CE264 –Spring 2017
Problem Set 2: Specification, Estimation, and Testing of Binary Choice Models

Due: Thursday, February 8, 2017
(via electronic submission on bCourses)

As with most of the assignments in this class, this assignment includes three parts: The first deals with real data; for this assignment you will estimate binary choice models . The second part will help you move forward on the research question or policy of interest that you wish to address with the methods that you learn in this course . The third part involves more traditional homework problems where you answer well-defined questions.

The problem set is to be done in groups of 2 or 3 currently enrolled CE264 students with one solution being turned in by the group (i.e. choose one group member to upload the solution to bCourses). You must list the student ID number (SID) of each team member on homework submissions.

Part 1 (70 Points): Model Development

Objectives

The objectives are for you to become familiar with the basic foundations of discrete choice analysis, focusing here on the simple case of binary choice. As emphasized throughout this course, this means both understanding the theory and also being able to work with real data. You will learn how to use the estimation software , Pylogit , in this assignment . You will also gain experience on the art of model development , including basic statistical tests used in the binary choice model specification process. We've provided the python script for a very basic model specification file (binomialLogit 01.ipynb). Use this script as a starting point for any model specification that you wish to examine.

Dataset: Air Travel Survey

The Air Travel Survey dataset was collected by travel demand modeling company RSG, Inc. in 2010 to model the behavior of air travelers when purchasing air itineraries . The role of Frequent Flyer Program (FFP) membership in choosing flight itineraries was of particular interest for this survey. Individuals were asked about a recent air trip they made, during which background information was gathered , such as income , trip purpose , and FFP membership . Each person was then asked to choose between two hypothetical itineraries for the trip they recently made . This process was repeated eight times for each person . Each record represents a single choice made by a person and includes information such as travel time , number of connections , and air fare . While the original survey collected much more information about each individual , this assignment makes use of a subset of the complete dataset. The dataset is further detailed in Appendix C.

How to Approach Model Development

First think about the behavior and your a-priori expectations about the influence of the explanatory variables. How are these factors likely to affect the utilities? You may find it useful to do some preliminary statistical analysis on the data for this. Then specify and estimate a model that reflects your a-priori hypotheses. It is generally a good idea to start simple with the most important aspects of the model , and then work to refine the model specification.You will explore many different specifications.

For example, a basic model may be a straightforward specification with a few explanatory variables , and another model may make extensive use of interaction terms (e.g. fare/income, female*trip purpose, etc.) to capture heterogeneity. We ask in the assignment that you present in your write up the best model specification . How do you determine "best"? This can be difficult, but generally there are three factors you are weighing:

1. Expectations: Does the model match my a-priori expectations of the behavior? Does it include the factors that I believe would be important in the decision ? Is heterogeneity (i.e., variability in preferences based on personal characteristics and context) adequately represented?

2. Statistics: Use of statistical tests (t-test and likelihood ratio-test) and goodness of fit. For this assignment, you will be using t-tests to provide information on whether an explanatory variable should be included in the model (i.e., is it statistically significant?). You will use likelihood ratio-tests to provide information on whether a parameter should be generic (the same across alternatives) or alternative-specific (different across alternatives) or whether a parameter should vary based on socio-economic characteristic or context. Note that absolute goodness of fit value is not important (depending on the circumstances, a fit of 0.1 can be a reasonable model), but rho-bar squared can be used to compare models estimated from the same dataset (as you are doing in this assignment) that cannot be otherwise compared via a t-test or likelihood ratio test (i.e., in cases where one model is not a simple restriction of the other).

3. Applications: Will the model serve the purpose of the intended application?

Note that these factors are often conflicting. For example, a variable important in application (e.g., travel time) may either have an incorrect sign (failed expectations) *or* have the correct sign but be statistically insignificant (failed statistics). What to do?

In an ideal world, you would be able to refine the specification to resolve any conflicting issues. When you can't, you need to make hard choices. An easy choice is to not arbitrarily draw the line at 95% confidence on the statistics (e.g., a t-stat of 2.0), so it is okay to keep variables in the final model that have t-stats below 2.0 (later in the course we'll give some theoretical justification for this). Unfortunately, exactly how low can you go is hard to say! On the other hand, wrong signs are fatal, for example it is not acceptable to have a positive parameter on the travel time variable. This would be telling airlines they can improve ridership by increasing flight time!

