Homework Assignment No. 1

Binary Choice Models - Application to Electric Vehicle Adoption

Sebastián Cancino Espinoza

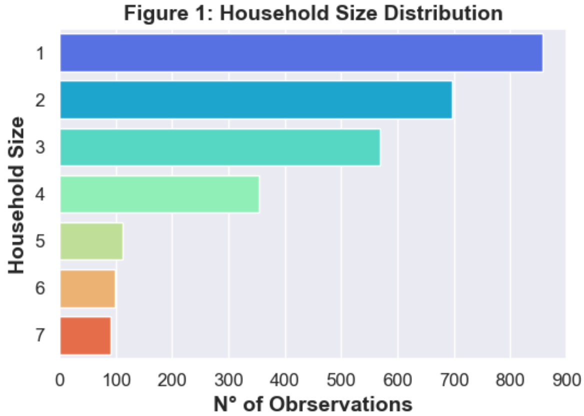
1. Examine Characteristics of the Data Set

Each observations of the Data Set consists on characteristics of a household and the yearly expected cost of get the electric vehicle. The goal is to understand which attributes are important in the decision of adopt a new electric vehicle or not. We will start exploring the dataset to understand the characteristics of the households in the sample.

Demographic Characteristics of the Data Set

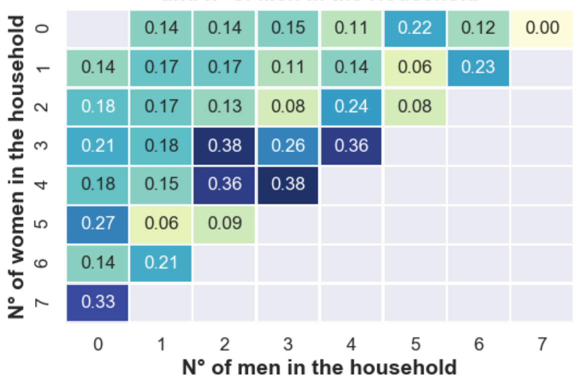
The number of people that lives in each household varies from 1 to 7 with an average of ~2.58 people per household. The vast majority of the observations corresponds to households composed of 1 - 4 people, as can be seen in the Figure 1. There is a great amount of people that lives alone.

Figure 1: Household Size Distribution



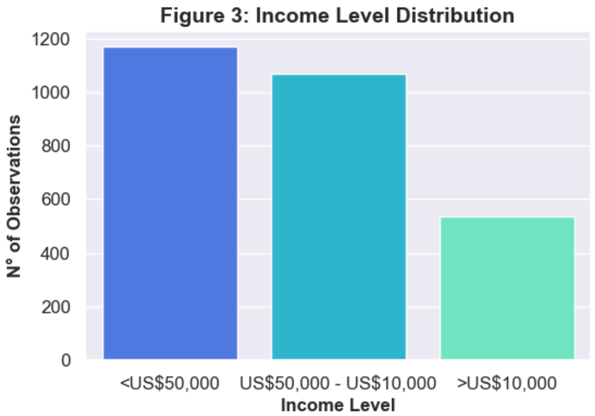
Naturally, the WOMEN and MEN variables distribution is pretty similar across the Data Set and the adoption percentage seems to increase as both increases (particularly the WOMEN variable). Nevertheless, this increase may be due to the fact that there are less households that has a higher number of people, what affects the confidence of the estimation.

