Assignment 2 – Regression Analysis

22 Winter GEOG 111B

Matthew Mangawang

Objectives

The main learning objective for this assignment is to understand how to build regression models, assess their performance, and extract behavioral indications. You are required to use the R code from the labs in this course and build your own analysis using your code. Examples of the R code used in the lecture are also posted on Gauchospace.

Assignment Description

Use the data in `NWTD_nodupes.rds`. You will estimate:

- a linear regression model,
- a count data regression model,
- a binary regression model using Logit, and
- a multicategory model using Multinomial Logit.

1. For linear regression use as dependent variable `duration` (this is the amount of time in an activity). Use as explanatory variables any other variable in the file.

- In the interpretation part of linear regression, you need to discuss:
  - what is the equation of your linear regression model? If you use any symbols in your equation, please state in words what each symbol stands for;
  - explain every coefficient;
  - tell me if every coefficient is significantly different from zero (i.e. interpret p-value);
  - how good is your model? (Explain your R-square only. Other regression diagnostics are optional)

2. For count data regression model use as dependent variable `n_stops` (this is the number of stops in a tour). Use as explanatory variables any other variable in the file.
3. For binary regression, create a dummy variable indicating one activity type based on `DissCat`. This is a four category variable for the activity type. The dummy could be one of these activities: Dining, Entertainment, Shopping_major, or Shopping_routine. Use the dummy variable as dependent variable. Use as explanatory variables any other variable in the file.
4. For the multicategory regression model use as dependent variable `DissCat` (this is a four category variable for the activity type). Use as explanatory variables any other variable in the file.

- In the interpretation parts of the above three models (count data regression, binary regression, multicategory regression), you need to:
  - explain every coefficient;

- tell me if every coefficient is significantly different from zero (i.e. interpret t-value and/or p-value);
- run lrtest() to do the likelihood ratio tests and tell me whether your model is better than a null model;
- eliminate insignificant independent variables and build a new trimmed model. Run lrtest() again and tell me whether your new trimmed model is better than the initial model;
- discuss if there are any differences in the coefficients and significance of the common independent variables between the new trimmed model and the initial model.

Load the packages and data

Load the packages

```
# I only load tidyverse here. Load other packages that you need
library(tidyverse)
```

Load the data

In this assignment, you will only use `NWTD_nodupes.rds`. The data and codebook can be found in the `Data` folder on GauchoSpace.

```
NWTD_Data <- read_rds("~/GEOG 111B Data/NWTD_nodupes.rds")  #change the path if necessary
```

Assignment 2 Report (100 points)

Matthew Mangawang

1. Introduction

Requirement: In this section, you should give a brief introduction to the story and description of what comes next. This should be a very short summary describing the main objective (the story) of your assignment.(20 points, word limit: 100-300)

In this assignment, I will be using the NWTD data from the NWTD_nodups.rds file given in class. I will be using 4 regression models (linear, count data, binary, multicategory) to attempt to analyze the reationship between a selected dependent variable and its independent variables. I will use the duration, n_stops, and DissCat as dependent variables, along with start_time, n_vehicles, n_people, n_children, and age as independent variables in the models. I will aim to use the regression data and analysis functions to find out which data is statistically significant and how that could help me understand behavioral travel patterns.

2. Descriptive Statistics

Requirement: Write the code to generate descriptive statistics table(s) including at least **mean, median, min, max, and standard deviation** of some major variables you used in the next regression analysis. You need to write a short analysis of the table. (20 points, word limit: 150-400)

```
# Write the code to do descriptive statistics on some major variables you used
# in the next regression analysis.  You should output the descriptive
# statistics table in the knitted pdf right below the code chunk.  If you want
# to include more than one table, you can create new code chunks in this
# section to split the output tables and your interpretations.

# install.packages('summarytools')
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 4.1.2
```

```
##
## Attaching package: 'summarytools'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

```
NWTD <- readRDS("~/GEOG 111B Data/NWTD_nodupes.rds")
NWTD2 <- subset(NWTD, select = c(duration, start_time, n_vehicles, n_people, n_children,
    age_num))
dfSummary(NWTD2)
```

