CE 264 Problem Set 2: Specification, Estimation, and Testing of Binary Choice Models

Franklin Zhao (3033030808)
Ruitong Zhu (3033103852)

8 Feb. 2018

Part 1 Model Development

1. What are your a-priori expectations about the relationships between the variables in the dataset and the choice of one itinerary versus another? How are the various variables expected to affect the utilities of each itinerary?

In our a-priori expectations, the choice of itineraries depends on a linear combination of some important variables in the dataset, including air fare, travel time, number of connections and the FFP membership of a certain company. These are basically the factors we would take into account when choosing an itinerary. Since we prefer flights with fewer connections, less travel time and lower price, we would expect these three factors are negatively related to the utility of a certain flight. Also we would expect people to choose flights provided by the company of which they have membership. Therefore the membership should affect the utility positively.

2. What model specification corresponds to your a-priori hypotheses? Show the equation.

The utility equation for our a-priori hypotheses is shown in Equation (1), which is a linear combination of a few factors with generous parameters across alternatives.

$$U_a = \beta_C \times Connections + \beta_T \times TravelTime + \beta_F \times Fare + \beta_M \times Membership \tag{1}$$

3. What is your "best" model specification? Show the equation and use well-defined variables with meaningful names with a brief description of each variable.

The equation and variables of the "best" model is shown in Equation (2) and Table 1.

$$\begin{aligned} U_{b1} = \ & \beta_c \times Connections_{A1} + \beta_m \times Membership_{A1} + \beta_d \times Departure + \beta_a \times \\ & Affordability + \beta_v \times UnitePrice + \beta_p \times Performance + \beta_t \times Time \\ U_{b2} = \ & \beta_c \times Connections_{A2} + \beta_m \times Membership_{A2} + \beta_d \times Departure + \beta_a \times \\ & Affordability + \beta_v \times UnitePrice + \beta_p \times Performance + \beta_t \times Time \end{aligned} \tag{2}$$

Variable Description:

$\mathbf{Connections_{Ai}}$: With alternative specific parameters, representing the number of connections

during the flight for alternative i.
Membership$_{\mathbf{Ai}}$: With alternative specific parameters, representing the level of FFP membership for corresponding airline for alternative i (1: Not a member, 2: Basic member, 3: Elite member).
Departure: Departure time for the flight.
Affordability: $Fare/Income^{Payment}$, a combined variable describing whether the ticket is affordable for a certain person. *Payment* is categorical: 1 = I paid, personally, 2 = My company paid or reimbursed me 3 = It was free through the airline (either through a frequent flyer program, a voucher or from getting bumped), 4 = Family or friend, 5 = Other.
UnitPrice: $Fare/ClassTicket$, a combined variable describing the price of the ticket, taking into account the class of the ticket. *ClassTicket* is categorical: 1 = Economy or coach, 2 = Premium economy, 3 = Business, 4 = First Class.
Performace: $Performance/Purpose$, a variable describing the on-time performance of an airline and whether a customer values it. *Purpose*: Trip purpose, categorical variable: 1 = Business, 2 = Attend conference, 3 = Vacation, 4 = Visit friends or relatives, 5 = Attend school, 6 = Other.
Time: $TravelTime \times Age$, a variable describing the modulated travel time. *Age*, categorical variable: 1 = 15-19, 2 = 20-24, 3 = 25-34, 4 = 35-44, 5 = 45-54, 6 = 55-64, 7 = 65-74, 8 = 75 years or older.

Table 1: Variables for the "best" model

Name	Unit	Estimate	T-stats	P-value
Connections$_{Ai}$		A1: -0.6651; A2: -0.6755	12.639	0.000
Membership$_{Ai}$		A1: 0.4876; A2: -0.4134	10.271	0.000
Departure	Hours after midnight	0.0283	3.063	0.002
Affordability		0.0023	2.740	0.006
UnitPrice	\$100	-0.0085	25.660	0.000
Performance		0.2335	7.626	0.000
Time	Hours	-0.0527	7.072	0.000

*Variables are generous across the alternatives if not specified.

