Geog111B Assignment 2

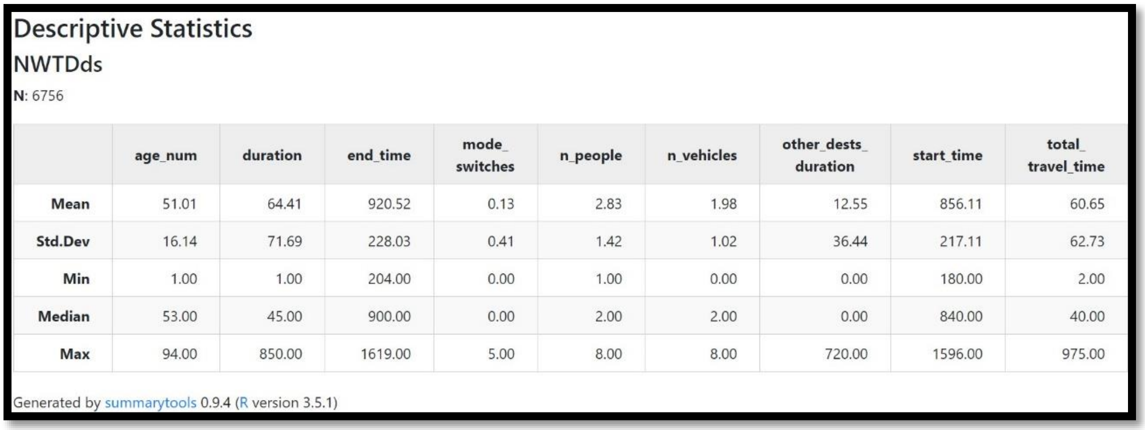
# Prediction of Category and Time Factors of Activities and Traveling Behaviors

Jiamin Tan

## Introduction

In this study, we are interested in whether some factors of a certain activity are related to the time variables and the type of the activity, and we also want to know if certain household factors can help people to predict the type of the activity. By building regression models using the NWTD_nodupes.rds data, we found that a linear regression can be used to describe the activity duration using end time of activities and types of activities. To predict the number of stops in a tour, a negative binomial model can be used knowing time factors and whether the trip is for commuting. In addition, to predict the type of the activity, knowing certain time factors and information about the household, the logit models for binary and multinominal activity types can be used. We can conclude that the variables describing the activity and travelling time is significantly related to category of the trips.

## Descriptive Statistics

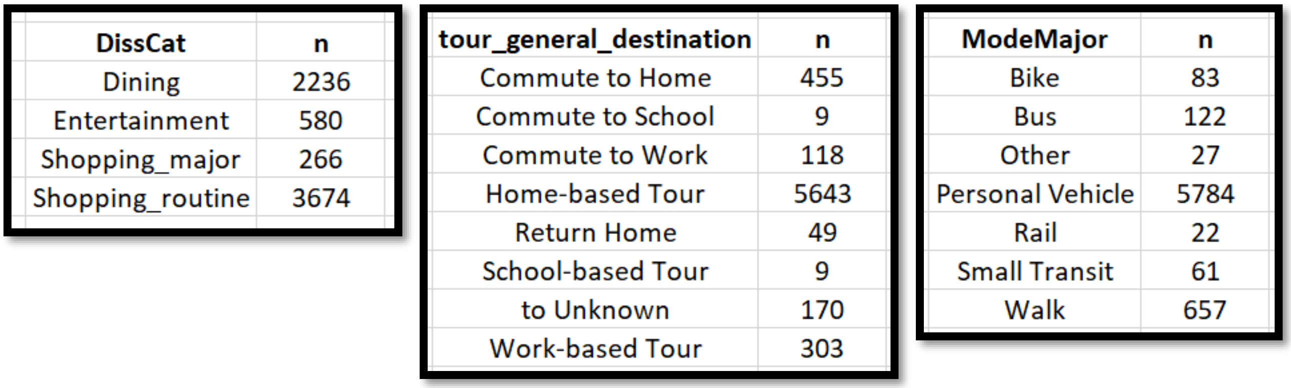


Descriptive Statistics
NWTDds
N: 6756

| | age_num | duration | end_time | mode_switches | n_people | n_vehicles | other_dests_duration | start_time | total_travel_time |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 51.01 | 64.41 | 920.52 | 0.13 | 2.83 | 1.98 | 12.55 | 856.11 | 60.65 |
| Std.Dev | 16.14 | 71.69 | 228.03 | 0.41 | 1.42 | 1.02 | 36.44 | 217.11 | 62.73 |
| Min | 1.00 | 1.00 | 204.00 | 0.00 | 1.00 | 0.00 | 0.00 | 180.00 | 2.00 |
| Median | 53.00 | 45.00 | 900.00 | 0.00 | 2.00 | 2.00 | 0.00 | 840.00 | 40.00 |
| Max | 94.00 | 850.00 | 1619.00 | 5.00 | 8.00 | 8.00 | 720.00 | 1596.00 | 975.00 |

Generated by summarytools 0.9.4 (R version 3.5.1)

**Table 1 – Descriptive Statistics of Some Variables**



| DissCat | n |
|---|---|
| Dining | 2236 |
| Entertainment | 580 |
| Shopping_major | 266 |
| Shopping_routine | 3674 |

| tour_general_destination | n |
|---|---|
| Commute to Home | 455 |
| Commute to School | 9 |
| Commute to Work | 118 |
| Home-based Tour | 5643 |
| Return Home | 49 |
| School-based Tour | 9 |
| to Unknown | 170 |
| Work-based Tour | 303 |

| ModeMajor | n |
|---|---|
| Bike | 83 |
| Bus | 122 |
| Other | 27 |
| Personal Vehicle | 5784 |
| Rail | 22 |
| Small Transit | 61 |
| Walk | 657 |

**Table 2,3,4 (Left to Right) – Descriptive Statistics of Categorical Variables**

From Table 1 above, we know that the average age all the respondents is about 51 years old with standard deviation of 16.14 years old, and there are 2.83 people and 1.98 cars in a household on average in this survey. The average duration of an activity is 64.41 minutes. The mean start time of an activity is around 2 pm, and the mean end time of an activity is around 3 pm. The average of total travel time is about an hour with standard deviation of 62.73 minutes, and on average people spend around 12 minutes to do other things during the trip. The average number of switch mode is 0.13. From Table 2, 3, and 4, we know that in the survey the most activity that people do is dining, and the most trips made are home based trips. People used the personal vehicles the most often and use the rail the least.