```
## Data Frame Summary
## NWTD2
## Dimensions: 6756 x 6
## Duplicates: 48
```

```
##
## -------------------------------------------------------------------------------------------------------
## No   Variable     Stats / Values              Freqs (% of Valid)    Graph               Valid      Mi
## ---- ------------ --------------------------- --------------------- ------------------- ---------- --
## 1    duration     Mean (sd) : 64.4 (71.7)     292 distinct values   :                   6756       0
##      [numeric]    min < med < max:                                  :                   (100.0%)   (0
##                   1 < 45 < 850                                      :
##                   IQR (CV) : 51.2 (1.1)                             :
##                                                                     : : .
##
## 2    start_time   Mean (sd) : 856.1 (217.1)   844 distinct values         : :           6756       0
##      [numeric]    min < med < max:                                        : : . :       (100.0%)   (0
##                   180 < 840 < 1596                                        : : : :
##                   IQR (CV) : 343 (0.3)                                  : : : : : .
##                                                                       . : : : : : :
##
## 3    n_vehicles   Mean (sd) : 2 (1)           0 :  302 ( 4.5%)                          6756       0
##      [integer]    min < med < max:            1 : 1769 (26.2%)      IIIII               (100.0%)   (0
##                   0 < 2 < 8                   2 : 3090 (45.7%)      IIIIIIIII
##                   IQR (CV) : 1 (0.5)          3 : 1117 (16.5%)      III
##                                               4 :  376 ( 5.6%)      I
##                                               5 :   58 ( 0.9%)
##                                               6 :   30 ( 0.4%)
##                                               7 :   10 ( 0.1%)
##                                               8 :    4 ( 0.1%)
##
## 4    n_people     Mean (sd) : 2.8 (1.4)       1 : 1116 (16.5%)      III                 6756       0
##      [integer]    min < med < max:            2 : 2301 (34.1%)      IIIIII              (100.0%)   (0
##                   1 < 2 < 8                   3 : 1241 (18.4%)      III
##                   IQR (CV) : 2 (0.5)          4 : 1263 (18.7%)      III
##                                               5 :  549 ( 8.1%)      I
##                                               6 :  192 ( 2.8%)
##                                               7 :   44 ( 0.7%)
##                                               8 :   50 ( 0.7%)
##
## 5    n_children   Mean (sd) : 0.6 (1)         0 : 4631 (68.5%)      IIIIIIIIIIIII       6756       0
##      [integer]    min < med < max:            1 :  874 (12.9%)      II                  (100.0%)   (0
##                   0 < 0 < 6                   2 :  886 (13.1%)      II
##                   IQR (CV) : 1 (1.7)          3 :  262 ( 3.9%)
##                                               4 :   85 ( 1.3%)
##                                               5 :   16 ( 0.2%)
##                                               6 :    2 ( 0.0%)
##
## 6    age_num      Mean (sd) : 51 (16.1)       94 distinct values              . :       6468       28
##      [numeric]    min < med < max:                                          : :       (95.7%)    (4
##                   1 < 53 < 94                                             : : :
##                   IQR (CV) : 20.2 (0.3)                                 : : : : :
##                                                                       . : : : : : : .
## -------------------------------------------------------------------------------------------------------
```

```r
# install.packages('psych')
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
describe(NWTD2)
```

```
##            vars    n   mean     sd median trimmed    mad min  max range  skew
## duration      1 6756  64.41  71.69     45   50.84  37.06   1  850   849  3.55
## start_time    2 6756 856.11 217.11    840  854.41 247.59 180 1596  1416  0.09
## n_vehicles    3 6756   1.98   1.02      2    1.91   1.48   0    8     8  0.84
## n_people      4 6756   2.83   1.42      2    2.70   1.48   1    8     7  0.81
## n_children    5 6756   0.57   0.97      0    0.37   0.00   0    6     6  1.71
## age_num       6 6468  51.01  16.14     53   51.73  14.83   1   94    93 -0.40
##            kurtosis   se
## duration      19.82 0.87
## start_time    -0.59 2.64
## n_vehicles     2.25 0.01
## n_people       0.48 0.02
## n_children     2.44 0.01
## age_num       -0.03 0.20
```

Here we can see the summary table of some of the variables I chose that I will use later in my regression models. What stood out to me from the statistical summary was the low mean number of children (0.57) and the high mean of the age of respondents (51.01). Because of the relatively high age of the respondents, I assume family size would be smaller and that is reflected in both mean number of people and children. Both start time and duration are values I expected from work in the previous assignment.

3. Model estimation

Requirement: You will be creating several model estimation tables and writing one paragraph about each. (40 points, word limit: 300-600)

3.1 Linear regression

