# Modeling Number of Cars with a Cumulative Logit Model

Categorical Analysis

Jennifer Eberling

## Introduction

I found my data set on Kaggle from an unknown source. The description states that it came from bike stores across the world, and it has 13 variables. The variables contain demographic information on customers that have visited the stores.

The variables, with base level listed first in the parentheses, are

- ID
- Cars - number of Cars (0,1,2,3,4)
- Commute.Distance – commute distance to workplace (0-1,1-2,2-5,5-10,10+ miles)
- Education (Partial High School, High School, Partial College, Bachelors, Graduate Degree)
- Occupation (Clerical, Management, Manual, Professional, Skilled Manual)
- Children (0,1,2,3,4,5)
- Region (Europe, North America, Pacific)
- Purchased.Bike – whether the customer purchased a bike (No, Yes)
- Home.Owner – whether the customer owns a home (No, Yes)
- Marital.Status (Married, Single)
- Gender (Female, Male)
- Income – annual income (treated numerically)
- Age – in years (treated numerically)

Id was dropped, and all categorical variables were initially treated as nominal variables. There are 1000 complete observation in this cleaned dataset. This paper will detail the process of creating a cumulative logit model with Cars as the response variable.

## Data Exploration

The lowest counts for the number of cars variable are in the three and four car categories (Table 1). In order to make this more evenly distributed, I combined three and four cars so that the smallest cell count is 145 instead of 60 (Table 2). Moving forward, this category will be labeled as '3' and not '3+.'

| | Number of Cars | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Counts | 243 | 267 | 345 | 85 | 60 |

Table 1

| Counts | Number of Cars | | | |
|---|---|---|---|---|
| | 0 | 1 | 2 | 3+ |
| | 243 | 267 | 345 | 145 |

Table 2

Looking at cars versus commute distance, we see the highest count is people who live 0-1 miles from their work and have 0 cars (Table 3). Moving across the ‘0’ row results in a decrease in counts, meaning that people living further from their work are more likely to have at least one car. People with ‘2’ or ‘3’ cars have the highest counts in 5-10 miles and 10+ miles respectively.

| | | Commute Distance | | | | |
|---|---|---|---|---|---|---|
| | | 0-1 Miles | 1-2 Miles | 2-5 Miles | 5-10 Miles | 10+ Miles |
| Number of Cars | 0 | 166 | 31 | 36 | 6 | 4 |
| | 1 | 85 | 48 | 72 | 49 | 13 |
| | 2 | 80 | 75 | 38 | 109 | 43 |
| | 3 | 35 | 15 | 16 | 28 | 51 |

Table 3

To better understand the 0-1 Miles count, it is important to look at a table of Region versus commute distance (Table 4). In Europe, about 63% (188/300) of customers have a commute distance of 0-1 Miles. This is in stark contrast to North America’s 25% (126/508) and the Pacific’s 27% (52/192). Of the three regions, Europe has the lowest average income at $40,266.67 compared to North America’s $62,716.54 and the Pacific’s $63,541.67.

| | | Number of Cars | | | | |
|---|---|---|---|---|---|---|
| | | 0-1 Miles | 1-2 Miles | 2-5 Miles | 5-10 Miles | 10+ Miles |
| Region | Europe | 188 | 38 | 40 | 16 | 18 |
| | North America | 126 | 108 | 103 | 109 | 62 |
| | Pacific | 52 | 23 | 19 | 67 | 31 |

Table 4

When it comes to the five categories of occupation, clerical and manual workers are the least likely to have three or more cars (Table 5). They also happen to be the categories with the lowest mean incomes.

| | | Occupation | | | | |
|---|---|---|---|---|---|---|
| | | Clerical | Management | Manual | Professional | Skilled Manual |
| Number of Cars | 0 | 79 | 14 | 34 | 54 | 62 |
| | 1 | 50 | 31 | 47 | 73 | 66 |
| | 2 | 47 | 66 | 38 | 79 | 115 |
| | 3 | 1 | 62 | 0 | 70 | 12 |

Table 5

The table of cars versus occupation tells us that customers with ‘Partial High School’ as their highest education level are most likely to have two cars while those with ‘Graduate Degree’ are most likely to have zero cars (Table 6).

| | Education | | | | | |
|---|---|---|---|---|---|---|
| | | Partial High School | High School | Partial College | Bachelor’s Degree | Graduate Degree |
| Number of Cars | 0 | 1 | 13 | 30 | 84 | 115 |
| | 1 | 1 | 37 | 121 | 96 | 12 |
| | 2 | 68 | 104 | 78 | 65 | 30 |
| | 3 | 6 | 25 | 36 | 61 | 17 |

