Travel Demand Analysis
Exercise 4

310702051 葉家榮

2022-05-10

This assignment is also uploaded in the following website, for detailed information and code.
https://chiajung-yeh.github.io/Travel-Demand-Analysis/discrete-choice-modeling.html

Problem

You are provided with a data set, from a survey of 210 individuals' choices of travel mode between Sydney, Melbourne and New South Wales. There are four alternative choices, along with four choice-specific covariates for each choice. The variable definition is provided as follows:

- Mode = choice; Air, Train, Bus, or Car
- Ttme = terminal waiting time, 0 for car
- Invc = in vehicle cost
- Invt = travel time, in vehicle
- GC = generalized cost measure
- Hinc = household income
- Psize = party size in mode chosen

Use Apollo to answer the following questions. Note that you need to transform the data set format from long to wide and add an ID variable.

Data Transformation

Note that the provided data is long format, which the attributes of each alternative are listed in different rows. However, the required format for developing model by using `apollo` package is wide, and thus, the data should first be transformed. Also, ID variable is required in the model, to identify each of the respondent.

Before transforming the data, it is better to observe the original data in advance. Table 1 shows the long data format (original) of a respondent, the column "MODE" represents Air, Train, Bus, and Car respectively. If it is coded as 1, it means that the respondent chooses that mode. Thus, every 4 rows belongs to a respondent, and there must exists only one "1" in "MODE" column. Take Table 1 for instance, the first row records the attributes of mode "Air", and its terminal waiting time (TTME) is 69 minutes, while the second row records the attributes of "Train", and its in vehicle cost (INVC) is 31 dollars, and so forth. Also, note that terminal waiting time (TTME), vehicle cost (INVC), in vehicle travel time (INVT), generalized cost (GC) are often regarded as generic variables, which have identical impacts on the utility for different modes. Conversely, household income (HINC) is often considered to be alternative specific variable, which has a different impact on utility by modes.

As the illustration above, the long format should be transformed to wide and add the ID variable before developing the choice model by means of `apollo` package. Table 2 shows part of the wide format data after

Table 1: Example for Long Data

MODE	TTME	INVC	INVT	GC	HINC	PSIZE
0	69	59	100	70	35	1
0	34	31	372	71	35	1
0	35	25	417	70	35	1
1	0	10	180	30	35	1

Table 2: Example for Wide Data (Part of columns, attributes of Air and Train)

ID	TTME.Air	INVC.Air	INVT.Air	GC.Air	TTME.Train	INVC.Train	INVT.Train	GC.Train	MODE
1	69	59	100	70	34	31	372	71	Car

being reshaped. For each of row, it represents one of the attributes and the final choice of a respondent. For instance, for respondents ID "1", the terminal waiting time (TTME) of "Air" is 69 minutes, while it takes 34 minutes for "Train". Also, the final choice is shown in the last column of Table 2. For respondents ID "1", he chooses "Car".

Model without Intercept

Problem Description

Run a model with generalized cost and in-vehicle time, without intercepts.
(1) Do the estimated coefficients have the expected signs?
(2) Are both coefficients significantly different from zero?
(3) How closely do the average probabilities match the shares of travelers choosing each alternative?
(4) The ratio of coefficients usually provides economically meaningful information. The willingness to pay (wtp) through higher travel cost for a one-minute reduction in travel time is the ratio of the travel time coefficient to the travel cost coefficient. What is the wtp from this model? Is it reasonable in magnitude?

Model Result

Consider a model only with generalized cost and in-vehicle time, the model (**Model 1**) can be formulated as below.

$$V_{mode} = \beta_{GC} * GC_{mode} + \beta_{INVT} * INVT_{mode}$$

The result of Model 1 is shown in Table 3. The estimation of β_{GC} (`b_gc` in the table) is -0.0124, which means a higher generalized cost would cause a negative utility for the mode. Also, the t-value of β_{GC} suggests that the coefficient is significantly different from 0. It is reasonable, for people would not be inclined to choose the mode as the cost increases. On the other hand, the coefficient of β_{INVT} (`b_invt` in the table) is also negative as expected though, the statistics test shows that it cannot significantly reject the null hypothesis. To sum up, the estimated coefficients have the expected signs, and particularly for the coefficient of generalized cost ***(1, 2)***.