**Linear Regression Model**

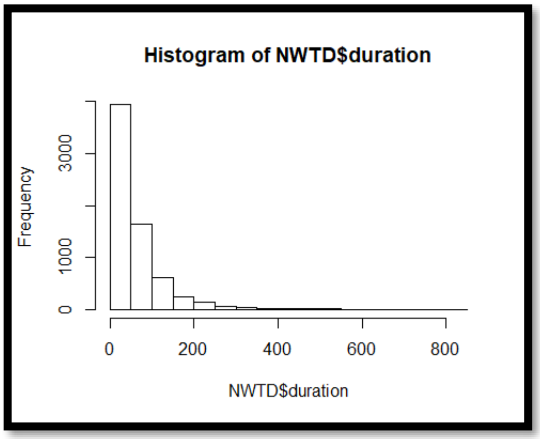


**Figure 1(a) – Histogram of duration**

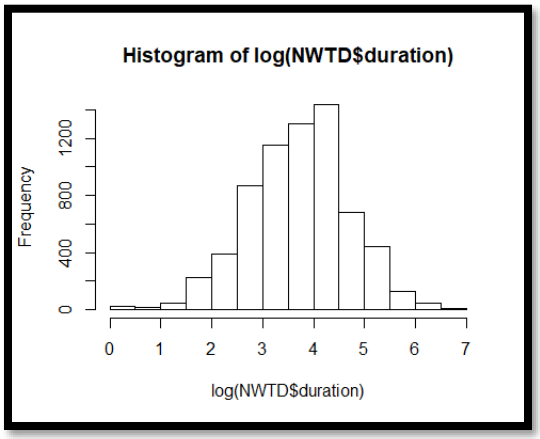
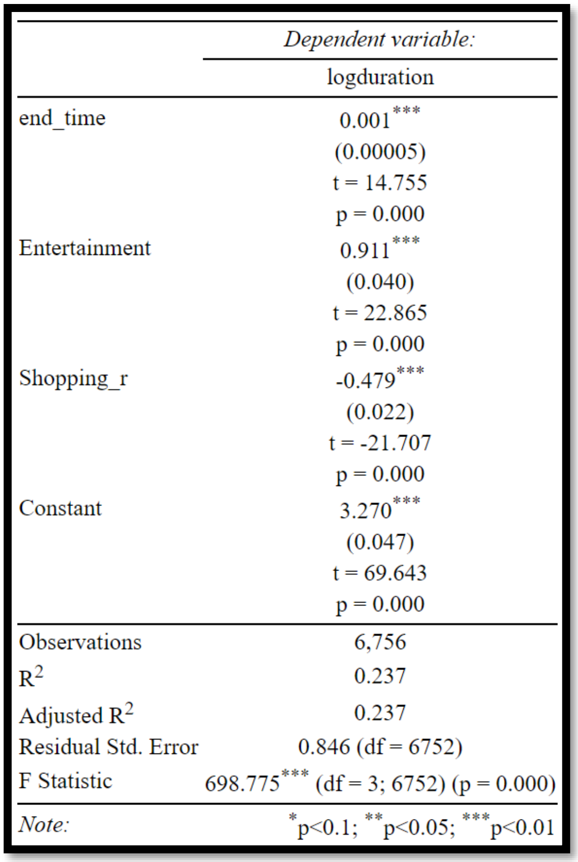


**Figure 1(b) – Histogram of log(duration)**

We want to find if there is any relationship between the dependent variable "duration" and independent variables "end_time" and "DissCat". For each categorical value from "DissCat", a dummy variable was created in the vector form <Entertainment, Shopping_routine, Dining> with binary digits. Specifically, Entertainment is represented by <1,0,0>; Shopping_routine is represented by <0,1,0>; Dining is represented by <0,0,1>; and Shopping_major is represented by <0,0,0>. By examining the distribution of "duration", we found that the histogram looks like a log normal distribution while distributions of other variables generally follow the normal distribution. Therefore, we transform the dependent variable with log transformation, and the linear regression model shows the relationship of log(duration).



| | *Dependent variable:* |
|---|---|
| | logduration |
| end_time | $0.001^{***}$ |
| | (0.00005) |
| | t = 14.755 |
| | p = 0.000 |
| Entertainment | $0.911^{***}$ |
| | (0.040) |
| | t = 22.865 |
| | p = 0.000 |
| Shopping_r | $-0.479^{***}$ |
| | (0.022) |
| | t = -21.707 |
| | p = 0.000 |
| Constant | $3.270^{***}$ |
| | (0.047) |
| | t = 69.643 |
| | p = 0.000 |
| Observations | 6,756 |
| $R^2$ | 0.237 |
| Adjusted $R^2$ | 0.237 |
| Residual Std. Error | 0.846 (df = 6752) |
| F Statistic | $698.775^{***}$ (df = 3; 6752) (p = 0.000) |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ |

**Table 5 – Final Result of the Linear Regression Model**

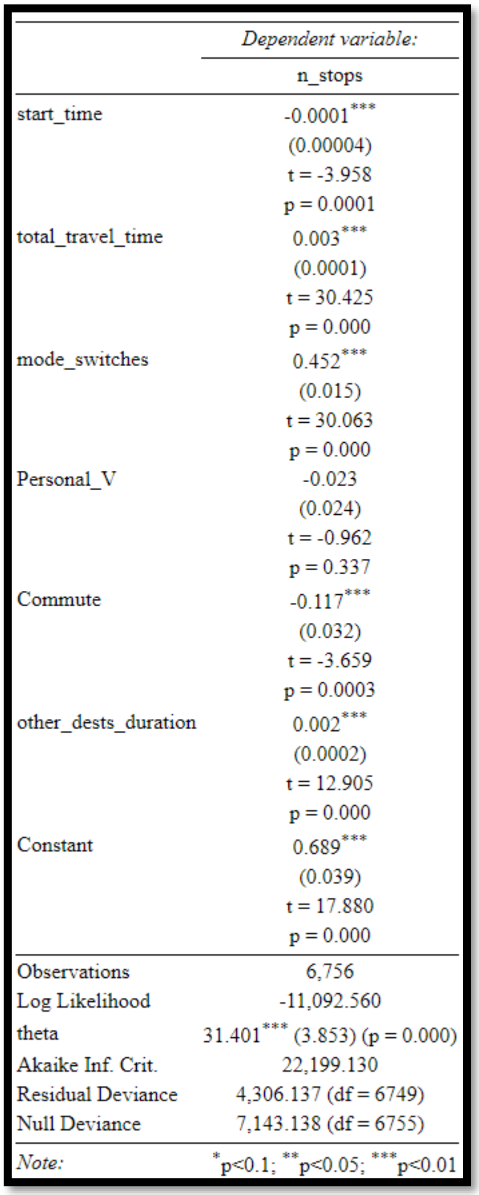
The variable "Dining" was proved to be insignificant in the fitted linear regression model. The specific result of that model is not covered in this report, but it is recorded in the RMarkdown file in the Appendix. Thus, the final linear model is written as:

**log(duration) = 0.001end_time + 0.911Entertainment - 0.479Shopping_routine + 3.270**

All the independent variables and the constant in this equation are significant at the significance level of 0.05. A one-unit increase in the variable "end_time" will cause the value of log(duration) to increase 0.001. A trip categorized as "Entertainment" will cause the value of log(duration) to increase 0.911, while a trip categorized as "Shopping_routine" will cause the value of log(duration) to decrease 0.479. The result obeys common sense that the later the activity end, the longer the activity lasts. Also, entertainments such as watching a film generally take longer time than a routine grocery shopping. The $R^2$ value is 0.237 which means our model explains 23.7% of the variation of the log(duration) around the mean. We consider this model is at least better than the NULL model which explains less than 10% of the variation of the response variable around the mean.