4. Does your "best" model match your a-priori expectations?

Those two models are not perfect match but they do share some common factors. For example, both of the models have the variable 'Membership' and 'Connections'.

5. Does your "best" model include the factors that you think are important in the decision?

Yes. The "best" model includes all the four factors we have in our a-priori hypotheses though in a different way. As mentioned above, both models share the some identical variables. Besides, for the variables "Fare" and "travel_time", these two factors also appear in the model but are converted as a part of a newly combined variable. For example, the factor "Fare" has been combined with another factor "income" into a new factor "Affordability" of the linear model.

Yet the "best" model also excludes some factors we think might also be influential, though might be not as important as others, like the "arrival_time", since arrival time can be something we would consider. we prefer to take flights arriving in the afternoon or evening, especially for a trip.

6. What heterogeneity is captured in your model? What heterogeneity is not captured by your model? Is heterogeneity adequately represented in your model?

The heterogeneity captured in the model includes the age, income, purpose, payment method, FFP membership and class of tickets of customers, and the only heterogeneity factors we neglected in the model is gender.

We think that the heterogeneity has already been quite adequately represented in the model, since from our perspective, flight choice is indifferent in terms of gender and all the factors besides have been considered.

7. What issues did you come across while developing your “best” model?

The biggest issue while developing our “best” model is how to combine different factors into a valid variable of the linear equation, if not consider the inconvenience of the “pylogit” package.

89. How did you use t-tests, likelihood ratio tests, and goodness-of-fit comparisons to arrive at your final model? Show how you used one t-test and how you used one likelihood ratio test to choose between different model specifications. Include your calculations.

T-test: Here we will give an example on trying to add another variable “departure time” and “arrival time” to the a-prior model, and the outcome of the new model is shown in Equation (3) and Table 2.

$$\begin{aligned} U_t = \quad & \beta_C \times Connections + \beta_T \times TravelTime + \beta_F \times Fare + \beta_M \times Membership \\ & + \beta_d \times Departure + \beta_a \times Arrival \end{aligned} \tag{3}$$

Table 2: T-test results

Variable	T-stat
Departure time	2.665
Arrival time	0.324

Taking the 5% significant value where t_{cr}=1.96, for the variable “departure time” we would reject the null hypotheses that “β_d=0, but for the variable “arrival time” we cannot reject the null hypotheses. Therefore we would include the first variable in the model and exclude the other.

Likelihood-ratio-test: The original a-priori model has generous coefficients because the alternatives are all itineraries. To verify this, we then do the likelihood ratio tests on those variables and here is an example for the variable “connections.

$$\begin{aligned} U_{l1} = \quad & \beta_{C1} \times Connections + \beta_T \times TravelTime + \beta_F \times Fare + \beta_M \times Membership \\ U_{l2} = \quad & \beta_{C2} \times Connections + \beta_T \times TravelTime + \beta_F \times Fare + \beta_M \times Membership \end{aligned} \tag{4}$$

Equation (4) is the unrestricted model, the null hypothesis of which is “$\beta_{C1} = \beta_{C2}$”.

Now Let us compare this model with the a-priori model. The likelihood-ratio test statistic for the null hypothesis of generic attributes is given by:

$$\chi^2 = -2(L(\hat{\beta_G}) - L(\hat{\beta_{AS}})) = -2(-3968.984 + 3964.356) = 9.256 \tag{5}$$

Since $\chi^2 > \chi^2_C = 5.991$, we can reject the null htpothesis. So in the refined model the coefficient of the variable “connections” remains to be alternative specific.

Goodness of fit comparisons: The statistic we pick as a criteria for goodness-of-fit comparisons is the adjusted likelihood ratio index ($\overline{\rho}^2$). Based on the given equation, a higher index indicates a better fit. So while exploring different specifications, the policy is to always pick one with higher index. For the four models mentioned above we have indexes listed in the Table 3.

