15.772  < 2e-16 ***
HOMEHR      -0.913970   0.010606 -86.178  < 2e-16 ***
PBHR        -0.908433   0.041687 -21.792  < 2e-16 ***
EDULEVEL2   -0.120777   0.080866  -1.494 0.135543
EDULEVEL3    0.002501   0.070618   0.035 0.971754
TPASS1      -0.299757   0.086501  -3.465 0.000547 ***
HOMEOWNER2   0.117099   0.062425   1.876 0.060910 .
TRIPS       -0.052959   0.013470  -3.932 8.89e-05 ***
NEARMALLKM  -0.008621   0.006091  -1.415 0.157186
NEARCBDKM    0.008245   0.005519   1.494 0.135469
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8812 on 1261 degrees of freedom
Multiple R-squared:  0.863,    Adjusted R-squared:  0.8617
F-statistic: 661.9 on 12 and 1261 DF,  p-value: < 2.2e-16
```



Summary:

- **Highly Significant** (p-value < 0.001): CHOREHR, RECHR, OTHERHR, HOMEHR, PBHR, TPASS1, TRIPS
- **Moderately Significant** (p-value < 0.1): HOMEOWNER2
- **Not Significant**: EDULEVEL2, EDULEVEL3, NEARMALLKM, NEARCBDKM

Interpretation:

- Time spent on activities like **chores, recreation, home activities**, and **physical activity** is negatively associated with **work duration**, meaning as people spend more time on these activities, their work hours tend to decrease.
- **Having a transit pass** also significantly reduces work duration, possibly due to not having a flexible working environment and tendency to align with bus's schedule.
- **Daily total trips** also show a negative association with work duration, indicating that more frequent trips might be linked to shorter working hours.
- **Education level** and **distance to nearest amenities (mall, CBD)** do not have strong or significant effects on work duration in this model.

Interpretation of the Model Co-efficients:

- **CHOREHR**: Each additional hour spent on chores reduces work duration by **0.938 hours** (highly significant, p < 2e-16).
- **RECHR**: Each additional hour spent on recreation reduces work duration by **0.886 hours** (highly significant, p < 2e-16).
- **OTHERHR**: Each additional hour spent on other activities reduces work duration by **0.906 hours** (significant, p < 2e-16).
- **HOMEHR**: Each additional hour spent at home reduces work duration by **0.914 hours** (highly significant, p < 2e-16).
- **PBHR**: Each additional hour spent on personal business reduces work duration by **0.908 hours** (significant, p < 2e-16).
- **EDULEVEL2**: Medium education level has a **marginal effect** on work duration (not significant, p = 0.135).
- **EDULEVEL3**: Higher education level has **no significant effect** on work duration (p = 0.971).
- **TPASS1**: Having a transit pass reduces work duration by **0.3 hours** (significant, p = 0.000547).
- **HOMEOWNER2**: Rentership increases work duration by **0.117 hours**, but the effect is **borderline significant** (p = 0.0609).
- **TRIPS**: Each additional trip reduces work duration by **0.053 hours** (significant, p = 8.89e-05).
- **NEARMALLKM**: Greater distance to the nearest mall slightly reduces work duration, but the effect is **not significant** (p = 0.157).
- **NEARCBDKM**: Greater distance to the nearest Central Business District (CBD) slightly increases work duration, but the effect is **not significant** (p = 0.135).

5.2 Mode Choice Model Results

		Stepwise Multinomial Logit Results					
y.level	**term**	**Coefficient**	**Std. Error**	**Odds Ratio**	**t-stat**	**p-value**	**significance**