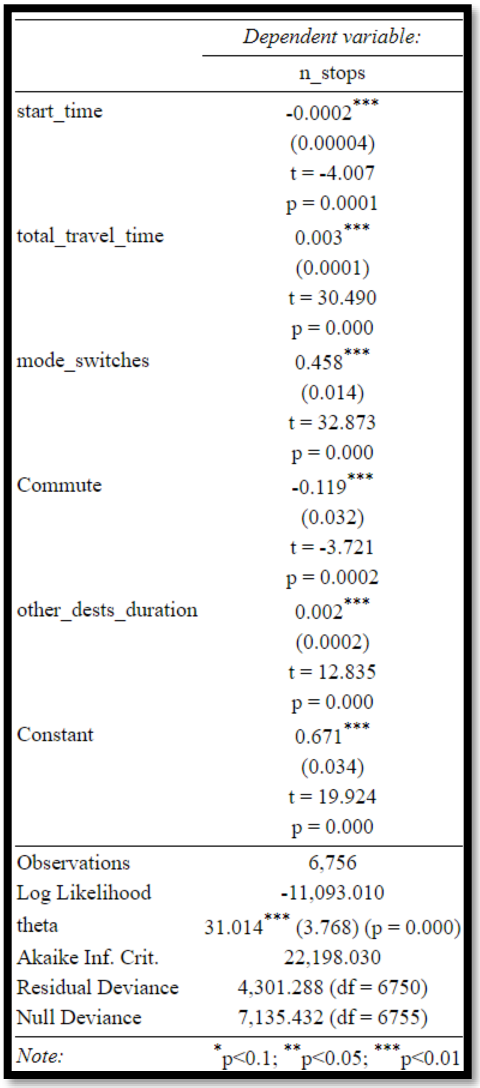
**Negative Binomial Regression Model for Count Data**

In this section, we want to quantify the relationship between the dependent variable “n_stops” and independent variables start time of the activity, total travel time, times of mode switches, whether the respondent use a personal vehicle, whether it is a commuting trip, and time took for other activities during the trip. We first defined dummy variables for if the respondent used a personal vehicle or other modes, and if the trip is for commuting. All the trip traveled by “Personal_vehicle” is represented by 1, and trips traveled by any other modes listed in the “ModeMajor” are represented by 0. Trips categorized as “Commute to Work/School/Home” are defined as 1 in the dummy variable “Commute”, and others are defined as 0.



| | *Dependent variable:* |
|---|---|
| | n_stops |
| start_time | $-0.0001^{***}$ |
| | (0.00004) |
| | t = -3.958 |
| | p = 0.0001 |
| total_travel_time | $0.003^{***}$ |
| | (0.0001) |
| | t = 30.425 |
| | p = 0.000 |
| mode_switches | $0.452^{***}$ |
| | (0.015) |
| | t = 30.063 |
| | p = 0.000 |
| Personal_V | -0.023 |
| | (0.024) |
| | t = -0.962 |
| | p = 0.337 |
| Commute | $-0.117^{***}$ |
| | (0.032) |
| | t = -3.659 |
| | p = 0.0003 |
| other_dests_duration | $0.002^{***}$ |
| | (0.0002) |
| | t = 12.905 |
| | p = 0.000 |
| Constant | $0.689^{***}$ |
| | (0.039) |
| | t = 17.880 |
| | p = 0.000 |
| Observations | 6,756 |
| Log Likelihood | -11,092.560 |
| theta | $31.401^{***}$ (3.853) (p = 0.000) |
| Akaike Inf. Crit. | 22,199.130 |
| Residual Deviance | 4,306.137 (df = 6749) |
| Null Deviance | 7,143.138 (df = 6755) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table 6(a) – Original Negative Binomial Regression Model**



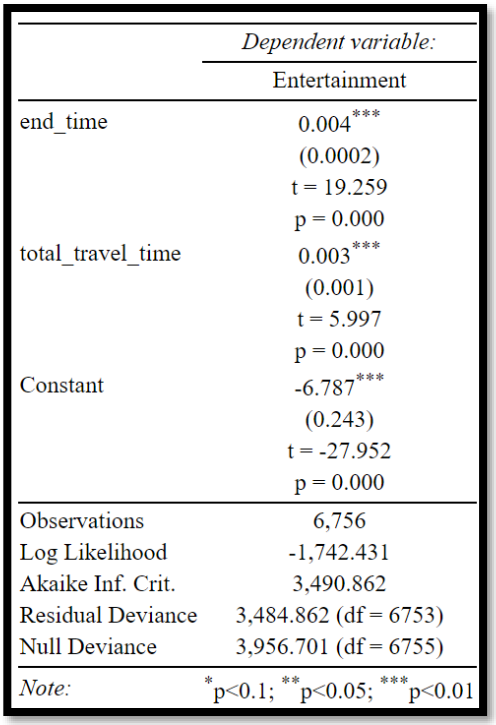
| | *Dependent variable:* |
|---|---|
| | n_stops |
| start_time | $-0.0002^{***}$ |
| | (0.00004) |
| | t = -4.007 |
| | p = 0.0001 |
| total_travel_time | $0.003^{***}$ |
| | (0.0001) |
| | t = 30.490 |
| | p = 0.000 |
| mode_switches | $0.458^{***}$ |
| | (0.014) |
| | t = 32.873 |
| | p = 0.000 |
| Commute | $-0.119^{***}$ |
| | (0.032) |
| | t = -3.721 |
| | p = 0.0002 |
| other_dests_duration | $0.002^{***}$ |
| | (0.0002) |
| | t = 12.835 |
| | p = 0.000 |
| Constant | $0.671^{***}$ |
| | (0.034) |
| | t = 19.924 |
| | p = 0.000 |
| Observations | 6,756 |
| Log Likelihood | -11,093.010 |
| theta | $31.014^{***}$ (3.768) (p = 0.000) |
| Akaike Inf. Crit. | 22,198.030 |
| Residual Deviance | 4,301.288 (df = 6750) |
| Null Deviance | 7,135.432 (df = 6755) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table 6(b) – Trimmed Negative Binomial Regression Model**

Table 6(a) shows that all the independent variables except “Personal_V” (whether the respondent used a personal vehicle for the trip or not) are significant at the significance level 0.05. Therefore, we trimmed the model by leaving out “Personal_V”, and we can see that in Table 6(b), all the independent variables are significantly different from 0 at the significance level of 0.05. We then perform a likelihood ratio test to see if the two models are significantly different from each other, and the p-value derived from the test is 0.3426 which is greater than the significance level of 0.05. Therefore, we fail to reject the hypothesis, and we can conclude that the two models are not significantly different. Given that the log likelihood of the two models are almost the same (-11092.560 vs -11093.010), the trimmed model is slightly better because it has fewer estimators than the original model.

To interpret the coefficients of the trimmed model above, we look at the table 6(b). If there is a one-unit increase in start time of the activity, the log count of the number of stops during the trip will decrease by 0.0002. If there is a one-unit increase in total travel time, number of mode switches, and total time of other activities in the tour, the log count of the number of stops will increase by 0.003, 0.458, and 0.002 respectively. In terms of the dummy variable “Commute”, the expected log count of number of stops for commuting trips is 0.119 lower than non-commuting trips. This makes sense because people tend to go from the origin to the destination when commuting which is unlikely for them to make several stops.