One other thing to keep in mind is that the absolute value of any parameter is not meaningful, for example the variable with the largest parameter does not mean it is the most important variable. To see this, note that if you change the units of the variable (e.g., from minutes to hours), this will simply scale the parameter the equivalent amount. What is useful to compare are ratios of parameters, which represent marginal rate of substitution. For example, in a linear model:

$$U_t = \cdots + \beta_t \times Time + \beta_c \times Cost + \cdots$$

$\frac{\beta_t}{\beta_c}$ is the marginal rate of substitution between cost and time and is an estimate of the value of time of the respondents. Looking at ratios such as these can be extremely useful in model development when we have a-priori expectations of the relationship (e.g., value of time should be on the order of the wage rate).

Appendices

There are three appendices to help you with the assignment, each of which you should read.

A. Brief Orientation to Pylogit
B. Theoretical reminders (on the statistical tests you are expected to perform)
C. Air Travel Survey Dataset

Report Content

Your report should be ***no more*** than 2 pages (no appendices) and include:

1. A presentation of your best model specification: This should be a cleaned up version of the Pylogit output that can be clearly understood by the reader (e.g., understandable description of the explanatory variables).

2. Discussion and analysis of your final model: Does the model match you're a- priori expectations? Why or why not? What heterogeneity have you captured or not captured? What issues did you come across? How did you use the statistical tests to arrive at your final model? Since learning the statistical tests is an important objective of this assignment, describe in general how you used these tools and also explicitly describe your use of ONE t-test and ONE likelihood ratio test, including the calculations.

This may be overly repetitive, but again, the specific questions to be answered are:

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?
2. What model specification corresponds to your a-priori hypotheses? Show the equation.
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 (dummy variable, categorical variable, etc). For example, a categorical variable, say educational level, should include the categories it entails in addition to the base category(ies).
4. Does your "best" model match your a-priori expectations?
5. Does your "best" model include the factors that you think are important in the decision?
6. What heterogeneity is captured in your model? What heterogeneity is not captured by your model? Is heterogeneity adequately represented in your model?
7. What issues did you come across while developing your "best" model?
8. How did you use t-tests, likelihood ratio tests, and goodness-of-fit comparisons to arrive at your final model?
9. Show how you used one t-test and how you used one likelihood ratio test to choose between different model specifications. Include your calculations.

For the table:

1. For each variable in the model, include the units and what alternative's utility equation the variable belongs to.
2. Use actual variable names (not the shorthand used in one's computer code). For example, use "travel time" instead of "tt".
3. Include the parameter estimates, t-stats, and p-values?
4. The report table should present everything that needs to be known about the model (clear variable names, choice indicator defined, dummy variables defined, clear base category(ies) for categorical variables, etc.). The table should convey all required information to understand and assess your "best" model.

Don't forget parts 2 and 3 to the problem set (outside of the 2 pages for the model report).

Part 2 (15 Points): Research Project

In Problem Set 1, we asked you to describe a specific research question or policy of interest that you wish to explore in greater detail using the quantitative tools that you learn through this course. Now, we want you to frame the research problem in terms of a binary choice, identifying explicitly both the dependent variable of interest, and all independent explanatory variables that you believe influence the choice. Specify the utility of the two alternatives as some function of the explanatory variables, accompanied by a very brief description of your a priori expectations regarding each of the model parameters. Don't feel constrained by whether an explanatory variable is observed or not, or if a particular utility specification is estimable. Be creative! Finally, show how you are addressing the research question using the model specification.

For example, in Problem Set 1 your research problem may have been, "What is the impact of real time bus arrival information (e.g. NextBus) on transit ridership?" One possible way (among many) to reframe the problem as a binary choice is to let the dependent variable be: does an individual choose transit or not. Possible explanatory variables that influence an individual's decision may include attributes of the transit alternative (availability of real-time information, some measure of reliability), characteristics of the decision-maker (level of patience), situational and contextual constraints (decision-maker is in a hurry, or is travelling with kids), etc. In picking a functional form for the utility of the two alternatives: choose transit or not, we employ a plain vanilla linear-in-parameters specification of the utility choosing transit:

$$U_{transit} = \cdots + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots$$

where the β's represent the model parameters and the x's denote the explanatory (exogenous) variables, and we set the utility of not choosing transit equal to zero. We encourage you to be more thoughtful in how you incorporate the explanatory variables.