Table 6

**Cumulative Logit Model**

To create my cumulative logit model, I performed a forward selection based on the AIC with a proportional odds assumption. The first variable added to the model was commute distance with an AIC of 2480.624. The ‘Variables Added’ column of Table 7 shows the variables in the order they were added to the model. The ‘Model’s AIC’ column reports the AIC of the model at that point. For example, Occupation was the third variable added. The model with Commute.Distance, Education, and Occupation in it has an AIC of 2036.45. Age, Gender, and Marital.Status were the only variables that were not added to the model with this selection method.

| **Variable Added** | **Model’s AIC** |
|---|---|
| Commute.Distance | 2480.624 |
| Education | 2348.766 |
| Occupation | 2036.45 |
| Income | 1960.879 |
| Children | 1901.528 |
| Region | 1886.925 |
| Purchased.Bike | 1875.33 |
| Home.Owner | 1870.065 |

Table 7

I then performed stepwise selection based on the AIC to see if removing any one variable after adding one would increase the AIC. No removal was necessary though, so the resulting model was the same.

I tested models that added a range of interaction terms including interactions between Commute Distance and Region, Income and Homeownership, and Education and Income. I found that the models with interaction terms for Occupation and Education, Commute Distance and Children, and Commute Distance and Occupation result in error, likely because these are the variables with the most categories, and the models are too complex for the limited data. Because of this restriction and the complex interpretations of interaction terms, I decided not to include any interaction terms in my final model.

Next, for each of my nominal variables for which it made sense, I tested whether they should be treated as ordinal or not. I tested Commute Distance with values of (0,1,2,5,10) and (0,1,2,3,4), Education with values of (0,1,2,3,4), and Children with values of (0,1,2,3,4,5). I did not test Occupation, because it was not reasonable to assign ordered values to a job title. For each of the selected variables, I created a model treating the variable as ordinal and as nominal, then performed a likelihood ratio test. Using a p-value of .05, only Commute Distance (p = 0.5993) had a p-value > .05 and should be treated ordinally using values of (0,1,2,5,10), so that is what I continued with in my full model:

<u>When i=0 (Cars = 0)</u>

$$\text{Logit}[\hat{P}(\text{Y}\leq\text{i})] = -2.337 - .1294(\text{X}_1) + 1.368(\text{X}_2) + 2.432(\text{X}_3) + 3.75(\text{X}_4) + 5.967(\text{X}_5) - 2.33(\text{X}_6) - .06829(\text{X}_7) - .254(\text{X}_8) - .2353(\text{X}_9) - .00003471(\text{X}_{10}) + .3395(\text{X}_{11}) + .1541(\text{X}_{12}) - .7278(\text{X}_{13}) - .275(\text{X}_{14}) - 1.796(\text{X}_{15}) - .3527(\text{X}_{16}) - 1.004(\text{X}_{17}) + .5133(\text{X}_{18}) + .4261(\text{X}_{19})$$

<u>When i=1 (Cars = 1)</u>

$$\text{Logit}[\hat{P}(\text{Y}\leq\text{i})] = -.1016 - .1294(\text{X}_1) + 1.368(\text{X}_2) + 2.432(\text{X}_3) + 3.75(\text{X}_4) + 5.967(\text{X}_5) - 2.33(\text{X}_6) - .06829(\text{X}_7) - .254(\text{X}_8) - .2353(\text{X}_9) - .00003471(\text{X}_{10}) + .3395(\text{X}_{11}) + .1541(\text{X}_{12}) - .7278(\text{X}_{13}) - .275(\text{X}_{14}) - 1.796(\text{X}_{15}) - .3527(\text{X}_{16}) - 1.004(\text{X}_{17}) + .5133(\text{X}_{18}) + .4261(\text{X}_{19})$$

<u>When i=2 (Cars = 2)</u>

$$\text{Logit}[\hat{P}(\text{Y}\leq\text{i})] = 2.865 - .1294(\text{X}_1) + 1.368(\text{X}_2) + 2.432(\text{X}_3) + 3.75(\text{X}_4) + 5.967(\text{X}_5) - 2.33(\text{X}_6) - .06829(\text{X}_7) - .254(\text{X}_8) - .2353(\text{X}_9) - .00003471(\text{X}_{10}) + .3395(\text{X}_{11}) + .1541(\text{X}_{12}) - .7278(\text{X}_{13}) - .275(\text{X}_{14}) - 1.796(\text{X}_{15}) - .3527(\text{X}_{16}) - 1.004(\text{X}_{17}) + .5133(\text{X}_{18}) + .4261(\text{X}_{19})$$