### Binary Regression Model Using Logit

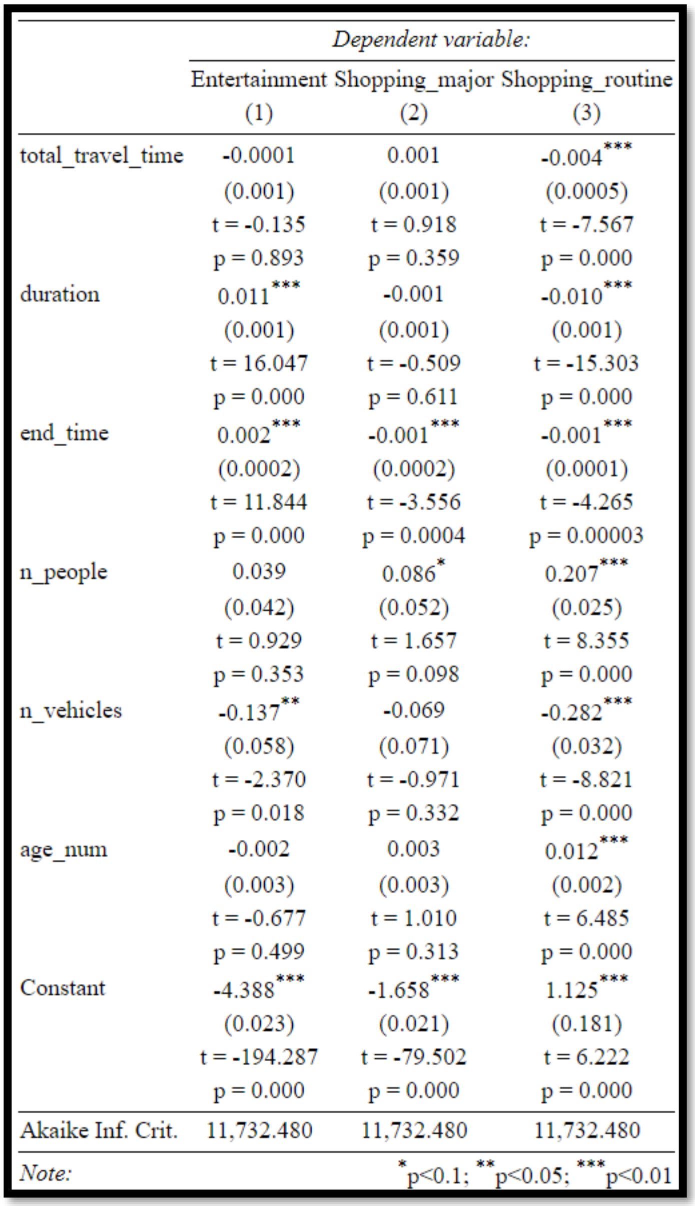


| | *Dependent variable:* |
|---|---|
| | Entertainment |
| end_time | $0.004^{***}$ |
| | (0.0002) |
| | t = 19.259 |
| | p = 0.000 |
| total_travel_time | $0.003^{***}$ |
| | (0.001) |
| | t = 5.997 |
| | p = 0.000 |
| Constant | $-6.787^{***}$ |
| | (0.243) |
| | t = -27.952 |
| | p = 0.000 |
| Observations | 6,756 |
| Log Likelihood | -1,742.431 |
| Akaike Inf. Crit. | 3,490.862 |
| Residual Deviance | 3,484.862 (df = 6753) |
| Null Deviance | 3,956.701 (df = 6755) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table 7 – Result of Binary Regression Model Using Logit**

We are now interested in whether the end time and the total travel time of an activity are decisive for the type, entertainment in this case, of activity that people participate. As described in the linear regression section at the beginning, in the "Entertainment" dummy variable, 1 is defined for entertainment activities, and 0 is defined as other types of activities. There is not trimmed model in this section, because we can see from the table 7 above that obviously, both of the independent variables are significantly different from 0 at the significance level of 0.05. The likelihood ration test comparing this model to the null model gives a p-value less than 2.2 x $10^{-16}$, so we reject the null hypothesis and conclude that this model is better than the null model with a higher likelihood value (-1742.4 vs -1978.3). To interpret the coefficients, if there is a one-unit increase in "end_time", then the log odds of "Entertainment" will increase by 0.004. Similarly, if there is a one-unit increase in "total_travel_time", then the log odds of "Entertainment" will increase by 0.003. In other words, as the end time and/or the total travel time of an activity increases, the probability of this activity categorized as entertainment increases.

## Multinomial Logit Regression Model



| | *Dependent variable:* | | |
|---|---|---|---|
| | Entertainment | Shopping_major | Shopping_routine |
| | (1) | (2) | (3) |
| total_travel_time | -0.0001 | 0.001 | -0.004*** |
| | (0.001) | (0.001) | (0.0005) |
| | t = -0.135 | t = 0.918 | t = -7.567 |
| | p = 0.893 | p = 0.359 | p = 0.000 |
| duration | 0.011*** | -0.001 | -0.010*** |
| | (0.001) | (0.001) | (0.001) |
| | t = 16.047 | t = -0.509 | t = -15.303 |
| | p = 0.000 | p = 0.611 | p = 0.000 |
| end_time | 0.002*** | -0.001*** | -0.001*** |
| | (0.0002) | (0.0002) | (0.0001) |
| | t = 11.844 | t = -3.556 | t = -4.265 |
| | p = 0.000 | p = 0.0004 | p = 0.00003 |
| n_people | 0.039 | 0.086* | 0.207*** |
| | (0.042) | (0.052) | (0.025) |
| | t = 0.929 | t = 1.657 | t = 8.355 |
| | p = 0.353 | p = 0.098 | p = 0.000 |
| n_vehicles | -0.137** | -0.069 | -0.282*** |
| | (0.058) | (0.071) | (0.032) |
| | t = -2.370 | t = -0.971 | t = -8.821 |
| | p = 0.018 | p = 0.332 | p = 0.000 |
| age_num | -0.002 | 0.003 | 0.012*** |
| | (0.003) | (0.003) | (0.002) |
| | t = -0.677 | t = 1.010 | t = 6.485 |
| | p = 0.499 | p = 0.313 | p = 0.000 |
| Constant | -4.388*** | -1.658*** | 1.125*** |
| | (0.023) | (0.021) | (0.181) |
| | t = -194.287 | t = -79.502 | t = 6.222 |
| | p = 0.000 | p = 0.000 | p = 0.000 |
| Akaike Inf. Crit. | 11,732.480 | 11,732.480 | 11,732.480 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