Say we hypothesize that real-time arrival information influences transit ridership through its effect on the disutility of waiting time. The availability of real-time information may reduce the unpleasantness associated with not knowing when the next bus will arrive, but may also drive potential passengers away from transit in cases where they learn that the next bus is not due for a onerous, thereby increasing the probability that a decision-maker chooses transit. If the opposite is true, then the availability of real -time arrival information decreases the probability that a decision-maker chooses transit. If the hypothesis test is inconclusive, then it appears that real-time arrival information has no impact on transit ridership, at least through its effect on waiting times.

The questions to be answered in this part are:

1. As a binary choice problem, what would be your dependent variable?
2. What independent variables do you think influence the choice?
3. Specify the utilities of the two alternatives.
4. What are your a priori expectations of the coefficients for each specified variable?
5. How are you addressing the research question using the model specification?

Part 3 (15 Points): Supplemental Problem

In a linear city, there are two stores located at points C and D, as shown in Figure 1. There are 500 daily trips in total to both stores. Two hundred trips originate from the residential areas on the left of C, and 300 trips originate from the area to the right of D. (The area between C and D is not residential.)

The shoppers make one trip per day to one of the stores, either store 1 or store 2, but not both. Those who live on the left of point C are uniformly distributed between points A and B. The shoppers who live on the right of point D are exponentially distributed:

$$f_x(x) = \gamma e^{-\gamma x}, x > 0$$

where γ is an unknown constant (parameter) and x is the distance from point D (to the right). The utility to individual n from choosing store i, $where\ i \in \{1, 2\}$ is given by:

$$U_{in} = \beta_1(distance_{in}) + \beta_2 \ln(size_i) + \epsilon_{in}$$

, where $distance_{in}$ is the distance of individual n from store i, $size_i$ is the size of store i, and ϵ_{in} is an error term, which is independently and identically distributed Gumbel (Type I Extreme Value) across all alternatives and across the population.

Given distances d_{AB}, d_{BC} and d_{CD}, what is the expected number of trips taken each day to store 1 and store 2? How does the value of γ affect the number of trips to stores 1 and 2? Your answers should be neat and written in order such that one can read the answer from top to bottom without having to turn to multiple areas to see how one arrives at their answer. Should clearly circle or denote their answer. Should show all work.

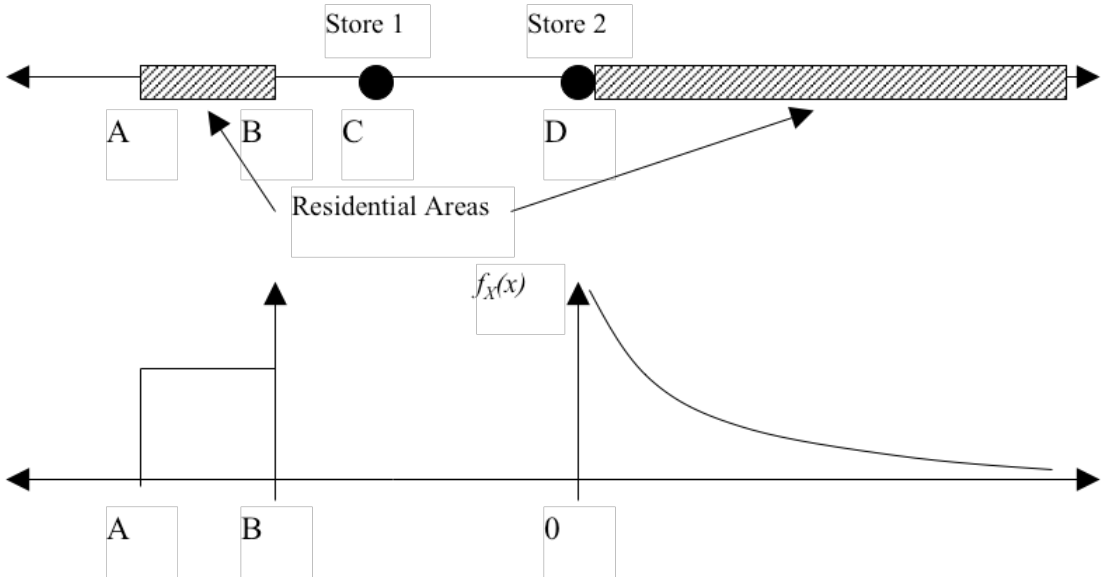


Figure 1: Population distribution for the linear city

Appendix A: Brief Orientation to Pylogit

Pylogit requires, as input, a model specification file and a data file, which are found on bcourses. We have provided a sample model specification file that you can use as a starting point. The file is heavily commented, to help you understand the syntax.