1							
1	(Intercept)	-3.181	1.174	0.042	-2.710	0.007	**
1	HOMELOC2	1.147	0.288	3.148	3.979	0.000	***
1	HOMELOC3	13.923	0.000	1113423.652	13016996.915	0.000	***
1	LICENSE1	0.188	0.471	1.206	0.398	0.690	
1	AGEGROUP2	-0.812	0.469	0.444	-1.731	0.083	.
1	AGEGROUP3	-0.187	0.450	0.829	-0.417	0.677	
1	AGEGROUP4	0.347	0.469	1.415	0.740	0.460	
1	AGEGROUP5	-0.081	0.523	0.922	-0.155	0.877	
1	NEARCBDKM	0.267	0.060	1.306	4.478	0.000	***
1	EMP2	1.403	0.533	4.069	2.632	0.008	**
1	EMP3	1.776	0.433	5.905	4.098	0.000	***
1	EMP4	-0.450	1.029	0.638	-0.437	0.662	
1	EMP5	-2.414	1.159	0.089	-2.083	0.037	*
1	VEHNUMLVL1	1.402	0.435	4.065	3.227	0.001	**
1	VEHNUMLVL2	1.964	0.449	7.126	4.376	0.000	***
1	VEHNUMLVL3	2.137	0.528	8.472	4.049	0.000	***
1	INCOMELEVEL2	1.473	0.999	4.362	1.474	0.141	
1	INCOMELEVEL3	0.662	0.992	1.939	0.667	0.505	
1	INCOMELEVEL4	0.475	1.007	1.609	0.472	0.637	
1	Gender2	-0.392	0.213	0.675	-1.840	0.066	.
1	TRIPS	-0.064	0.056	0.938	-1.134	0.257	

Key Findings:

- **Mode 1 (Auto)** is more likely to be chosen over **Active Mode** for individuals with a driving license, higher vehicle ownership, higher income, middle aged, male and those living in suburban or rural areas.

- Interpretation of the Model Co-efficients:

    1. **HOMELOC2**: Living in Home Location 2 increases the odds of choosing Auto by 3.15 times (p-value < 0.001).
    2. **HOMELOC3**: Living in Home Location 3 strongly increases the likelihood of choosing Auto, with odds of 1113423.652 (p-value < 0.001).
    3. **LICENSE1**: Having a driving license increases the odds of choosing Auto by 1.206 times, but the effect is not significant ($p = 0.690$).
    4. **AGEGROUP2**: Older individuals (Age group 2) are less likely to choose Auto, with a 0.444 decrease in the odds ($p = 0.083$).
    5. **NEARCBDKM**: As distance to the nearest CBD increases, the odds of choosing Auto increases by 1.306 times (p-value < 0.001).
    6. **EMP2**: Employment status category 2 increases the odds of choosing Auto by 4.069 times (p-value $= 0.008$).
    7. **EMP3**: Employment status category 3 significantly increases the odds of choosing Auto by 5.905 times (p-value < 0.001).
    8. **VEHNUMLVL1**: Having one vehicle increases the odds of choosing Auto by 4.065 times (p-value $= 0.001$).
    9. **VEHNUMLVL2**: Having two vehicles increases the odds of choosing Auto by 7.126 times (p-value < 0.001).
    10. **VEHNUMLVL3**: Having three vehicles increases the odds of choosing Auto by 8.472 times (p-value < 0.001).

2							
2	(Intercept)	2.213	0.769	9.145	2.879	0.004	**
2	HOMELOC2	0.190	0.287	1.210	0.664	0.507	
2	HOMELOC3	-4.150	0.000	0.016	-5747135.245	0.000	***
2	LICENSE1	-1.456	0.315	0.233	-4.617	0.000	***
2	AGEGROUP2	-0.180	0.390	0.835	-0.461	0.645	
2	AGEGROUP3	-1.243	0.414	0.288	-3.006	0.003	**
2	AGEGROUP4	0.030	0.419	1.030	0.071	0.943	
2	AGEGROUP5	-0.593	0.482	0.552	-1.231	0.218	
2	NEARCBDKM	0.299	0.061	1.348	4.904	0.000	***
2	EMP2	-0.044	0.579	0.957	-0.077	0.939	
2	EMP3	1.573	0.438	4.822	3.595	0.000	***
2	EMP4	-1.022	0.881	0.360	-1.160	0.246	
2	EMP5	-0.374	0.563	0.688	-0.665	0.506	
2	VEHNUMLVL1	0.416	0.322	1.516	1.292	0.196	
2	VEHNUMLVL2	1.036	0.342	2.819	3.030	0.002	**
2	VEHNUMLVL3	0.947	0.437	2.578	2.166	0.030	*
2	INCOMELEVEL2	0.115	0.600	1.122	0.192	0.848	
2	INCOMELEVEL3	-0.783	0.629	0.457	-1.246	0.213	
2	INCOMELEVEL4	-1.514	0.646	0.220	-2.343	0.019	*
2	Gender2	-0.630	0.205	0.533	-3.075	0.002	**
2	TRIPS	-0.137	0.052	0.872	-2.618	0.009	**

Note:

Reference category for MODE is level 3

* *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Key Findings:

- **Mode 2 (Transit)** is more likely to be chosen over **Active Mode** for individuals with higher vehicle ownership, low income, urban/suburban dwellers, retirees, and non-license owners.