**Table 8 – Result of Multinominal Logit Regression**

We want to find out that if the category of an activity can be determined by total travel time, duration, end time of the activity, number of people and vehicles in the household, and the age of the respondent. The reference category for this model is "Dining". There is no trimmed model in this section because we can see from Table 8 that for the dependent variable Shopping_routine, all the independent variables are significantly different from 0 at the significance level of 0.05. The likelihood ration test gives a p-value less than 2.2 x $10^{-16}$, so we reject the null hypothesis and conclude that this model is at least better than the null model with a greater likelihood value (-5845.2 vs -6680.6). To interpret the coefficients, we will look at “total_travel_time”, “duration”, “end_time”, “n_people”, “n_vehicles”, and “age_num”. A one-unit increase in “total_travel_time” will lead to a) a decrease of log odds of “Entertainment” by 0.0001 comparing to “Dining”; b) an increase of log odds of “Shopping_major” by 0.001 comparing to “Dining”; and c) a decrease of log odds of “Shopping_routine” by 0.004 comparing to “Dining”. A one-unit increase in “duration” will lead to a) an increase of log odds of “Entertainment” by 0.011 comparing to “Dining”; b) a decrease of log odds of “Shopping_major” by 0.001 comparing to “Dining”; and c) a decrease of log odds of “Shopping_routine” by 0.010 comparing to “Dining”. A one-unit increase in “end_time” will lead to a) an increase of log odds of “Entertainment” by 0.002 comparing to “Dining”; b) a decrease of log odds of “Shopping_major” by 0.001 comparing to “Dining”; and c) a decrease of log odds of “Shopping_routine” by 0.001 comparing to “Dining”. A one-unit increase in “n_people” will lead to a) an increase of log odds of “Entertainment” by 0.039 comparing to “Dining”; b) an increase of log odds of “Shopping_major” by 0.086 comparing to “Dining”; and c) an increase of log odds of “Shopping_routine” by 0.207 comparing to “Dining”. Lastly, a one-unit increase in “n_vehicle” will lead to a) a decrease of log odds of “Entertainment” by 0.137 comparing to “Dining”; b) a decrease of log odds of “Shopping_major” by 0.069 comparing to “Dining”; and c) a decrease of log odds of “Shopping_routine” by 0.282 comparing to “Dining”.

In other word, comparing to “Dining” an increase in duration, end time of the activity and the number of people in the household will lead to larger chance of an entertainment activity. An increase in total travel time and number of people in the household will lead to greater chance of a major shopping activity, and the increase of the number of people in the household will increase the chance of a routine shopping activity.

**Conclusion**

From studies above, we know that a linear regression model can be used to predict the duration of activities using end time and the type of the activity. A negative binomial model can predict the count of stops during a trip also through variables describing the time and categories of the activity or the trip. Lastly, we can predict the types of activities by knowing certain information of not only the about the travelling time and categories but also the information of the household that the respondent comes from. We can see that the activity time and categories are highly related to each other and knowing one of them can be useful for predicting the other.

**Appendix**

# Tan_Jiamin_Geog111B_Lab2

Jiamin Tan

February 24, 2020

```
library(tidyverse)
library(stargazer)
library(ggfortify)
library(sandwich)
library(lmtest)
library(MASS)
library(lmtest)
library(summarytools)
library(nnet)
library(reshape2)

NWTD <- (readRDS("NWTD_nodupes.rds"))
```

### *Descriptive Statistics*

```
# duration, end_time, start_time, total_travel_time, mode_switches,
# n_people, n_vehicles, age_num, other_dests_duration
# Disscat, ModeMajor, tour_general_destination
NWTDds <- NWTD %>%
  dplyr::select(duration, end_time, start_time, total_travel_time, mode_switc
hes,n_people, n_vehicles, age_num, other_dests_duration)

stats_table <- descr(NWTDds, stats = c("mean", "sd",
                                       "min", "med", "max"))
view(stats_table)

## Switching method to 'browser'

## Output file written: C:\Users\JIAMIN~1\AppData\Local\Temp\RtmpABy6pi\filef
924654d64f8.html

as.data.frame(summary(as.factor(NWTD$DissCat)))

##                    summary(as.factor(NWTD$DissCat))
## Dining                                          2236
## Entertainment                                    580
## Shopping_major                                   266
## Shopping_routine                                3674

as.data.frame(summary(as.factor(NWTD$ModeMajor)))

##                    summary(as.factor(NWTD$ModeMajor))
## Bike                                                83
## Bus                                             122
## Other                                            27
## Personal Vehicle                               5784
## Rail                                             22
## Small Transit                                    61
## Walk                                            657
```

```
as.data.frame(summary(as.factor(NWTD$tour_general_destination)))
```

```
##                   summary(as.factor(NWTD$tour_general_destination))
## Commute to Home                                                 455
## Commute to School                                                 9
## Commute to Work                                                 118
## Home-based Tour                                                5643
## Return Home                                                      49
## School-based Tour                                                 9
## to Unknown                                                      170
## Work-based Tour                                                 303
```

### *1. Linear Regression Model*

```
# simple linear regression model with end_time only
lreg.mod1 <- lm(duration ~ end_time, data = NWTD)
# variable significance and R sqaure
summary(lreg.mod1) ## R-squared = 0.09359
```

```
## 
## Call:
## lm(formula = duration ~ end_time, data = NWTD)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -125.49  -38.21  -13.98   17.96  751.50 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -24.121212   3.453747  -6.984 3.14e-12 ***
## end_time      0.096174   0.003642  26.408  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 68.25 on 6754 degrees of freedom
## Multiple R-squared:  0.09359,    Adjusted R-squared:  0.09345 
## F-statistic: 697.4 on 1 and 6754 DF,  p-value: < 2.2e-16
```

```
# is this SLR model better than the NULL model(interception)?
anova(lreg.mod1) ## p-value < 0.05, it is better
```

```
## Analysis of Variance Table
## 
## Response: duration
##              Df   Sum Sq Mean Sq F value    Pr(>F)    
## end_time     1  3248825 3248825  697.37 < 2.2e-16 ***
## Residuals 6754 31464878    4659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# try to add categorical variables of activity type
unique(NWTD$DissCat) ## need 4 - 1 = 3 dummy variables
```

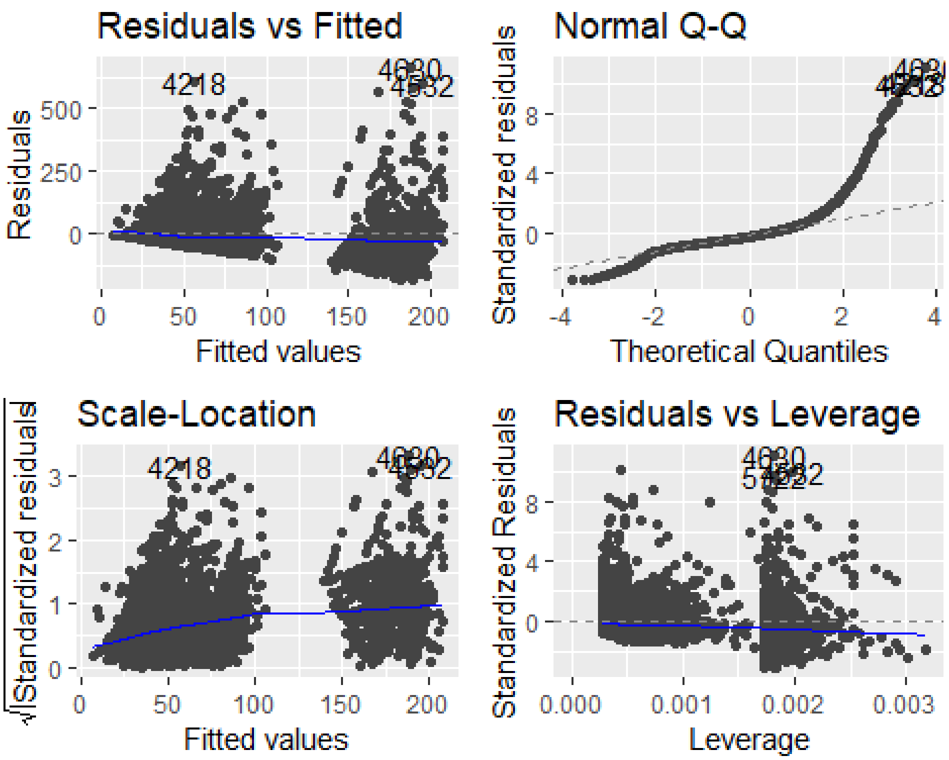
```
## [1] "Entertainment"    "Shopping_routine" "Dining"
## [4] "Shopping_major"
```

```
# is there missing values?
any(is.na(NWTD$DissCat))
```