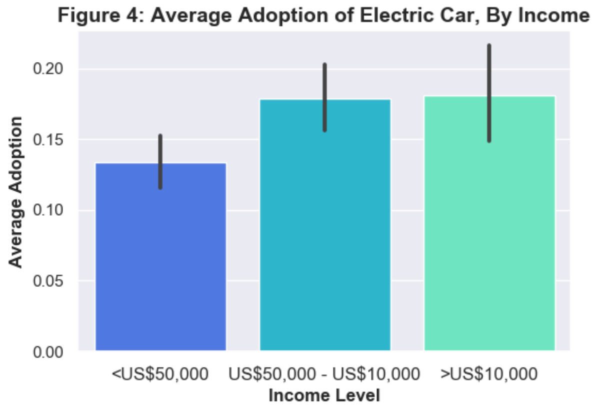
Figure 2: Adoption Perc. by n° of Women and n° of Men in the Household



All variables that refers to number of people (e.g. WOMEN, MEN, NWORK) seems to be relatively high correlated with HHSIZE. In fact, WOMEN + MEN = HHSIZE. Hence, we expect that only a few of them will appear in the final model.

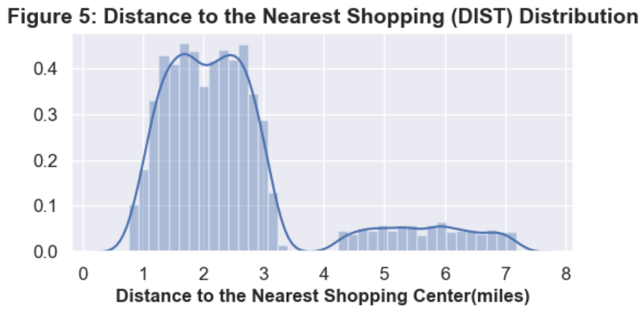
The INCOME distribution shows that there are more households on the lower and mid income level. We could naturally ask if households with higher income levels tend to adopt easily this technology. It appears that there is a ~5% increment in the adoption percentage from low to mid income level, and has a slightly increase from mid to high income. Again, the fact that there are less observations on the high-income levels increases the confidence intervals.





Characteristics of the Data Set Related to Transportation Method

The DIST variable distribution shows clearly two groups that could be converted to a categorical variable representing short and long distances (Figure 5).



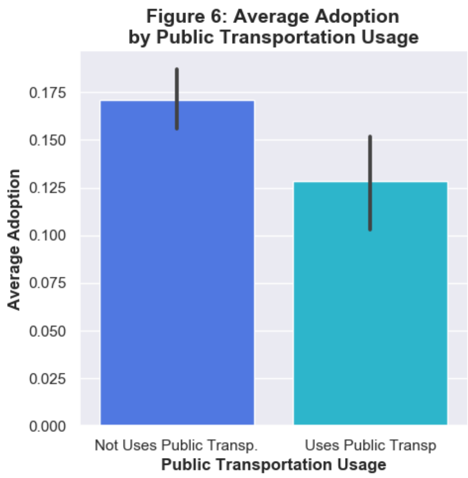


Table 1: Average Public Transportation Usage Per Income Level

Income Level (US$/Year)	Average Usage of Public Transportation
<US$50,000	26,03%
US$50,000 – US$100,000	25,42%
>US$100,000	26,25%

Households in which at least one person uses public transport has least probability of adopt the electric vehicle (Figure 6). We could explain this by saying that people that uses public transportation has a lower income level and therefore, they can't afford an electric vehicle, Nevertheless, if we take a look to the average usage of public transportation across the income levels

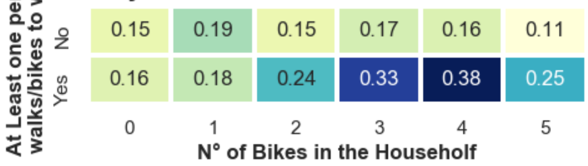


Figure 7: Average Adoption by n° of Bikes and walks/bikes to work

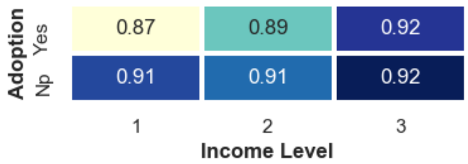


Figure 8: Average Yearly Cost by Income Level and Adoption

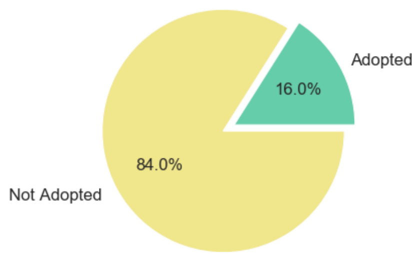


Figure 9: Percentage of Adoption of the Electric Car

on Table 1, we see that it is nearly the same for all of them, and in fact a little greater for the high income level.