Pylogit requries the data file to be processed in long format, in order to optimize estimation time as much as possible. We are providing you with the syntax that converts a data file from wide format into long format. However, for the upcoming assignments, you will be responsible for the conversion from wide to long format. The following link (https://github.com/timothyb0912/pylogit/blob/master/examples/notebooks/Main%20PyLogit%20Example.ipynb) entails a thorough description of the data structure conversion process in addition to a walk-through example. This link is critical to help you understand and get familiar with the syntax and code structure.

Appendix B: Theoretical Reminders

The Binary Logit Model

Binary logit can be used to model discrete choice situations where the choice set involves only two alternatives. Assuming random utility maximization, the probability that individual n chooses alternative i from choice set $C_n = \{i, j\}$, can be expressed as:

$$\begin{aligned} P_n(i|C) &= P\left(U_{in} \geq U_{jn}, \forall j \in C_n\right) \\ &= P\left(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}\right) \\ &= P(\epsilon_{in} - \epsilon_{jn} \geq V_{jn} - V_{in}) \end{aligned}$$

, and assuming the errors are distributed i.i.d. Extreme Value leads to the binary logit formula:

$$P_n(i|C) = \frac{e^{\mu V_{in}}}{1 + e^{\mu V_{in}}}$$

, where V_{in} and V_{jn} represent the deterministic parts of the utilities (U_{in}, U_{jn}) for alternatives i and j, respectively. The choice of the i.i.d. Extreme Value distribution is motivated by its good analytical properties, allowing for a closed-form solution of the choice probability formula (i.e., the probability is trivial to calculate). Note that for identification reasons the scale parameter μ (which is related to the inverse of the variance of the distribution of the error terms) is typically normalized to 1 (see Ben-Akiva and Lerman (1985), page 71).

Asymptotic t-test

The asymptotic *t*-test can be used to test whether a particular parameter θ differs from zero or some other known value k. The null hypothesis is represented by:

$$H_0: \theta = k$$

We reject the null hypothesis if

$$\left|\frac{\hat{\theta} - k}{\hat{\sigma}}\right| > t_{cr}$$

, where $\hat{\theta}$ is the parameter estimate, $\hat{\sigma}$ is the standard error of $\hat{\theta}$, and $t_{cr} = 1.65$ at the 10% significance level and $t_{cr} = 1.96$ at the 5% significance level. The .html output file contains the estimated values of the parameters in the utility functions, with the associated standard errors and *t*-statistics. An asterisk (*) is appended if the *t*-test fails at 95% confidence (i.e. the null hypothesis that $\theta = 0$ cannot be rejected).

Goodness-of-Fit Measures

Likelihood ratio index ($\boldsymbol{\rho^2}$): The likelihood ratio index ρ^2 is an informal goodness-of-fit index that measures the fraction of an initial log-likelihood value explained by the model. Mathematically,

$$\rho^2 = 1 - \left|\frac{L(\hat{\beta})}{L(0)}\right|$$

, where $L(\hat{\beta})$ is the maximum log-likelihood of the model whose goodness-of-fit we wish to estimate, and $L(0)$ is the log-likelihood of a model with all parameters equal to zero. The likelihood ratio monotonically increases in the number of parameters K. For a binary choice model, it must lie between 0 and 1 (see Ben-Akiva and Lerman (1985), page 91).

Adjusted likelihood ratio index ($\boldsymbol{\bar{\rho}^2}$): Adjusted likelihood ratio index $\bar{\rho}^2$ is an informal goodness-of-fit index that is similar to ρ^2 but corrected for the number of parameters estimated.

$$\bar{\rho}^2 = 1 - \left|\frac{L(\hat{\beta}) - K}{L(0)}\right|$$

Thus, the adjusted likelihood ratio index penalizes models with a larger number of parameters.

Likelihood ratio test

The likelihood ratio test can be used to compare the goodness-of-fit between two models when any one of the models is a nested specification of the other model (or alternately , when one of the models is some kind of extension of the other model). The base model is called the restricted model , and the extended model the unrestricted model. The log-likelihood values of the two models are compared in the likelihood ratio test. Let L^U and L^R denote the values of the log-likelihood function at its maximum for the unrestricted and restricted models, respectively , and r be the number of independent restrictions imposed on the parameters in computing L^R. The corresponding test statistic is given by:

$$-2(L^R - L^U)$$

which is asymptotically distributed as χ^2 with r degrees of freedom. If $\chi^2 > \chi_c^2$ we reject the null hypothesis that the restrictions are true. In other words, we reject the restricted mode and select the unrestricted model (see Ben-Akiva and Lerman (1985), page 28).