```
# Output the table of regression results in the knitted pdf right below the
# code chunk. Same for other code chunks below.

model1 = lm(duration ~ start_time + n_vehicles + n_people + age_num, data = NWTD2)
summary(model1)
```

```
##
## Call:
## lm(formula = duration ~ start_time + n_vehicles + n_people +
##     age_num, data = NWTD2)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -73.40 -41.19 -21.12  13.15 779.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 75.841109   5.960582  12.724  < 2e-16 ***
## start_time  -0.004570   0.004137  -1.105  0.26940
## n_vehicles  -0.777599   0.975834  -0.797  0.42556
## n_people     1.385663   0.750154   1.847  0.06477 .
## age_num     -0.194344   0.060834  -3.195  0.00141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.77 on 6463 degrees of freedom
##   (288 observations deleted due to missingness)
## Multiple R-squared:  0.003419,   Adjusted R-squared:  0.002803
## F-statistic: 5.544 on 4 and 6463 DF,  p-value: 0.0001879
```

The equation of the linear regression model is

$$tripduration(y) = -0.005Var_1 + -0.778Var_2 + 1.386Var_3 + -0.194Var_4 + 75.841$$

where Var_1 denotes start time, Var_2 denotes n_vehicles, Var_3 denotes n_people, Var_4 denotes age_num, and 75.84 is the y-intercept.

Above is the given linear regression model created using the selected variables: trip duration, start_time, n_vehicles, n_people, and age_num. The predicted value and y-intercept of the dependent variable (trip duration) is 75.841. The values in front of variables 1 - 4 are the regression coefficients of the model that respond to each of the 4 selected independent variables. The variables with the lowest p-values are the intercept, and age, which are both well below 0.05, probably implying that we can reject the null hypothesis of these variables and are likely to be a meaningful addition to my model. However, n_people, start_time, and n_vehicle variables all have p-values greater than 0.05, probably meaning that these values are less statistically meaningful to my model. My model has an adjusted R-squared value of 0.002, which is very low, but this makes sense given that 2 of my independent variables had such high P-values. So, based on my chosen variables, my model is not very good.

3.2 Count data regression model

```
# Produce a count data regression model and run lrtest()

# install.packages('MASS') install.packages('lmtest')
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
negbinmodel <- glm.nb(n_stops ~ start_time + n_vehicles + n_people + n_children +
    age_num, data = NWTD)
summary(negbinmodel)
```

```
##
## Call:
## glm.nb(formula = n_stops ~ start_time + n_vehicles + n_people +
##     n_children + age_num, data = NWTD, init.theta = 8.265730977,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1498  -0.8776  -0.2705   0.3554   7.6222
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.285e+00  6.086e-02  21.114  < 2e-16 ***
## start_time  -2.413e-04  4.261e-05  -5.663 1.49e-08 ***
## n_vehicles  -6.694e-02  1.109e-02  -6.034 1.60e-09 ***
## n_people    -1.254e-02  1.160e-02  -1.081   0.2798
## n_children   6.757e-03  1.538e-02   0.439   0.6605
## age_num     -1.136e-03  6.332e-04  -1.793   0.0729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(8.2657) family taken to be 1)
##
##     Null deviance: 5554.0  on 6467  degrees of freedom
## Residual deviance: 5454.7  on 6462  degrees of freedom
##   (288 observations deleted due to missingness)
## AIC: 23677
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  8.266
```