$LL(0)$ of Model 1 is -291.12, while $LL(C) = LL(\beta)$ is -279.74 and thus $\rho^2 = 0.039$, which implies there is only a slight difference between log-likelihoods when just take generalized cost and in-vehicle travel time into consideration. The contingency table and market share of prediction and real mode are illustrated in Table

Table 3: Model 1 Result

	Estimate	s.e.	t.rat.(0)	Rob.s.e.	Rob.t.rat.(0)
constant	0.0000	NA	NA	NA	NA
b_gc	-0.0124	0.0037	-3.3101	0.0045	-2.7694
b_invt	-0.0004	0.0003	-1.3438	0.0004	-1.1661

Table 4: Contigency Table of Model 1

		Prediction				
		Air	Bus	Car	Train	Percentage
REAL	Air	21	2	35	0	**27.6 %**
	Bus	19	4	7	0	**14.3 %**
	Car	29	0	30	0	**28.1 %**
	Train	33	0	26	4	**30 %**
	Percentage	**48.6 %**	**2.9 %**	**46.7 %**	**1.9 %**	**100 %**

4. From the table, we can find that there is a huge gap between market share of the real and prediction value. The predictive market share of "Air" as well as "Car" are severely over-estimated, while "Bus" and "Car" has an extreme low proportion. The accuracy of prediction of the mode is merely 28.1 %*(3)*.

The willingness to pay through higher travel cost for a one-minute reduction in travel time is the ratio of the travel time coefficient to the travel cost coefficient. It can be calculated as below:

$$\frac{-0.0004}{-0.0124} = 0.0322$$

, which means the willingness to pay for one-minute reduction travel time costs \$0.0322, that is approximately NT\$ 0.9, namely NT\$ 54 an hour. It is very unreasonable, since the estimated value is much lower than the wage*(4)*.

Model with Constants

Problem Description

Add alternative-specific constants to the model. Normalize the constant for the alternative bus to 0.
(1) How well do the estimated probabilities match the shares of travelers choosing each alternative?
(2) Calculate the wtp. Is it reasonable?

Model Result

Consider a model with alternative-specific constants, generalized cost and in-vehicle time, the model (**Model 2**) can be formulated as below.

$$V_{mode} = \beta_{mode} + \beta_{GC} * GC_{mode} + \beta_{INVT} * INVT_{mode}$$

Table 5: Model 2 Result

	Estimate	s.e.	t.rat.(0)	Rob.s.e.	Rob.t.rat.(0)
asc_car	0.3016	0.2400	1.2568	0.2504	1.2044
asc_train	0.8402	0.2373	3.5411	0.2232	3.7648
asc_air	-0.6861	0.5103	-1.3445	0.6207	-1.1054
asc_bus	0.0000	NA	NA	NA	NA
b_gc	-0.0117	0.0052	-2.2694	0.0062	-1.9048
b_invt	-0.0022	0.0009	-2.4051	0.0011	-1.9889

Table 6: Contigency Table of Model 2

		Prediction				
		Air	Bus	Car	Train	Percentage
REAL	Air	20	0	14	24	**27.6 %**
	Bus	19	1	0	10	**14.3 %**
	Car	24	0	29	6	**28.1 %**
	Train	18	0	3	42	**30 %**
	Percentage	**38.6 %**	**0.5 %**	**21.9 %**	**39 %**	**100 %**

The result of Model 2 is shown in Table 5. The estimation of β_{GC} (`b_gc` in the table) is -0.0117, while the coefficients of in-vehicle travel time is -0.0022, and both of the t-statistics suggest the null hypothesis be rejected. It implies a higher cost or longer travel time would result in the lower utility of the mode, and people would tend to choose other alternatives.

In addition, “Train” is the only alternative-specific constants that its t-value suggests reject the null hypothesis. It implies that there is a significant difference on utility between “Train” and “Bus” under other conditions (generalized cost and in-vehicle travel time) to be fixed. However, there is no apparent difference between the constants of other modes and that of the “Bus”.