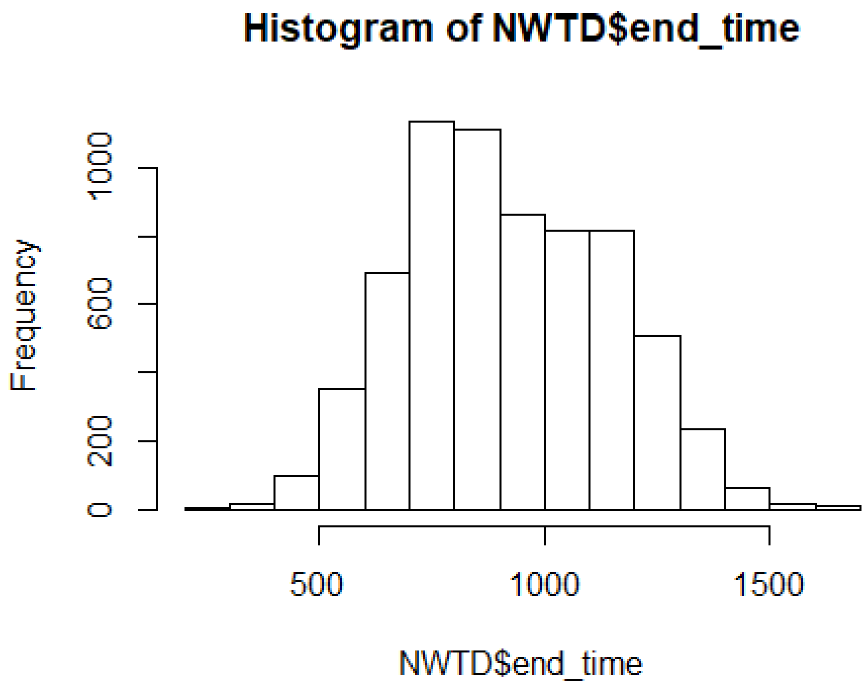
```
## [1] FALSE
```

```
# dummy variable 1: Entertainment
NWTD$Entertainment <- ifelse(NWTD$DissCat =='Entertainment', 1, 0)
# dummy variable 2: Shopping_r
NWTD$Shopping_r <- ifelse(NWTD$DissCat =='Shopping_routine', 1, 0)
# dummy variable 3: Dining
NWTD$Dining <- ifelse(NWTD$DissCat =='Dining', 1, 0)
## Shopping_major will have the value 0,0,0 in this case
# fit a multi-variable linear regression model using variables above
lreg.mod2 <- lm(duration ~ end_time + Entertainment + Shopping_r + Dining, da
ta = NWTD)
# variable significance and R sqaure
summary(lreg.mod2) ## R-squared = 0.292, much better than SLR
```

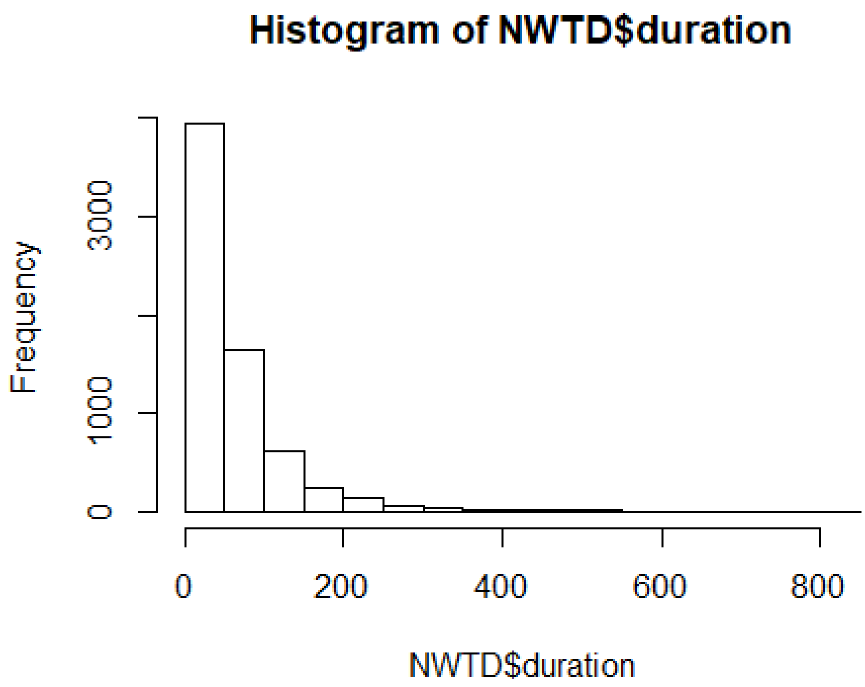
```
##
## Call:
## lm(formula = duration ~ end_time + Entertainment + Shopping_r +
##     Dining, data = NWTD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -181.38  -30.28  -11.68   14.26  662.03
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.459580   4.761559   3.457  0.00055 ***
## end_time       0.056447   0.003351  16.847  < 2e-16 ***
## Entertainment 99.537833   4.525544  21.995  < 2e-16 ***
## Shopping_r   -22.114232   3.831439  -5.772 8.19e-09 ***
## Dining        -1.602474   3.916194  -0.409  0.68241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.34 on 6751 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.2915
## F-statistic: 695.9 on 4 and 6751 DF,  p-value: < 2.2e-16
## Dining is not significant, so we remove it and fit the MLR again
lreg.mod3 <- lm(duration ~ end_time + Entertainment + Shopping_r, data = NWTD
)
summary(lreg.mod3) ## R-squared = 0.2919, equivilant to the one above

##
## Call:
## lm(formula = duration ~ end_time + Entertainment + Shopping_r,
##     data = NWTD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -181.37  -30.27  -11.69   14.24  662.04
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.075000   3.349876    4.50  6.9e-06 ***
## end_time        0.056396   0.003348   16.84  < 2e-16 ***
## Entertainment 100.978883   2.842040   35.53  < 2e-16 ***
## Shopping_r    -20.684745   1.573369  -13.15  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.34 on 6752 degrees of freedom
## Multiple R-squared:  0.2919, Adjusted R-squared:  0.2916
## F-statistic: 927.9 on 3 and 6752 DF,  p-value: < 2.2e-16

## all the variables are significant in this model
# is this model better than the NULL model?
anova(lreg.mod3) ## p-value < 0.05, it is better

## Analysis of Variance Table
##
## Response: duration
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## end_time         1  3248825 3248825  892.45 < 2.2e-16 ***
## Entertainment    1  6256135 6256135 1718.56 < 2.2e-16 ***
## Shopping_r       1   629189  629189  172.84 < 2.2e-16 ***
## Residuals     6752 24579554    3640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# diagnostics
autoplot(lreg.mod3) ## normality severely violated, others are okay
```