$\hat{P}(\text{Y}\leq\text{i})$ = Estimated probability of being in a category equal or lesser than a given category I<br>
$\text{X}_1$ = Distance.Ordinal<br>
$\text{X}_2$= Indicator for High School<br>
$\text{X}_3$= Indicator for Partial College<br>
$\text{X}_4$= Indicator for Bachelors

$X_5$= Indicator for Graduate Degree
$X_6$= Indicator for Management
$X_7$= Indicator for Manual
$X_8$= Indicator for Professional
$X_9$= Indicator for Skilled Manual
$X_{10}$= Income
$X_{11}$= Indicator for 1 child
$X_{12}$= Indicator for 2 children
$X_{13}$= Indicator for 3 children
$X_{14}$= Indicator for 4 children
$X_{15}$= Indicator for 5 children
$X_{16}$= Indicator for North America
$X_{17}$= Indicator for Pacific
$X_{18}$= Indicator for Purchasing a bike
$X_{19}$= Indicator for Homeowner

The model has an AIC of 1865.938.

## Discussion

There seems to be a contradiction in the model. Controlling for all else, we estimate the odds of having at least x cars to increase multiplicatively by exp(0.03471) = 1.035319 for every $1000 increase in income and we also see that a graduate degree is associated with decreased odds of having at least x cars by exp(2.217) = 9.18 times what having an undergraduate degree does. Essentially, the odds of having more cars decreases as education level increases but it increases as income increases. This is an unexpected relationship, because there is a general positive trend between increased education level and Income (Figure 1).

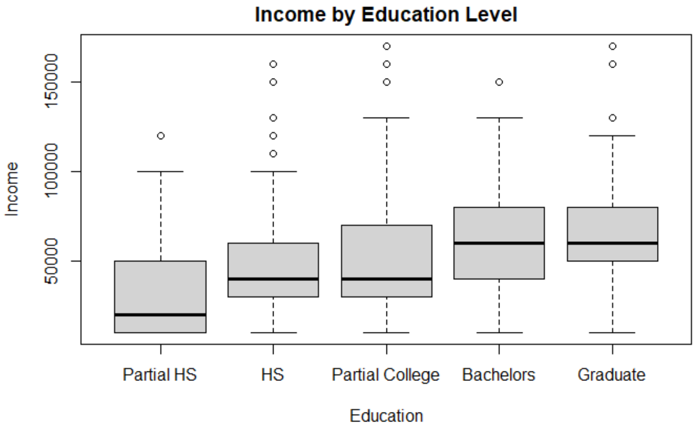

Figure 1

Another unexpected relationship is the one between the estimated odds of having at least x cars and Children. The indicators for having 1 and 2 children are positive, meaning that having 1 or 2 children results in a decreased odds of having at least x cars compared to those who have 0 children. This relationship is flipped for those who have 3,4, or 5 children. Those coefficients are negative, so having 3,4, or 5 children results in an increased odds of having at least x cars. It makes sense that larger families would need more cars. It is less clear why having 1 or 2 children would decrease the odds of having more cars when compared to someone with 0 children. It could be a result from potential lack of expendable income, but more research is needed.

**Analysis**

To check for potential overfitting, I attempted a test the proportional odds assumption. However, the model does not run when "parallel = F," suggesting the model may be too complex for the thousand observations without the proportional odds assumption.

I also checked the standardized residuals. Using a cutoff of ±3, there were 69 outliers. The largest was 1268.557262 and belonged to a man who commuted a long distance and has 3 children but 0 cars. All 23 of the residuals outside of ±8 were for people with 0 Cars. Of the initial 69 outliers, 48 had 0 Cars, 17 had 1 Car, and 4 had 2 Cars. This means that the model is not performing as well for people with 0 Cars as it does for those with more Cars. There is some piece of information that it is missing.

**Predictions**

To test my model with real life scenarios, I used my own information to find the estimated probability of me owning 0,1,2, and 3+ Cars. For someone living in North America a ½ mile from their work with an undergrad degree, a professional job, $20,000 salary, and 0 children, the probabilities are:

- $\hat{P}(\mathrm{Y} = 0) = 0.5279957$
- $\hat{P}(\mathrm{Y} = 1) = 0.3847396$
- $\hat{P}(\mathrm{Y} = 2) = 0.0823671$
- $\hat{P}(\mathrm{Y} = 3) = 0.0048976$

The situation with the highest probability is that I have 0 cars, which is the case. The second highest is that I have 1 car, which is the situation that a lot of my coworkers are in. It is unlikely that someone in my position would have need for or be able to afford 2 or more cars.