$LL(0)$ of Model 2 is -291.12, while $LL(\beta)$ is -266.94 and thus $\rho^2 = 0.083$, which is a little improved compared to Model 1. The contingency table and market share of prediction and real mode are illustrated in Table 6. From the table, we can find that there still exists a large gap between the real market share and the prediction one. The predictive market share of “Car” is about 21.9 %, which is relative close to the real proportion (28.1 %). However, the model wrongly estimates the market share of “Bus” which accounts for 14.3%, but the estimation is only 0.5 %. The overall accuracy of Model 2 is approximately 43.8 %***(1)***, it is indeed not ideal for the estimation result, but much higher than that of Model 1.

In terms of willingness to pay of one-minute reduction in travel time, the calculation is listed below:

$$\frac{-0.0022}{-0.0117} = 0.1880$$

, which means the willingness to pay is \$0.1880 per minute (\$11.28 per hour). It may be reasonable, since the current minimum wage for most of states in the U.S is around \$7.25 per hour to \$15 per hour, and \$11.28 per hour is in the range *(2)* (BUT unfortunately, we have no clue to prove that the wage is reasonable at the time data collected).

Table 7: Model 3 Result

	Estimate	s.e.	t.rat.(0)	Rob.s.e.	Rob.t.rat.(0)
asc_car	0.4111	0.2348	1.7507	0.2385	1.7234
asc_train	0.6920	0.2262	3.0591	0.2087	3.3149
asc_air	-1.0767	0.4877	-2.2077	0.6499	-1.6566
asc_bus	0.0000	NA	NA	NA	NA
b_gc_inc	-0.0735	0.0838	-0.8772	0.0921	-0.7977
b_invt	-0.0032	0.0008	-3.9439	0.0011	-3.0196

Models with sociodemographic variables

Problem Description

Now try some models with sociodemographic variables entering.
(1) Enter generalized cost divided by household income, instead of generalized cost. With this specification, the magnitude of the generalized cost is inversely related to household income, such that high income households are less concerned with generalized travel costs than lower income households. Does dividing generalized cost by income seem to make the model better or worse?
(2) Instead of dividing generalized cost by household income, enter alternative-specific generalized cost effects. Do these generalized cost terms enter significantly?
(3) Try other models. Determine which model you think is best from these data.

Model Result

Consider a model with generalized cost divided by household income, the model (**Model 3**) can be formulated as below.

$$V_{mode} = \beta_{mode} + \beta_{GC} * (\frac{GC_{mode}}{HINC_r}) + \beta_{INVT} * INVT_{mode}$$

, where r is for every respondent (sample).

The result of Model 3 is shown in Table 7. The estimation of generalized cost divided by household income (`b_gc_inc` in the table) is -0.0735; however, its t-value is only -0.8772, meaning that the null hypothesis that the coefficient is 0 cannot be rejected. The coefficients of in-vehicle travel time is -0.0032, and the t-statistics (-3.9439) suggest the null hypothesis be rejected. It implies a longer travel time would significantly result in the lower utility of the mode.

In this model, if the generalized cost divided by household income is larger, it indicates that the travel cost are not affordable for that target group. By this concept, we consider that this new variable would have a more negative impact on the utilities. However, the statistic test cannot prove this assumption, and $LL(\beta)$ of Model 3 is -269.18, slightly worsen than Model 2. In conclusion, dividing generalized cost by income seem have no significant benefits on the model result ***(1)***.

Now, taking all the general cost as alternative-specific variables, which means to consider the impact of general cost on each mode is totally different. From practical perspective, it may be true, since the "feelings" of increasing the cost for different mode might not be the same, and thus the change of utility might have a difference. Based on this assumption, the model (**Model 4**) can be formed as below.

$$V_{mode} = \beta_{mode} + \beta_{GC_{mode}} * GC_{mode} + \beta_{INVT} * INVT_{mode}$$