Table 3: Four models comparision

Model U_a	$\overline{\rho}^2$
A-priori hypotheses U_a	0.184
U_t for t-test	0.184
U_{l1}/U_{l2} for likelihood-ratio-test	0.185
“best” model U_{b1}/U_{b2}	0.227

According to Table 3, currently the best model can reach 0.227 in terms of $\overline{\rho}^2$.

Part 2 Research Project

1. As a binary choice problem, what would be your dependent variable?

Research question: How will driving costs influence the choice of transportation mode?

First let us frame the research problem in terms of a binary choice, and be specific about the transportation mode. Let us say, whether driving costs at Berkeley will influence a Berkeley’s graduate student’s choice on driving or riding a bike to campus everyday, assuming the distance is fair enough and there is no other transportation mode (e.g., no transit, no BART, and too far to walk). Then the dependent variables will be $\{B, C\}$, where B is to choose riding a bike, and C is to choose driving a car.

2. What independent variables do you think influence the choice?

The independent explanatory variables include parking fee, fuel cost, bike cost, maintenance cost, time cost, traffic condition, whether there are hills between campus and home of the decision maker, characteristics of the decision maker, financial status of the decision maker, and physical condition of the decision maker.

3. Specify the utilities of the two alternatives.

The utilities of the two alternatives are shown in Equation (6)

$$\begin{aligned} U_B &= \beta_0 + \beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7 + \beta_8 X_8 \\ U_C &= \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_9 X_9 + \beta_{10} X_{10} \end{aligned} \tag{6}$$

Variables specification:
$\mathbf{X_1}$**:** parking fee, which can be specified as money spent on parking every month.
$\mathbf{X_2}$**:** fuel cost, similarly can be specified as money spent on car fuel every month.
$\mathbf{X_3}$**:** bike cost (money spent on the bike).
$\mathbf{X_4}$**:** maintenance cost. The cost for car maintenance every month (parking and fuel are not included).
$\mathbf{X_5}$**:** time cost. The average time spent getting to the campus everyday.
$\mathbf{X_6}$**:** traffic condition, which can be specified as the average traffic density of the regular route to campus.
$\mathbf{X_7}$**:** hills, which can be specified as the length of the road that has a slope angle greater than 10^o.
$\mathbf{X_8}$**:** characteristics, which can be specified as the level of patience (0–5; “5” is the most patient).
$\mathbf{X_9}$**:** financial status (0–5; “5” has the best financial status).
$\mathbf{X_{10}}$**:** physical condition (0–5; “5” has the best physical condition).

4. What are your a priori expectations of the coefficients for each specified variable?

$\beta_\mathbf{0}$ is the alternative specific constant, which captures the difference of U_B and U_C when all else are equal. $\beta_\mathbf{1}$ is a negative number.
$\beta_\mathbf{2}$ is a negative number.
$\beta_\mathbf{3}$ is a negative number.
$\beta_\mathbf{4}$ is a negative number.
$\beta_\mathbf{5}$ is a negative number.
$\beta_\mathbf{6}$ is a negative number.
$\beta_\mathbf{7}$ is a negative number.
$\beta_\mathbf{8}$ is a positive number.
$\beta_\mathbf{9}$ is a positive number.
$\beta_\mathbf{10}$ is a positive number

5. How are you addressing the research question using the model specification?

For our research question, we can use the hypothesis testing. The null hypothesis will be that the driving costs will mot influence a Berkeley’s graduate student’s mode choice, and the alternative hypothesis will be the opposite. Then we use t-test and p-value method to test coefficients β_1, β_2, and β_4 (which corresponds to the driving cost variables X_1, X_2, and X_4). If we reject the null hypothesis, then the driving cost will increase the probability that a decision maker chooses riding a bike. If we fail to reject the null hypothesis, then driving cost may be a irrelevant factor, which has no impact on mode choice.