Next, I ran my model with a set of inputs I could see for myself in the future to compare those probabilities to my current ones. In this version of the future, I have a 2-5 mile commute to my management job that makes a $120,000 salary. I have moved to New Zealand after graduating from Villanova University with my graduate degree, and I am now a proud homeowner with 0 children. My estimated probabilities are now

- $\hat{P}(\mathrm{Y} = 0) = 0.02408233$
- $\hat{P}(\mathrm{Y} = 1) = 0.1633929$
- $\hat{P}(\mathrm{Y} = 2) = 0.6301142$
- $\hat{P}(\mathrm{Y} = 3) = 0.1824106$

Now what used to be my highest probability, having 0 cars, is now by far my lowest at 2.4%. The highest probability is having 2 cars (63.01%), and 1 and 3 have similar probabilities (16.34% and 18.24% respectively) despite being rather different lifestyle choices.

**Limitations**

In the data exploration stage of this project, I noticed that the data looked extremely clean – even for a “cleaned” dataset. For example, all income terms are rounded to the nearest thousand and are all in the range of $10,000 to $170,000. Another example is that income is identically distributed between the genders (Figure 2). I was also unable to identify how the variable Cars was defined. Cars could be per family or per person, and that is an important distinction for this analysis.

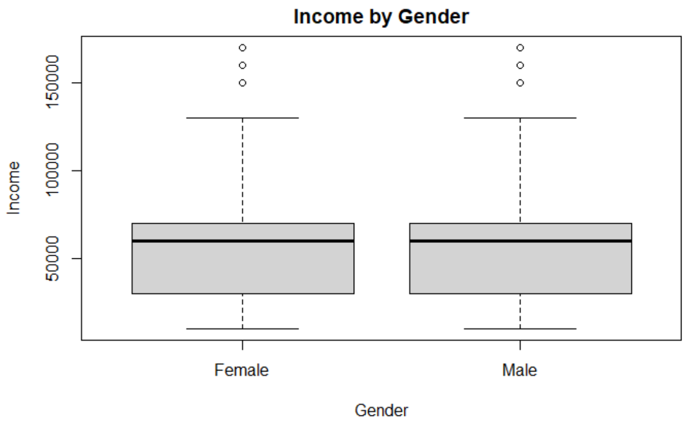

Figure 2

It is also unclear if the data comes from one store chain or multiple. This, as well as the type of bike being purchased, could change our model. For instance, buying a mountain bike may imply the necessity of an extra car to transport the bike, while the purchase of an electric bike may suggest a desire to move away from car ownership/use.

Another variable that could provide insight is whether the customer lives in an urban or rural environment. We observed in Table 6 that people with a partial high school education are most likely to have 2 cars whereas people with a graduate degree are most likely to have 0. This could be due to the demands of their environment rather than a personal choice. It is also important to consider how bike store customers differ from the general public since they are our sampling population. A final limitation to note is that this data excludes other vehicles that could be substituted for a car, such as a motorcycle.

### Extraneous Logit Models

After creating a cumulative logit model, I created a logit model where 0 = has no car and 1 = has any number of cars to compare which variables were added. Forward selection with Commute Distance as nominal added all variables to the model, but Commute Distance was added 8th instead of 1st. The AIC was 557.95. Using backwards selection, Region, Age, Commute Distance, Children, Homeowner, Occupation, and Education were added and it had an AIC of 550.5.

Treating Commute Distance as ordinal (0,1,2,5,10), forward selection still added all variables to the model with Commute Distance added 8th. The AIC was 559.1. Using backwards selection, Age, Region, Commute Distance (Ordinal), Children, Homeowner, Occupation, and Education were added, and it had an AIC of 551.8. Treating Commute Distance results in higher AIC models, and Age is added before Region.

Thus, Commute Distance is more helpful in predicting the number of cars someone has rather than if they have one at all. For that, Age and Region give more information.

### Conclusion

Understanding how and why people move is important to understand before making changes to or investing in transportation policies. This analysis of a cumulative logit model has highlighted the role of a person’s commute distance in how many cars they have, although it is not as important as age or region when predicting if someone has at least one car or not. There are conflicting trends within the model when it comes to income and education, with a higher income associated with more cars, a higher education associated with less, but a higher income still associated with higher education. Despite this, it performed reasonably when applied to a real-life subject.

The data is limited by its sampling population – customers of bike stores as opposed to the general public – and there are a number of variables that might help the cumulative logit model fit better, such as a variable to reflect population density and ownership of other transportation methods.

The model was already too complex for several interaction terms, so a larger dataset could allow me to get a more accurate model by adding more terms. This would be especially helpful for modeling situations where people have 0 cars, the observations that this final model most often failed with.

Ultimately, the model does tell us a lot about what factors decide how many cars people have and the amount of information that we are still missing when it comes to an individual's situation and transportation decisions.