Table 8: Model 4 Result

	Estimate	s.e.	t.rat.(0)	Rob.s.e.	Rob.t.rat.(0)
asc_car	0.4284	0.5607	0.7640	0.5911	0.7248
asc_train	1.4311	0.5377	2.6617	0.5630	2.5418
asc_air	-0.6012	0.7383	-0.8143	0.8048	-0.7470
asc_bus	0.0000	NA	NA	NA	NA
b_gc_car	0.0004	0.0051	0.0703	0.0055	0.0654
b_gc_train	-0.0065	0.0040	-1.6498	0.0041	-1.5971
b_gc_air	-0.0061	0.0076	-0.8085	0.0111	-0.5538
b_gc_bus	0.0000	NA	NA	NA	NA
b_invt	-0.0036	0.0012	-2.9728	0.0018	-2.0116

The result of Model 4 is shown in Table 8. Note that the constant term and generalized cost coefficient of bus is regarded as base (0). We can find that the generalized cost coefficients are not significantly different from 0, that is, the impact of generalized cost on utility are not different among modes ***(2)***. It implies that taking generalized cost as alternative-specific variable might not be a good choice, instead, regarding it as the generic variable is appropriate.

Last, develop a model that can well predict the market share of each mode. Here, we use terminal waiting time (TTME), in-vehicle cost (INVC), in-vehicle travel time (INVT), and household income (HINC) to be the independent variable. Note that the household income is used to be as alternative-specific variables, while others are the generic variables. The model (**Model 5**) is formed as below.

$$V_{mode} = \beta_{TTME} * TTME_{mode} + \beta_{INVC} * INVC_{mode} + \beta_{INVT} * INVT_{mode} + \beta_{HINC_{mode}} * HINC_r$$

, where r is for every respondent (sample).

The result of Model 5 is shown in Table 9. In Table 9, first observe the alternative-specific constants, we can find that "Car" and "Train" have a significant difference compared to "Bus". Under same conditions, the utility of "Car" is less than that of the "Bus", while the utility of "Train" is higher. And the alternative-specific constants of "Air" seems to have no significant difference between that of "Bus". As for the terminal waiting time, in-vehicle travel time, they are all significantly cause a negative impact on utility. And interestingly, the coefficient of terminal waiting time is more negative than that of in-vehicle travel time (-0.0957<-0.0036, note that the unit of these two variables are the same). It does make sense, waiting time is much more unaffordable than staying in the vehicle, for the out of vehicle environment, which may influenced by the weather and the crowds, is definitely not comfortable relative to the one in the vehicle. As for the household income, we can find that only "Train" has a significant negative sign compared to "Bus", which means that if the income is large, people would tend not to choose "Train".

$LL(0)$ of Model 5 is -291.12, while $LL(\beta)$ is -182.41 and thus $\rho^2 = 0.373$, which is hugely improved compared to all of the previous model. The contingency table and market share of prediction and real mode are illustrated in Table 10. From the table, we can find that the predictive market share is much closer to the real one, indicating that this model might have a good estimation on the mode choice. The overall accuracy of Model 5 is approximately 74.3 %, which is pretty well among all the models developed***(3)***.

Table 9: Model 5 Result

	Estimate	s.e.	t.rat.(0)	Rob.s.e.	Rob.t.rat.(0)
asc_car	-4.0296	0.6830	-5.8999	0.6440	-6.2570
asc_train	1.4113	0.5554	2.5410	0.4922	2.8673
asc_air	-0.0401	0.8223	-0.0488	0.9132	-0.0439
asc_bus	0.0000	NA	NA	NA	NA
b_ttme	-0.0957	0.0104	-9.2386	0.0141	-6.8071
b_invt	-0.0036	0.0009	-4.1686	0.0011	-3.2719
b_invc	0.0000	0.0236	0.0000	0.0015	0.0000
b_hinc_car	0.0253	0.0156	1.6252	0.0128	1.9748
b_hinc_train	-0.0349	0.0167	-2.0878	0.0154	-2.2724
b_hinc_air	0.0233	0.0164	1.4186	0.0139	1.6764
b_hinc_bus	0.0000	NA	NA	NA	NA

Table 10: Contigency Table of Model 5

		Prediction				
		Air	Bus	Car	Train	Percentage
REAL	Air	40	0	13	5	**27.6 %**
	Bus	1	23	4	2	**14.3 %**
	Car	7	0	43	9	**28.1 %**
	Train	6	1	6	50	**30 %**
	Percentage	**25.7 %**	**11.4 %**	**31.4 %**	**31.4 %**	**100 %**