Households with people who walks or bikes to work (Figure 7) has higher chance to adopt the vehicle than those who don't, and this chance increases with the number of bikes in the household. It's important to note that the majority of households in the dataset doesn't has any bike, and this can be affecting the estimation.

The yearly mean cost of adopting the electric vehicle is US$9,095. We see (Figure 8) that households that actually adopted the electric car, were offered to lower yearly costs, especially on low- and mid-income levels. Also, we see that higher income levels tend to accept higher yearly costs, as expected.

The Data Set is highly biased towards households that not adopted the electric vehicle. In fact, only 16.0 % of the households adopted the technology (Figure 9).

Additional Notes:

- Strangely, households that has solar panels has less probability of adopting the electric car.
- People seems to be influenced by their neighbors. Households that has neighbors that have electric cars has a ~5% more chance to adopt the technology according to the dataset.
- Households that has at least one senior member seems to adopt the technology easily.

Conclusion

Possible explanatory variables to the problem could be INCOME, MEN, WOMEN, PTRANSP, WALKBIKE, NBIKES, NEIGHB. Due to relatively high correlation, one of WOMEN MEN and HHSIZE may not be in the final model.

Since the dataset has a high percentage of households that didn't adopted the technology, it's difficult to find a variable (or variables) that has a high impact in the adoption percentage. For this same reason, we expect the 'only constant model' to return a lower chance, and thus a lower utility value, for the option of adopt the technology.

2. Estimate Base Market Share Model

The base market share model, or only constants model, consists of the utility functions:

$$V_{YES} = ASC_{YES},$$
$$V_{NO} = ASC_{No},$$

Where V_{YES} refers to the utility value of adopt the electric car, and V_{NO} refers to the utility value of not adopt the electric car. Since we cannot estimate both constats, we set the ASC_{No}constant equal to 0. Thus, the model that we estimate on Biogeme is:

$$V_{YES} = ASC_{YES},$$
$$V_{NO} = 0.$$

Table 2: Estimated Parameters for the Only Constants Model

Parameter	Value	t-Test
$\boldsymbol{ASC_{YES}}$	-1.658205	-32.023794

Biogeme outputs a negative value (Table 2) for the ASC_{YES} constant, with a t-Test value of ~-32, greater than 1, and, therefore significant. The negative sign of the constant reflects the fact that there are more people that did not adopted de electric car in the dataset. Thus, under the only constant model, choosing to adopt the electric vehicle gives a lower utility value when comparing to not adopting the alternative. If we compute the probability of an individual choosing to adopt the vehicle, which in this case is the same for all individuals in the sample, we obtain the same value that we obtained before when computing the market shares.

$$P(YES) = \frac{\exp(V_{YES})}{\exp(V_{YES}) + \exp(V_{NO})} = \frac{\exp(-1.658205)}{\exp(-1.658202) + \exp(0)} = \sim 0.16$$

This model has a $\bar{\rho}^2$ value equals to ~ 0.3652 with respect to the null model. This means that the only constants models perform 36,5% better than the null model.