```
##           Std. Err.:  0.554
##
##  2 x log-likelihood:  -23663.357
```

```r
lrtest(negbinmodel)
```

```
## Likelihood ratio test
##
## Model 1: n_stops ~ start_time + n_vehicles + n_people + n_children + age_num
## Model 2: n_stops ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   7 -11832
## 2   2 -11881 -5 98.729  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here I used a count data regression model using a negative binomial regression model. The independent variables I chose are the same ones from the linear regression model: start_time, n_vehicles, n_people, n_children, and age_num. The dependent variable for this model though is n_stops (the number of stops). Based on this model, we can see that the p-values are particularly low (<0.001) for the intercept, start_time, and n_vehicles coefficients, while the n_people and n_children values are relatively large. Those variables are ones that I intend on trimming in order to create a better model. Finally, we can see that the p-value when compared to a null model is very low (near 0), so we would reject the null hypothesis, and this model offers a significant improvement over a null model.

```r
# Produce a trimmed model and run lrtest()

negbintrim <- glm.nb(n_stops ~ start_time + n_vehicles + age_num, data = NWTD)
summary(negbintrim)
```

```
##
## Call:
## glm.nb(formula = n_stops ~ start_time + n_vehicles + age_num,
##     data = NWTD, init.theta = 8.262135822, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1654  -0.8778  -0.2709   0.3541   7.6596
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.252e+00  5.373e-02  23.302  < 2e-16 ***
## start_time  -2.382e-04  4.249e-05  -5.607 2.06e-08 ***
## n_vehicles  -7.352e-02  9.322e-03  -7.886 3.11e-15 ***
## age_num     -9.032e-04  5.747e-04  -1.572    0.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(8.2621) family taken to be 1)
##
##     Null deviance: 5553.4  on 6467  degrees of freedom
```

```
## Residual deviance: 5455.6  on 6464  degrees of freedom
##   (288 observations deleted due to missingness)
## AIC: 23675
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  8.262
##           Std. Err.:  0.553
##
##  2 x log-likelihood:  -23664.828
```

```
lrtest(negbinmodel, negbintrim)
```

```
## Likelihood ratio test
##
## Model 1: n_stops ~ start_time + n_vehicles + n_people + n_children + age_num
## Model 2: n_stops ~ start_time + n_vehicles + age_num
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   7 -11832
## 2   5 -11832 -2 1.4714     0.4792
```

Here we can see the results of the trimmed count data regression model that omits n__people and n__children variables. We can see htat the intercept, start__time, and n__vehicle coefficients remain fairly unchanged, while the age__num coefficient got smaller. The intercept, start__time, and age__num variables were relatively unchanged as well, while the n__vehicles standard error decreased. We can also see that the p-value decreased slightly for start__time, and decreased greatly for n__vehicles, while the p-value of age__num increased above 0.1. When performing the lrtest between the 2 models, we can see that the p-value given is 0.479, well above the 0.05 threshold, meaning that we fail to reject the null hypothesis. Both models fit the data, but we should use the trimmed one because the additional variables do not offer a significant improvement in fit.