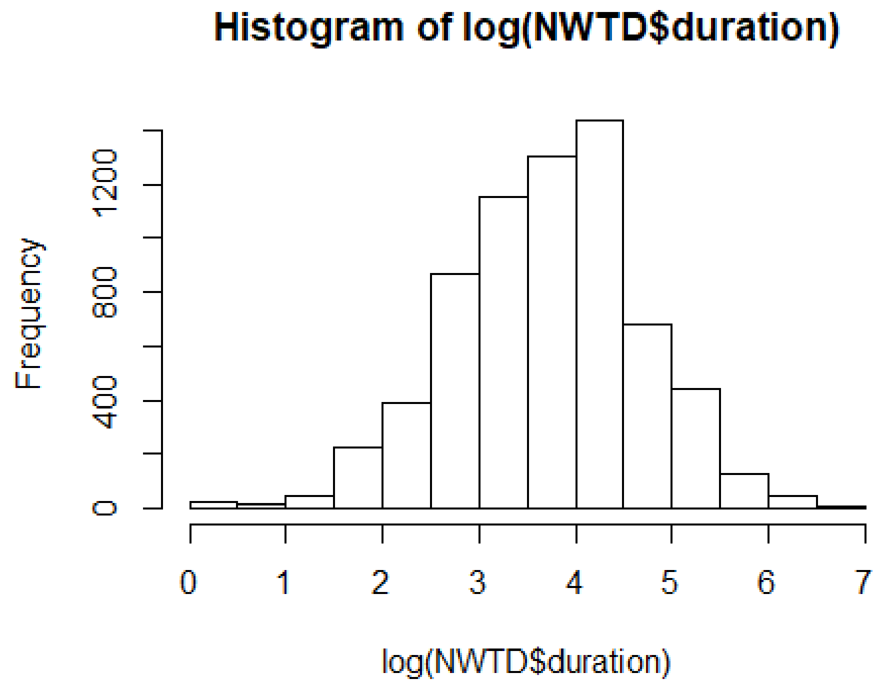
```r
# distribution of the continous
# end_time
hist(NWTD$end_time) ## normality looks okay
```



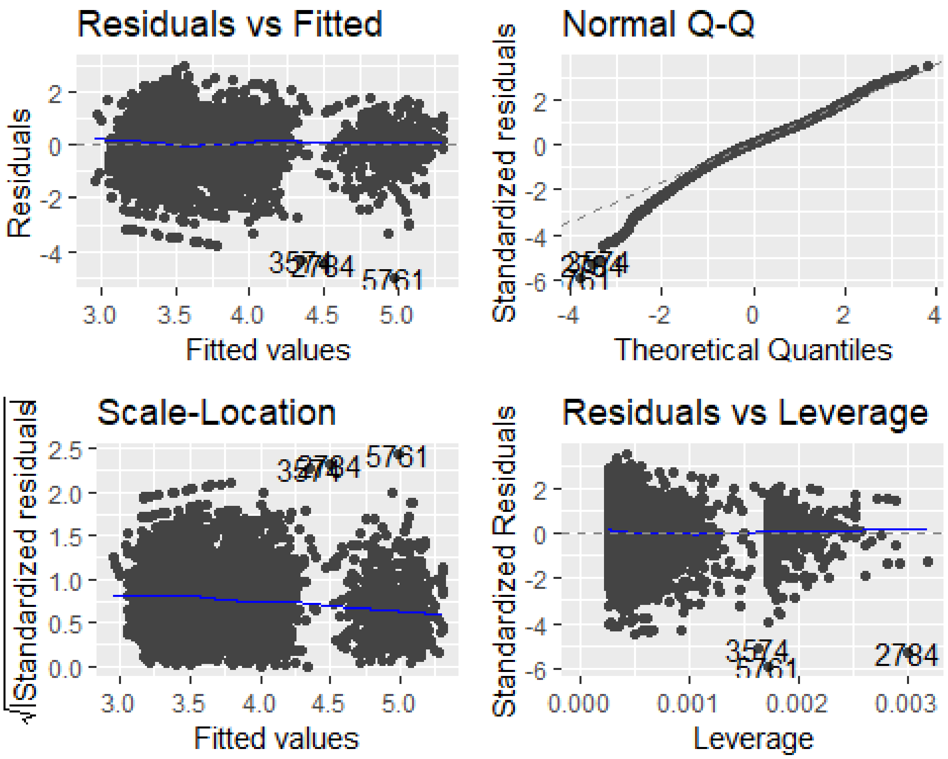
```
# duration
hist(NWTD$duration) ## looks like log normal!
```



```
hist(log(NWTD$duration)) ## consider log transformation
```



```
# fit with log(duration)
NWTD$logduration <- log(NWTD$duration)
lreg.mod4 <- lm(logduration ~ end_time + Entertainment + Shopping_r, data = N
WTD)
summary(lreg.mod4) ## R-squared = 0.2369, equivilant to the one above

##
## Call:
## lm(formula = logduration ~ end_time + Entertainment + Shopping_r,
##     data = NWTD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9811 -0.4571  0.0577  0.5343  2.9400
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.270e+00  4.695e-02   69.64   <2e-16 ***
## end_time       6.924e-04  4.693e-05   14.76   <2e-16 ***
## Entertainment  9.108e-01  3.983e-02   22.86   <2e-16 ***
## Shopping_r    -4.787e-01  2.205e-02  -21.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8457 on 6752 degrees of freedom
## Multiple R-squared:  0.2369, Adjusted R-squared:  0.2366
## F-statistic: 698.8 on 3 and 6752 DF,  p-value: < 2.2e-16
## all the variables are significant in this model
# is this model better than the NULL model?
anova(lreg.mod4) ## p-value < 0.05, it is better

## Analysis of Variance Table
##
## Response: logduration
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## end_time         1  480.0  480.00  671.20 < 2.2e-16 ***
## Entertainment    1  682.2  682.20  953.93 < 2.2e-16 ***
## Shopping_r       1  337.0  336.98  471.20 < 2.2e-16 ***
## Residuals     6752 4828.7    0.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# diagnostic
autoplot(lreg.mod4) ## linearity and normality imporved!
```



```
# table using stargazer
#lreg.table <- stargazer(lreg.mod4, out="lregmod4.html", style="all")
```