3. Estimate Binary Choice Model with Explanatory Variables

a) Now, we estimate a model with the 'to adopt' (ASC_{YES}) alternative specific constant and the following explanatory variables: SPOOL, GRAD, PTRANSP, NEIGHB, PSENIORS, NCHILD and COST. For this model, the utility functions are:

$$V_{YES} = ASC_{YES} + \beta_{SPOOL}SPOOL + \beta_{GRAD}GRAD + \beta_{PTRANSP}PTRANSP + \beta_{NEIGHB}NEIGHB + \beta_{PSENIORS}PSENIORS + \beta_{NCHILD}NCHILD$$

$$V_{NO} = 0$$

Table 3: General Statistics for the a) Model

Statistic	Value
Log Likelihood	-1206.21
ρ^2 for the null-model	0.372904
$\bar{\rho}^2$ for the null-model	0.368745
ρ^2 for the constant model	0.011371
$\bar{\rho}^2$ for the constant model	0.005630

Table 4: Estimated Parameters for model a)

Parameter	Value	t-test
ASC_{YES}	-1,48229	-8,20728
β_{COST}	-0,31625	-1,76612
β_{GRAD}	0,021214	0,264361
β_{NCHILD}	0,054645	1,240122
β_{NEIGHB}	0,427169	3,631858
$\beta_{PSENIORS}$	0,219748	1,58506
$\beta_{PTRANSP}$	-0,33378	-2,63687
β_{SPOOL}	-0,01266	-0,10627

The model has a $\bar{\rho}^2$ for the constant model of 0,010474 (Table 3), which suggest that it doesn't perform much better than the constant model. All estimated parameters, with the exception of β_{SPOOL} and β_{GRAD}, resulted significant in the test ($|t| < 1.00$).

As expected, we obtained a negative value for the β_{COST} parameter, indicating that higher cost means a lower probability of choosing to adopt the electric car. We obtained positive values for the β_{NCHILD} and $\beta_{PSENIORS}$ parameters, which suggests that maybe the number of people without a stable income matters in the decision of adopt an electric vehicle. As discussed in section 1, the $NEIGHB$ variable has a great effect on the decision, and also the $PTRANSP$ has a negative effect on the utility. One could think that people that uses public transportation worries about the environment and thus, could be more interested on getting an electric car, but this is not the case.

b) In the previous model, the parameters associated with the variables SPOOL and GRAD resulted non-significant. If we drop these variables, the model to estimate is:

$$V_{YES} = ASC_{YES} + \beta_{PTRANSP} PTRANSP + \beta_{NEIGHB} NEIGHB + \beta_{PSENIORS} PSENIORS + \beta_{NCHILD} NCHILD + \beta_{COST} COST$$

$$V_{NO} = 0$$

We see now in Table 6 that all variables that all variables that were significant in the previous model remains being significant. We can use a log-likelihood ratio test, which compares the goodness of fit of both models, to see which one performs better.

Table 5: General Statistics for the b) Model

Statistic	Value
Log Likelihood	-1206.25
ρ^2 for the null-model	0.372883
$\bar{\rho}^2$ for the null-model	0.369763
ρ^2 for the constant model	0.011339
$\bar{\rho}^2$ for the constant model	0.007234

Table 6: Estimated Parameters for model b)

Parameter	Value	t-test
ASC_{YES}	-1,48278	-8,36569
β_{COST}	-0,3154	-1,76184
β_{NCHILD}	0,054351	1,233625
β_{NEIGHB}	0,427133	3,631884
$\beta_{PSENIORS}$	0,227974	1,688889
$\beta_{PTRANSP}$	-0,33385	-2,63798

Likelihood Ratio Test

$$H_0: The\ restricted\ model\ (b)\ is\ equal\ to\ the\ unrestricted\ model\ (c).$$
$$H_1: The\ restricted\ model\ (b)\ is\ different\ to\ the\ unrestricted\ model\ (c).$$

We compute the statistic,

$$Statistic = -2 * [LL_R - LL_U] = -2 * [(-1206.249) - (-1206.209\ )] = 0.08085$$

and consider the following critic value from a chi-squared distribution with $\alpha = 0.05$ and 2 degrees of freedom:

$$\chi_{0.95,2} = 5.9915$$

We see that 0.08085 < 5.9915, hence, we cannot reject the null hypothesis, and can consider the models equal and confirm that the dropped variables didn't provide any information to the model.