For example, a likelihood ratio test can be used to test for generic attributes vs. alternative-specific attributes. A generic specification imposes restrictions of equality of coefficients on a more general model with alternative-specific attributes. The likelihood-ratio test statistic for the null hypothesis of generic attributes is given by:

$$-2\left(L(\hat{\beta}_G) - L(\hat{\beta}_{AS})\right)$$

where G and AS denoted the generic and alternative-specific models, respectively. It is χ^2 distributed with the number of degrees of freedom equal to the number of restrictions $(K_{AS} - K_G)$ see Ben-Akiva and Lerman (1985), page 168).

References

Ben-Akiva, M., and Lerman, S. (1985), "**Discrete choice analysis: Theory and application to travel demand**," Vol. 9, The MIT press.

Appendix C: The Air Travel Survey Dataset

Context

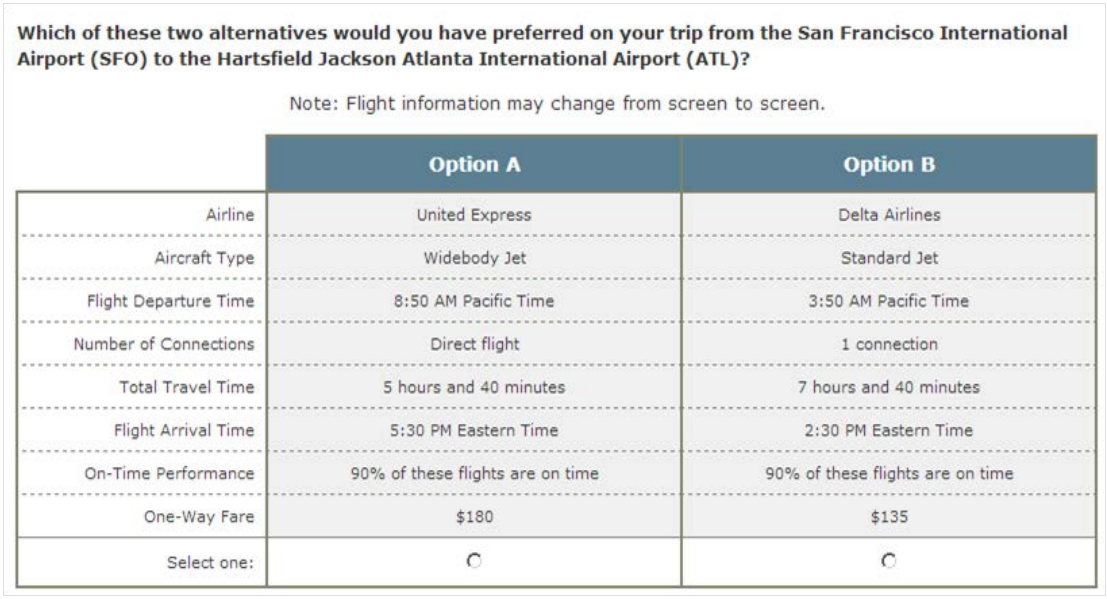
With the recent trend of on-line travel sites, consumers have become increasingly more aware of available air travel options. As a result, consumers are able to search for and book a ticket to suit their own preferences, rather than rely on a travel agent. Understanding the behavior of air passengers has recently been a major focus for travel demand research for airlines and researchers alike. Actual ticket purchase history of passengers is difficult to come by, so surveys have primarily been the dominant means of acquiring data. This survey seeks to determine the extent to which basic trip and socioeconomic characteristics, in particular Frequent Flyer Program (FFP) membership, affect the process of itinerary choice for air travel passengers.

FFPs were created following the de-regulation of the airline industry in 1978. American Airlines created the world's first airline loyalty program in 1981 and shortly thereafter other airlines followed suit. Initially the mechanism behind FFPs was the accrual of tokens, commonly referred to as "miles", from flying with a particular airline. When a certain level of miles was reached, one could redeem them for a free flight of her choice. The purpose of these programs was to create customer loyalty towards a particular airline. As individuals accumulate miles with one airline they face a high opportunity cost for flying with other carriers. Eventually the programs were expanded to include tiers of "status" for individuals that fly certain amounts each year. With these status levels came other benefits: first class upgrades, mileage multipliers, priority boarding, and free or reduced fees ranging from checked bags to last minute flight changes. Quantifying the loyalty effect created by FFP membership is the goal of this study.