### 2. Count Data Regression

```
# define dummy variables for Personal_Vehicle
unique(NWTD$ModeMajor)
```

```
## [1] "Personal Vehicle" "Bus"              "Walk"
## [4] "Small Transit"    "Rail"             "Other"
## [7] "Bike"
```

```
NWTD$Personal_V <- ifelse(NWTD$ModeMajor =='Personal Vehicle', 1, 0)
NWTD$Personal_V <- ifelse(is.na(NWTD$ModeMajor), 0, NWTD$Personal_V)
# define dummy variables for Commuting
unique(NWTD$tour_general_destination)
```

```
## [1] "Home-based Tour"   "Commute to Work"   "Work-based Tour"
## [4] "Commute to Home"   "Commute to School" "to Unknown"
## [7] "Return Home"       "School-based Tour"
```

```
NWTD$Commute <- ifelse(NWTD$tour_general_destination == 'Commute to Work' | N
WTD$tour_general_destination =='Commute to Home' | NWTD$tour_general_destinat
ion == 'Commute to School', 1, 0)
NWTD$Commute <- ifelse(is.na(NWTD$tour_general_destination), 0, NWTD$Commute)

# fit data with negative binomial regression
nbreg.mod1 <- glm.nb(n_stops ~ start_time + total_travel_time + mode_switches
+ Personal_V + Commute + other_dests_duration, data = NWTD)
```

```
## Warning in glm.nb(n_stops ~ start_time + total_travel_time + mode_switches
## + : alternation limit reached
```

```
# summary
summary(nbreg.mod1) ## Personal_V is insignificant
```

```
##
## Call:
## glm.nb(formula = n_stops ~ start_time + total_travel_time + mode_switches
+
##     Personal_V + Commute + other_dests_duration, data = NWTD,
##     init.theta = 31.40116233, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.7164  -0.6614  -0.1573   0.3346   4.3260
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           6.888e-01  3.853e-02  17.880  < 2e-16 ***
## start_time           -1.488e-04  3.760e-05  -3.958 7.55e-05 ***
## total_travel_time     2.863e-03  9.409e-05  30.425  < 2e-16 ***
## mode_switches         4.518e-01  1.503e-02  30.063  < 2e-16 ***
## Personal_V           -2.275e-02  2.366e-02  -0.962 0.336216
## Commute              -1.169e-01  3.195e-02  -3.659 0.000254 ***
## other_dests_duration  2.047e-03  1.586e-04  12.905  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for Negative Binomial(31.4012) family taken to be 1)
## 
##     Null deviance: 7143.1  on 6755  degrees of freedom
## Residual deviance: 4306.1  on 6749  degrees of freedom
## AIC: 22199
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##                Theta:  31.40 
##            Std. Err.:  3.85 
## Warning while fitting theta: alternation limit reached 
## 
##  2 x log-likelihood:  -22183.13
```

```
#nbreg.table1 <- stargazer(nbreg.mod1, out="nbregmod1.html", style="all")

# fit data without Personal_V
nbreg.mod2 <- glm.nb(n_stops ~ start_time + total_travel_time + mode_switches
+ Commute + other_dests_duration, data = NWTD)
```

```
## Warning in glm.nb(n_stops ~ start_time + total_travel_time + mode_switches
## + : alternation limit reached
```

```
# summary
summary(nbreg.mod2) ## Personal_V is insignificant
```

```
## 
## Call:
## glm.nb(formula = n_stops ~ start_time + total_travel_time + mode_switches
+ 
##     Commute + other_dests_duration, data = NWTD, init.theta = 31.01423378, 
##     link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -8.7317  -0.6595  -0.1602   0.3310   4.3183  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           6.707e-01  3.366e-02  19.924  < 2e-16 ***
## start_time           -1.506e-04  3.758e-05  -4.007 6.14e-05 ***
## total_travel_time     2.858e-03  9.373e-05  30.490  < 2e-16 ***
## mode_switches         4.578e-01  1.393e-02  32.873  < 2e-16 ***
## Commute              -1.187e-01  3.191e-02  -3.721 0.000198 ***
## other_dests_duration  2.036e-03  1.587e-04  12.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(31.0142) family taken to be 1)
## 
```

```
##     Null deviance: 7135.4  on 6755  degrees of freedom
## Residual deviance: 4301.3  on 6750  degrees of freedom
## AIC: 22198
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  31.01 
##           Std. Err.:  3.77 
## Warning while fitting theta: alternation limit reached 
## 
##  2 x log-likelihood:  -22184.03
```

```
# is the model better than NULL model?
lrtest(nbreg.mod2) ## yes it is
```

```
## Likelihood ratio test
## 
## Model 1: n_stops ~ start_time + total_travel_time + mode_switches + Commut
e + 
##     other_dests_duration
## Model 2: n_stops ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)    
## 1   7 -11092                         
## 2   2 -12414 -5 2643.9  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(nbreg.mod2, nbreg.mod1) ## fewer variables
```

```
## Likelihood ratio test
## 
## Model 1: n_stops ~ start_time + total_travel_time + mode_switches + Commut
e + 
##     other_dests_duration
## Model 2: n_stops ~ start_time + total_travel_time + mode_switches + Person
al_V + 
##     Commute + other_dests_duration
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   7 -11092                     
## 2   8 -11092  1 0.9008     0.3426
```

```
# stargazer
#nbreg.table <- stargazer(nbreg.mod2, out="nbregmod2.html", style="all")
```

### *3. Binary Regression Model*

```
blreg.mod1 <- glm(Entertainment ~ end_time + total_travel_time, family = bino
mial(link = "logit"), data = NWTD)
summary(blreg.mod1)
```

```
##
## Call:
## glm(formula = Entertainment ~ end_time + total_travel_time, family = binom
ial(link = "logit"),
##     data = NWTD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1766  -0.4628  -0.3083  -0.2186   3.0734
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.7873948  0.2428197 -27.952  < 2e-16 ***
## end_time           0.0041615  0.0002161  19.259  < 2e-16 ***
## total_travel_time  0.0034277  0.0005716   5.997 2.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3956.7  on 6755  degrees of freedom
## Residual deviance: 3484.9  on 6753  degrees of freedom
## AIC: 3490.9
##
## Number of Fisher Scoring iterations: 6
```

```
#blreg.table <- stargazer(blreg.mod1, out = "blregmod1.html", style = "all")
lrtest(blreg.mod1)
```

```
## Likelihood ratio test
##
## Model 1: Entertainment ~ end_time + total_travel_time
## Model 2: Entertainment ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -1742.4
## 2   1 -1978.3 -2 471.84  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### *4. Multinomial Logit Model*

```
mnreg.mod1 <- multinom(DissCat ~ total_travel_time + duration + end_time + n_
people + n_vehicles + age_num, data = NWTD, hessian = TRUE)
```

```
## # weights:  32 (21 variable)
## initial  value 8966.551928
## iter  10 value 6155.766701
## iter  20 value 5997.962377
## iter  30 value 5845.241930
## final  value 5845.241860
## converged
```

```
#mnreg.table <- stargazer(mnreg.mod1, out="mnregmod1.html", style="all")

lrtest(mnreg.mod1)
```

```
## # weights:  8 (3 variable)
## initial  value 9365.804704
## iter  10 value 6994.888571
## final  value 6994.876522
## converged
## # weights:  8 (3 variable)
## initial  value 8966.551928
## iter  10 value 6680.575437
## final  value 6680.561917
## converged
```

```
## Likelihood ratio test
##
## Model 1: DissCat ~ total_travel_time + duration + end_time + n_people +
##     n_vehicles + age_num
## Model 2: DissCat ~ 1
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  21 -5845.2
## 2   3 -6680.6 -18 1670.6  < 2.2e-16 ***
## ---
## Signif.
```