- Model Coefficients:

- **Intercept**: When all predictors are at baseline, the log-odds of choosing Transit over Active Mode is 2.213, indicating a preference for Transit when no other factors are considered.

- **HOMELOC3**: Living in Rural location significantly decreases the likelihood of choosing Transit over Active Mode, with an odds ratio of 0.016 (p-value < 0.001).

- **LICENSE1**: Having a driving license decreases the odds of choosing Transit over Active Mode by 0.233 times (p-value < 0.001).

- **AGEGROUP3**: Middle aged individuals (35-50) are less likely to choose Transit over Active Mode, with a 0.288 decrease in odds ($p = 0.003$).

- **NEARCBDKM**: As the distance to the nearest CBD increases, the odds of choosing Transit over Active Mode increases by 1.348 times (p-value < 0.001).

- **EMP3**: Being a retiree significantly increases the odds of choosing Transit by 4.822 times (p-value < 0.001).

- **VEHNUMLVL2**: Having two vehicles increases the odds of choosing Transit over Active Mode by 2.819 times ($p = 0.002$).

- **VEHNUMLVL3**: Having three vehicles increases the odds of choosing Transit over Active Mode by 2.578 times ($p = 0.030$).

- **INCOMELEVEL4**: Higher-income individuals (>100K/year) are less likely to choose Transit over Active Mode, with odds decreasing by 0.220 times ($p = 0.019$).

- **Gender2**: Females are less likely to choose Transit over Active Mode, with odds decreasing by 0.533 times ($p = 0.002$).

- **TRIPS**: More total trips per day decrease the likelihood of choosing Transit over Active Mode, with odds decreasing by 0.872 times ($p = 0.009$).

6. Conclusions

This research provides valuable insights into the complex relationships between sociodemographic factors, commute patterns, and the built environment, and their influence on daily work duration and transportation mode choice. By utilizing a variety of statistical models, including Ordinary Least Squares (OLS) regression, Generalized Linear Models (GLM), Iteratively Reweighted Least Squares (IRLS) regression, and Multinomial Logistic Regression (MNL), we have explored how different factors shape individuals' work habits and their preferred transportation modes.

The analysis reveals that sociodemographic factors such as age, gender, income, and vehicle ownership play significant roles in determining both the number of hours individuals spend working each day and their choice of commute mode. For instance, higher vehicle ownership, income, and middle age were strongly associated with a preference for private auto use over active modes. Conversely, lower income, the lack of a driving license, and living in urban or suburban areas increased the likelihood of using transit over active modes. Additionally, factors such as home location, employment status, and proximity to essential services (e.g., the nearest CBD or mall) were critical determinants of mode choice.

Moreover, this research underscores the influence of time-use patterns, such as time spent on chores, recreation, and other personal activities, which were found to negatively affect work duration. As individuals allocate more time to non-work activities, their total work hours tend to decrease, further shaping daily travel behavior.

From a transportation planning and policy perspective, these findings emphasize the need for tailored solutions that accommodate diverse work patterns and travel behaviors. Policies that promote flexible work hours, improved public transit accessibility, and better integration of the built environment can help mitigate congestion and reduce transportation disparities across different population segments. Future studies should further investigate these relationships to refine urban transportation policies and enhance the efficiency and equity of transport systems.

References

Boarnet, M. G., & Crane, R. (2001). *Travel by design: The influence of urban form on travel*. Oxford University Press.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation research part D: Transport and environment*, *2*(3), 199-219.

Fan, J. X., Wen, M., & Kowaleski-Jones, L. (2015). Sociodemographic and environmental correlates of active commuting in rural America. The Journal of Rural Health, 31(2), 176–185.

Givoni, M., & Rietveld, P. (2014). Do cities deserve more railway stations? The choice of a departure railway station in a multiple-station region. *Journal of Transport Geography*, *36*, 89-97.

Graham, D. J., & Glaister, S. (2004). Road traffic demand elasticity estimates: a review. *Transport reviews*, *24*(3), 261-274.

Hochschild, A., & Machung, A. (2012). *The second shift: Working families and the revolution at home*. Penguin.

Wachter, T. V. (2020). The persistent effects of initial labor market conditions for young adults and their sources. *Journal of Economic Perspectives*, *34*(4), 168-194.

AI Acknowlegement:

We have carefully reviewed and modified the AI-generated code to ensure it aligned with the project requirements and our understanding of the problems.