Data

The data represents a sample of 878 individuals that were asked about their recent travel behavior. The dataset was collected through an online survey conducted in 2010. Background information was collected for each individual, such as income, FFP membership, and age. FFP membership information was collected with regards to seven airlines. The airlines and airline codes used in the data file are as follows: American (AA), Continental (CO), Delta (DL), JetBlue (B6), Southwest (WN), United (UA) and US Airways (US). Each person was asked about a recent trip they have taken. Information from this trip was recorded, such as trip purpose, class of service, preferred arrival time, who paid for the ticket, and operational characteristics like departure time, connections, and airports. Eight pairs of hypothetical itineraries were generated based on the trip provided by the respondent and they were then asked to select between pairs of itineraries presented to them, as if they were flying the same trip. We can assume that properties of the original trip, like trip purpose, preferred arrival time, and class of service, remain constant for the hypothetical trips. The data file is in wide format and will be converted to long format via the sample code that we have provided.

A screenshot of the hypothetical trips that were presented to the individuals is shown below (see Figure 2). This trip was based on a $350 round trip, economy ticket from SFO to ATL. The individual characteristics are described in Table 1. The alternative-specific attributes for each option are described in Table 2.



Which of these two alternatives would you have preferred on your trip from the San Francisco International Airport (SFO) to the Hartsfield Jackson Atlanta International Airport (ATL)?

Note: Flight information may change from screen to screen.

	Option A	Option B
Airline	United Express	Delta Airlines
Aircraft Type	Widebody Jet	Standard Jet
Flight Departure Time	8:50 AM Pacific Time	3:50 AM Pacific Time
Number of Connections	Direct flight	1 connection
Total Travel Time	5 hours and 40 minutes	7 hours and 40 minutes
Flight Arrival Time	5:30 PM Eastern Time	2:30 PM Eastern Time
On-Time Performance	90% of these flights are on time	90% of these flights are on time
One-Way Fare	$180	$135
Select one:	○	○

Figure 2: Example screenshot from on-line survey

personID	Unique number for each person in the survey
choiceSituationID	Unique number for each choice situation in the survey
choice	Stated choice indicator: 1= alt. 1 , 2 = alt. 2
gender	Gender: 1 = male, 2 = female
age	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
purpose	Trip purpose, categorical variable: 1 = Business, 2 = Attend conference, 3 = Vacation, 4 = Visit friends or relatives, 5 = Attend school/college, 6 = Other
income	Annual income, categorical variable: 1 = Under $10,000, 2 = $10,000 - $19,999, 3 = $20,000 - $29,999, 4 = $30,000 - $39,999, 5 = $40,000 - $49,999, 6 = $50,000 - $74,999, 6 = $75,000 - $99,999, 7 = $100,000 - $149,999, 8 = $150,000 - $199,999, 9 = $200,000 - $249,999, 10 = $250,000 or more
classTicket	Class of ticket, categorical: 1 = Economy or coach, 2 = Premium economy, 3 = Business, 4 = First Class
payment	Who paid for the ticket, 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
(airline)_FFP	FFP membership for each (airline), categorical: 1 = Not a member, 2 = Basic member, 3 = Elite member

Table 1: Individual Characteristics

aXaircraft	Aircraft type for alternative X, categorical: 1 = Widebody (200+ passenger capacity with 2 aisles in coach), 2 = Standard Jet (100-200 passenger capacity), 3 = Regional Jet (50-100 passenger capacity), 4 = Propeller
aXdepartMAM	Departure time for alt. X, minutes after midnight (MAM)
aXconnections	Number of connections for alt. X (0, 1 or 2)
aXtravtime	Total travel time for alt. X (minutes)
aXarriveMAM	Arrival time for alt. X, minutes after midnight (MAM)
aXtimediff	Difference between arrival time and preferred arrival time, for alt. X (min) (e.g. +60 means 1 hour later than preferred)
aXperformance	On-time performance for alt. X (%)
aXfare	Air fare for alt. X ($)
aXairline	Airline for alternative X, categorical: 1 = American (AA), 2 = Continental (CO), 3 = Delta (DL), 4 = JetBlue (B6), 5 = Southwest (WN), 6 = United (UA), 7 = US Airways (US), 8 = Other

Table 2: Alternative-Specific Attributes