c) Now, we introduce some interaction variables into the model, creating dummy variables for the INCOME variable to see how it affects the importance of COST in the model. I.e. we could want to know if, for higher income levels, the COST variable has less importance.
The model to estimate is defined as follows:

$$V_{YES} = ASC_{YES} + \beta_{PTRANSP}PTRANSP + \beta_{NEIGHB}NEIGHB + \beta_{PSENIORS}PSENIORS + \beta_{NCHILD}NCHILD +$$
$$\beta_{COST}COST + \beta_{MINCCOST}MINCOME * COST + \beta_{HINCCOST}HINCOME * COST$$
$$V_{NO} = 0$$

Where MINCOME (HINCOME) is a dummy variable that takes the value 1 if the household bellows to the medium (high) income level, and 0 otherwise.

Table 7: General Statistics for the c) model

Statistic	Value
Log Likelihood	-1200.29
ρ^2 for the null-model	0.375979
$\bar{\rho}^2$ for the null-model	0.37182
ρ^2 for the constant model	0.016220
$\bar{\rho}^2$ for the constant model	0.01047

Table 8: Estimated Parameters for model c)

Parameter	Value	t-test
ASC_{YES}	-1,47013	-8,27634
β_{COST}	-0,57102	-2,89258
$\beta_{HINCCOST}$	0,401354	2,665577
$\beta_{MINCCOST}$	0,393642	3,102009
β_{NCHILD}	0,051069	1,154979
β_{NEIGHB}	0,429363	3,635429
$\beta_{PSENIORS}$	0,237258	1,752799
$\beta_{PTRANSP}$	-0,33516	-2,644

We see that both parameters included are significant (Table 8), but with positive value. The terms of the utility function that content the COST variable can be reordered as follows:

$$(\beta_{COST} + \beta_{MINCCOST} MINCOME \; + \; \beta_{HINCCOST} HINCOME) * COST$$

This can be interpreted as different values for the parameter associated with the cost variable, depending on the household income level. This interpretation is illustrated in Table 9.

Table 9: Interpretation of parameters for the INCOME – COST Interaction

Income Level	Expression	Value
Low	β_{COST}	-0,57102
Medium	$\beta_{COST} + \beta_{MINCCOST}$	-0,17738
High	$\beta_{COST} + \beta_{HINCCOST}$	-0,16967

Now, we can see that the cost has a negative effect on the utility values for all income levels, but the medium and high-income levels are clearly less sensitive to the cost than the low-income level. This is adding more information to the model, as we can see in Table 7, the rho-square-bar for the constant model increases ~5% if we compare with the previous model in b).

To test if this increment is significative, we can use the same ratio test that we use before, but now to compare models in c) and in b). For this test, the statistic is:

$$Statistic \; = \; -2 \; * \; [LL_R - LL_U] \; = \; -2 \; * \; [(-1206.249) - (-1200.293 \,)] \; = 11.9116$$

and we consider the same critic value from before:

$$\chi_{0.95,2} = 5.9915$$

This time, the statistic has a greater value than the critic value, and H_0 can be rejected. This way, we conclude that the model in c) is better in terms of goodness of fit.

4. Searching for Your Final Specification

a) To achieve the task of searching for the best specification, a forward pass approach was used. This was done considering all variables in the data set, plus some interaction terms, like the interaction between the WALKBIKE and NBIKES variables that was discussed in section 1, and also the INCOME – COST interactions used in section 2. Some other variables, where included in the analysis, such as LDIST that took value 1 if the distance to the nearest shopping center was greater than 4 miles, and 0 otherwise; and HCHILD, the one proposed. Other considerations, such as the linear dependence between the variables HHSIZE, WOMEN and MEN were taken into account when estimating the model. Also, some variables were transformed, ETHN was converted in three dummy variables, and POPDIST was scaled by dividing it by its maximum values. In each phase, the significance of the parameter and the rho-square-bar for the null-model, were used to decide if the variable entered the model or not. Using this approach, the best model found is:

$$\begin{aligned} V_{YES} = ASC_{YES} &+ \beta_{WOMEN} WOMEN + \beta_{AFRICAN} AFRICAN + \beta_{PTRANSP} PTRANSP + \beta_{POPDENSE} POPDENSE \\ &+ \beta_{SMARTPH} SMARTPH + \beta_{CABLE} CABLE + \beta_{NEIGHB} NEIGHB + \beta_{SMARTPH} SMARTPH \\ &+ \beta_{HCHILD} HCHILD + \beta_{WBIKE-NBIKES} WALKBIKE * NBIKES + \beta_{LINCCOST} LINC * COST \\ &+ \beta_{COST} COST \end{aligned}$$

$$V_{NO} = 0$$

Table 10: General Statistics for the final model

Statistic	Value
Log Likelihood	-1187.12
ρ^2 for the null-model	0.382826
$\bar{\rho}^2$ for the null-model	0.376587
ρ^2 for the constant model	0.027014
$\bar{\rho}^2$ for the constant model	0.021259

Table 11: Estimated Parameters for the final model

Parameter	Value	t-test
ASC_{YES}	-1,60365	-8,29043
$\beta_{AFRICAN}$	0,431169	2,326773
β_{CABLE}	0,199776	1,74446
β_{COST}	-0,20323	-1,09745
β_{HCHILD}	-0,16112	-1,39432
$\beta_{LINCCOST}$	-0,31902	-2,50988
β_{NEIGHB}	0,443073	3,721465
$\beta_{POPDENSE}$	-0,40343	-1,91767
$\beta_{PTRANSP}$	-0,34725	-2,72048
$\beta_{SMARTPH}$	0,13113	2,389468
$\beta_{WBIKE-NBIKES}$	0,163967	1,807834
β_{WOMEN}	0,084956	1,856009

Note for the interaction between COST and INCOME that, since in the previous model the parameters for medium and high income were similar, it was preferred to include a dummy variable for the low-income level (LINC), to prevent to include too much variables in the model.
With respect to the coefficients, can be noted that African people, has more probability of adopting the electric car. Also, the signs of the $\beta_{SMARTPH}$ and β_{CABLE}, may imply that households that are more familiar with technology are more likely to accept the vehicle. Finally, the fact that a household has children has a negative effect in the decision.

b) The marginal utility of cost is given by:

$$\frac{\partial \boldsymbol{V_{YES}}}{\partial \boldsymbol{COST}} = \boldsymbol{\beta_{COST}} + \boldsymbol{\beta_{LINCCOST}} \boldsymbol{LINC} = -0{,}20323 - 0{,}31902 * LINC$$

We can note that, for households of low-income levels, the marginal utility of cost has a greater negative effect when comparing to the medium and high-income levels. This means, as was said before, that lower income levels are more sensitive to the cost of the car.

5. Aggregate Effect on Adoption shares Due to Level-of-Service Changes

To determine the impact of a reduction in the cost of the electric vehicles in 25%, we compute, for each household, the probabilities of choosing to adopt the electric car in both cases, with the original cost and with the reduced cost. This is done by, passing the value of the variables of each observation though the utility functions for the model in 4.a), and using the following expression to compute the probability:

$$P_{YES} = \frac{exp(V_{YES})}{exp(V_{YES}) + exp(V_{NO})}$$

Then, we compute the difference in the probabilities of acceptation in both cases, for all observation, and take the mean of the difference, obtaining, in average an increment of 0,9% of choosing to adopt the electric car. Also, if we disaggregate this percentage by income level, we see that households of lower incomes have a greater increase in the probability of ~1.36% in comparison with the medium and high-income levels, with ~0,66% and ~0,70% respectively. We can say that low-income households are more sensitive to the cost than medium and high-income households.