3.3 Binary regression

```
# Produce a binary regression model and run lrtest()
NWTD$Diss_Cat <- ifelse(NWTD$DissCat == "Dining", 1, 0)
summary(as.factor(NWTD$Diss_Cat))
```

```
##    0    1
## 4520 2236
```

```
disscatlogit <- glm(Diss_Cat ~ start_time + n_vehicles + n_people + n_children +
    age_num, family = binomial(link = "logit"), data = NWTD)
summary(disscatlogit)
```

```
##
## Call:
## glm(formula = Diss_Cat ~ start_time + n_vehicles + n_people +
##     n_children + age_num, family = binomial(link = "logit"),
```

```
##     data = NWTD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5835  -0.9156  -0.8322   1.4048   1.9443
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5555410  0.1777207  -3.126  0.00177 **
## start_time   0.0003313  0.0001232   2.689  0.00717 **
## n_vehicles   0.2498720  0.0324814   7.693 1.44e-14 ***
## n_people    -0.1970841  0.0350253  -5.627 1.83e-08 ***
## n_children   0.0579429  0.0455761   1.271  0.20361
## age_num     -0.0079593  0.0018374  -4.332 1.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8225.6  on 6467  degrees of freedom
## Residual deviance: 8124.1  on 6462  degrees of freedom
##   (288 observations deleted due to missingness)
## AIC: 8136.1
##
## Number of Fisher Scoring iterations: 4
```

```
lrtest(disscatlogit)
```

```
## Likelihood ratio test
##
## Model 1: Diss_Cat ~ start_time + n_vehicles + n_people + n_children +
##     age_num
## Model 2: Diss_Cat ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -4062.1
## 2   1 -4112.8 -5 101.49  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, I created a binary regression model using logit, as discussed in lab. I used the same independent variables I used in the previous models, however, my dependent variable is from the DissCat column (a 4 category variable for activity type). The DissCat variable I chose was "Dining" and I created a dummy variable using this selection to use as the dependent variable in the regression model. Here, we can see that the n_vehicles, n_people, and age_num are the most statistically significant variables based on p-value, while intercept and start_time have values below 0.01. The n-children variable has the highest p-value and well above the 0.05 threshold. Finally, we can see that the p-value when compared to a null model is very low (near 0), so we would reject the null hypothesis, and this model offers a significant improvement over a null model.

```
# Produce a trimmed model and run lrtest

disscatlogittrim <- glm(Diss_Cat ~ start_time + n_vehicles + n_people + age_num,
```

```
    family = binomial(link = "logit"), data = NWTD)
summary(disscatlogittrim)
```

```
##
## Call:
## glm(formula = Diss_Cat ~ start_time + n_vehicles + n_people +
##     age_num, family = binomial(link = "logit"), data = NWTD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5381  -0.9148  -0.8303   1.4033   1.9556
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5603322  0.1775451  -3.156  0.00160 **
## start_time   0.0003276  0.0001232   2.660  0.00782 **
## n_vehicles   0.2326572  0.0293877   7.917 2.44e-15 ***
## n_people    -0.1642158  0.0234075  -7.016 2.29e-12 ***
## age_num     -0.0082976  0.0018157  -4.570 4.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8225.6  on 6467  degrees of freedom
## Residual deviance: 8125.7  on 6463  degrees of freedom
##   (288 observations deleted due to missingness)
## AIC: 8135.7
##
## Number of Fisher Scoring iterations: 4
```

```
lrtest(disscatlogit, disscatlogittrim)
```

```
## Likelihood ratio test
##
## Model 1: Diss_Cat ~ start_time + n_vehicles + n_people + n_children +
##     age_num
## Model 2: Diss_Cat ~ start_time + n_vehicles + n_people + age_num
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -4062.1
## 2   5 -4062.9 -1 1.6193     0.2032
```

Here, I trimmed the earlier binary regression model by removing n_children, which I deemed to be statistically irrelevant. All the coefficient values and standard errors remained fairly similar, while all the p-values decreased. We can also see in the lrtest that the p-value was 0.2032, which is means to fail to reject the null hypothesis. Both models should fit the data, but we should use the trimmed model because the n_children variable does not offer a significant improvement in fit.

3.4 Multicategory regression model

```r
# Produce a Multicategory regression model and run lrtest()

# install.packages('stargazer') install.packages(nnet)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.1.2
```

```r
multicat <- multinom(DissCat ~ start_time + n_vehicles + n_people + n_children +
    age_num, data = NWTD, hessian = TRUE)
```

```
## # weights:  28 (18 variable)
## initial  value 8966.551928
## iter  10 value 6669.623202
## iter  20 value 6560.372540
## final  value 6554.721545
## converged
```

```r
stargazer(multicat, type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     -------------------------------------------
##                     Entertainment Shopping_major Shopping_routine
##                          (1)           (2)             (3)
## ---------------------------------------------------------------
## start_time            0.001***      -0.001***       -0.001***
##                       (0.0002)       (0.0002)        (0.0001)
##
## n_vehicles             -0.093         -0.094         -0.286***
##                        (0.059)       (0.079)         (0.034)
##
## n_people               -0.079         0.133*          0.242***
##                        (0.063)       (0.080)         (0.036)
##
## n_children            0.260***        -0.093         -0.104**
##                        (0.082)       (0.113)         (0.047)
##
```