Part 3 Supplemental Problem

For the nth shopper:

$$\begin{aligned} U_{1n} = &\quad \beta_1(distance_{1n}) + \beta_2 ln(size_1) + \epsilon_{1n} \\ U_{2n} = &\quad \beta_1(distance_{2n}) + \beta_2 ln(size_2) + \epsilon_{2n} \end{aligned} \tag{7}$$

Since ϵ_1 and ϵ_2 – iid EV, then:

$$\begin{aligned} P_n(1|X) &= \frac{exp(V_{1n})}{exp(V_{1n})+exp(V_{2n})} \\ &= \frac{1}{1+exp(V_{2n}-V_{1n})} \\ &= \frac{1}{1+exp(\beta_1(distance_{2n}-distance_{1n})+\beta_2 ln\left(\frac{size_2}{size_1}\right))} \end{aligned} \tag{8}$$

$$\begin{aligned} P_n(2|X) &= \frac{exp(V_{2n})}{exp(V_{1n})+exp(V_{2n})} \\ &= \frac{1}{1+exp(V_{1n}-V_{2n})} \\ &= \frac{1}{1+exp(\beta_1(distance_{1n}-distance_{2n})+\beta_2 ln\left(\frac{size_1}{size_2}\right))} \end{aligned} \tag{9}$$

Hence, the expected number of trips taken each day to store 1 and store 2 will be:

$$\begin{aligned} E_1 \quad = &\quad \int_0^{d_{AB}} \frac{200}{d_{AB}} \times \frac{1}{1+exp(\beta_1(x+d_{BC}+d_{CD}-(x+d_{BC}))+\beta_2 ln\left(\frac{size_2}{size_1}\right))} dx + \\ &\quad \int_0^{\infty} \gamma e^{-\gamma x} \times \frac{1}{1+exp(\beta_1(x-(x+d_{CD}))+\beta_2 ln\left(\frac{size_2}{size_1}\right))} dx \\ = &\quad \frac{200}{1+exp(\beta_1 d_{CD})\left(\frac{size_2}{size_1}\right)^{\beta_2}} + \int_0^{\infty} \gamma e^{-\gamma x} \frac{1}{1+exp(-\beta_1 d_{CD})\left(\frac{size_2}{size_1}\right)^{\beta_2}} dx \end{aligned} \tag{10}$$

$$\begin{aligned} E_2 \quad = &\quad \int_0^{d_{AB}} \frac{200}{d_{AB}} \times \frac{1}{1+exp(\beta_1(x+d_{BC}-(x+d_{BC}+d_{CD}))+\beta_2 ln\left(\frac{size_1}{size_2}\right))} dx + \\ &\quad \int_0^{\infty} \gamma e^{-\gamma x} \times \frac{1}{1+exp(\beta_1(x+d_{CD}-x)+\beta_2 ln\left(\frac{size_1}{size_2}\right))} dx \\ = &\quad \frac{200}{1+exp(-\beta_1 d_{CD})\left(\frac{size_1}{size_2}\right)^{\beta_2}} + \int_0^{\infty} \gamma e^{-\gamma x} \frac{1}{1+exp(\beta_1 d_{CD})\left(\frac{size_1}{size_2}\right)^{\beta_2}} dx \end{aligned} \tag{11}$$

Since the only part contains x in the second terms of E_1 and E_2 is $\gamma e^{-\gamma x}$, we can derive the following equation:

$$\begin{aligned} &\int_0^{\infty} \gamma e^{-\gamma x} dx \\ = &\ \gamma \int_0^{\infty} e^{-\gamma x} dx \\ = &\ -e^{-\gamma x}|_0^{\infty} \\ = &\ 1 \end{aligned} \tag{12}$$

Turns out it is a constant. Hence, the value of γ will not affect the expected number of trips.

Contributions

Franklin Zhao: Part 2 & 3
Ruitong Zhu: Part 1