```
## age_num               -0.005*        0.003          0.011***
##                       (0.002)        (0.003)        (0.002)
##
## Constant             -2.027***      -1.699***       0.381**
##                       (0.038)        (0.016)        (0.160)
##
## ----------------------------------------------------------------
## Akaike Inf. Crit.  13,145.440     13,145.440      13,145.440
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

```r
lrtest(multicat)
```

```
## # 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 ~ start_time + n_vehicles + n_people + n_children + age_num
## Model 2: DissCat ~ 1
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  18 -6554.7
## 2   3 -6680.6 -15 251.68  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here is a multicategory regression model using multinominal logit. I used the DissCat categorical variable again as my dependent variable with the default reference category set as "Dining." The independent variables are the same as the other models. Now, we get coefficient values for every independent variable in relationship to each categorical variable. We can see that the p-values are all very low for the shopping_routine variable in relation to the independent variables, while the shopping_major and entertainment variables have 3 and 2 variables, respectively, with high p-values. Finally, we can see that the p-value when compared to a null model is very low (near 0), so we would reject the null hypothesis, and this model offers a significant improvement over a null model.

```r
# Produce a trimmed model and run lrtest()

multicattrim <- multinom(DissCat ~ start_time + n_people + n_children + age_num,
    data = NWTD, hessian = TRUE)
```

```
## # weights:  24 (15 variable)
## initial  value 8966.551928
## iter  10 value 6681.242235
```

```
## iter  20 value 6592.904461
## final  value 6592.903348
## converged
```

```r
stargazer(multicattrim, type = "text")
```

```
##
## ===============================================================
##                              Dependent variable:
##                   ---------------------------------------------
##                   Entertainment Shopping_major Shopping_routine
##                        (1)           (2)              (3)
## ---------------------------------------------------------------
## start_time          0.001***      -0.001***        -0.001***
##                     (0.0002)      (0.0002)         (0.0001)
##
## n_people           -0.138***       0.070           0.069**
##                     (0.048)       (0.061)          (0.028)
##
## n_children          0.316***       -0.034           0.058
##                     (0.073)       (0.103)          (0.043)
##
## age_num             -0.005*        0.002           0.010***
##                     (0.002)       (0.003)          (0.002)
##
## Constant           -2.075***      -1.737***         0.242
##                     (0.037)       (0.016)          (0.158)
##
## ---------------------------------------------------------------
## Akaike Inf. Crit.  13,215.810     13,215.810      13,215.810
## ===============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

```r
lrtest(multicat, multicattrim)
```

```
## Likelihood ratio test
##
## Model 1: DissCat ~ start_time + n_vehicles + n_people + n_children + age_num
## Model 2: DissCat ~ start_time + n_people + n_children + age_num
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  18 -6554.7
## 2  15 -6592.9 -3 76.364  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, I trimmed the previous multicategory regression model by removing the n_vehicles variable from the model, which had the highest p-values across the deendent variables. Most of the coefficients remained similar, but some of the shopping_routine coefficients became positive. The entertainment p-values decreased while the shopping_routine p-values stayed relatively unchanged. One of the shopping_major p-values moved above 0.1. Finally, when performing the lrtest between the 2 models, we can see that the p-value is basically 0, meaning that we would reject the null hypthesis. We could conclude that the model before trimming offers an improvement over the trimmed model.

4. Summary

Requirement: In this section, you should give a brief conclusion pointing out the most important findings. (20 points, word limit: 100-300)

Using the first 3 methods of regression modelling (linear, count data, binary), we can see that there was usually at least 1 variable that we could omit as statistically insignificant, due to high p-values. However, in the final multicategory regression model, we can see that there was less variation between the trimmed and untrimmed models and it was actually better to leave in all the variables I chose. Using all these models, we can start to analyze the relationships between dependent and independent variables. If analyzed even further, we could compare whatever variable chosen and develop a basis for prediction and effects on target variables.