CE 264 – Spring 2017
Problem Set 4: Forecasting

Due: Tuesday March 6th, 2017

In part 1 of this assignment, you will use sample enumeration techniques to forecast changes in aggregate demand in response to changes in one or more inputs to the choice model. You will not be estimating a model of your own: the model file (MNL-Forecasting.ipynb) will be used for both estimation and forecasting. The file (MNL-Forecasting.ipynb) entails the full specification you need for estimation. Make sure you understand what each parameter denotes, and how the file works. The python script also performs sample enumeration to predict market shares. More details on sample enumeration can be found in the Appendix. Part 2 asks about experiment design for your research question. Part 3 is a supplemental problem on how to select a "best" model given estimation results for multiple model specifications.

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 (60 Points): Forecasting

Objectives:

The objective here is to demonstrate how the many different model forms taught throughout the course may be applied in practice to inform important policy questions.

Dataset: The Bay Area Travel Survey (BATS) 2000

The Bay Area Travel Survey is a large-scale regional household travel survey conducted in the nine county San Francisco Bay Area of California. The Metropolitan Transportation Commission (MTC) has periodically sponsored BATS to provide data to support travel modeling and analysis of regional travel behavior. The target data collection period for BATS 2000 was of course the 2000 calendar year. The survey consisted of an activity-based travel diary that requested information on all in-home and out-of-home activities over a two-day period, including weekday and weekend pursuits. In all, more than 15,000 households participated in the survey.

For the purpose of this problem set, we shall be examining travel mode choice for morning commute by a sample of 10,000 Bay Area residents. The choice set for any individual consists of a maximum of six travel alternatives: drive-alone, shared ride, walk, bike, walk to transit and drive to transit. The variables in the dataset are enumerated in Table 1. The model specification provided is a modified version of the work-tour mode choice model component of the San Francisco County Transportation Authority's (SFCTA) travel demand forecasting model.

The SFCTA holds proprietary rights to the dataset, and the dataset may not be used for term projects or purposes outside of the class (research or otherwise) without explicit written permission from the instructor. There are restrictions on this dataset, and its use outside of the case studies may require consent from the appropriate sources.

obsID	Observation ID
choice	Chosen alternatives 1 = drive alone (da), 2 = shared ride (sr), 3 = walk, 4 = bike, 5 = walk to transit (wt), 6 = drive to transit (dt)
originTAZ	Travel Analysis Zone in which the trip origin lies
destTAZ	Travel Analysis Zone in which the trip destination lies
avail_X	Boolean variable indicating the availability of alternative X
tt_X	Travel time of alternative X (minutes)
dist_car	Travel distance of alternatives drive alone and shared ride (miles)
cost_X	Travel costs of alternative X ($)
accTime_X	Access time of alternative X (minutes)
accDist_dt	Access distance by car of alternative drive to transit (miles)
egrTime_X	Egress time of alternative X (minutes)
iWait_X	Initial wait at first transit stop for transit alternative X (minutes)
xWait_X	Transfer wait for transit alternative X (minutes)
weights	Population weights

Table 1: Description of variables in the dataset

Appendices

There is one appendix on theoretical reminders to help you with the assignment.

Tasks

1. Estimate the model specification file provided with the assignment (Note: you do ***not*** need to create your own specification).
   a) Briefly describe the specification in words.
   b) Do the parameter estimates make sense?
   c) What do you think of the specification, in particular the estimated parameters associated with travel costs?
   d) Is there anything that you would have done differently if specifying the model yourself?
2. We shall use estimation results from the model specification provided with the problem set to generate some forecasts using sample enumeration techniques. The SFCTA is proposing a congestion toll for downtown San Francisco. The toll will be applied in an area bounded by Divisadero and Castro streets in the west, 18th street in the south and the San Francisco Bay in the north and east. The toll will apply to all motorized private vehicles inbound to this area during the morning. The toll rate has not yet been fixed, and the SFCTA is interested in seeing how commuters will react to different amounts.

   The dataset contains information about the Travel Analysis Zones (TAZs) of the origin and destination for each trip. TAZs ***1- 42*** lie in the northeast cordon that SFCTA is proposing to toll. Make assumptions as you deem fit, but be sure to mention and justify them in your report. Remember that you do not need to estimate a model for generating forecasts. Since the effect of

the toll is expected to be on trips that have some component within the cordon, restrict your analysis to trips that have either the origin or the destination or both within the cordon (though the toll only applies to inbound trips).

a) Use the "MNL-Forecasting.ipynb" file and the estimation results for the model specification provided to generate ***both*** a table and a plot of mode shares for all six modal alternatives as the congestion toll is varied from $0 - $10, in increments of $1. Note that the "MNL-Forecasting.ipynb" file will have to be changed.

b) List and justify any assumptions made in performing the forecasts from Task 2a).

c) Comment on the results of your mode share forecasts in response to the varying congestion tolls.

3. Greenhouse gas (GHG) emissions have become a very important decision variable of late. One of the consequences of the congestion toll will be a shift from driving to other more sustainable modes such as public transit and biking. To get a rough estimate of the impact, assume that a car on average emits 0.916 lbs. of CO2 per mile traveled.

   a) Generate ***both*** a table and a plot that shows the reduction in CO2 emissions as a function of the congestion charge. Use the trip distances provided in the dataset to come up with an estimate.

   b) Clearly explain your calculations from task 3a) and show a sample calculation of how you determined the reduction in CO2 emissions for a given individual.

4. Answer the following questions:

   a) Should the travel times change as a result of the congestion charge? Why or why not?

   b) Based on your answer to question 4a), are your mode share and CO2 emission forecasts underestimates or overestimates?

Report Content

Submit a copy of your report in class. Your report should include:

1. The estimation results from the model provided with the assignment and the answers to the questions asked in Tasks 1 parts (a) – (d).
2. The table and plot of mode shares for each mode as a result of the varying congestion tolls. The answers to the questions in Task 2 parts (a) – (c) should also be included.
3. The table and plot of the changes in CO2 emissions as a result of the congestion charge. The explanation and sample calculation from Task 3b) should also be included.
4. The questions from Tasks 4a) and 4b).
5. A copy of "MNL-Forecasting.ipynb" once you've made appropriate changes to the syntax to account for the effect of the congestion toll.

Please keep the report as short as possible while meeting all of the requirements above. Don't forget parts 2 and 3 of the problem set!

Part 2 (20 Points): Research Project

By now, you should have an idea for a research project that you're dedicated to pursuing through the remainder of the course. The objective of this assignment is to get you thinking about data collection.

1. Identify some of the key attributes and characteristics (3-4 each) that you think are central to the decision-making process.

2. Discuss the relative merits and demerits of collecting Stated Preferences (SP) data versus collecting Revealed Preferences (RP) data within the context of your group project.

3. Describe how you would collect the data in an RP setting. In order to estimate a model, remember that you require the attributes of not only the chosen alternative but non-chosen alternatives as well.

4. Construct a sample question for an SP survey. Outline the hypothetical situation for the benefit of the survey respondent, enumerate example values for the relevant attributes of each of the alternatives, and write out the questions asking for the pertinent characteristics of the decision-maker.

You should brainstorm with other members from your group, but each member is required to submit an independent copy.

Part 3 A) (10 Points): Supplemental Problem

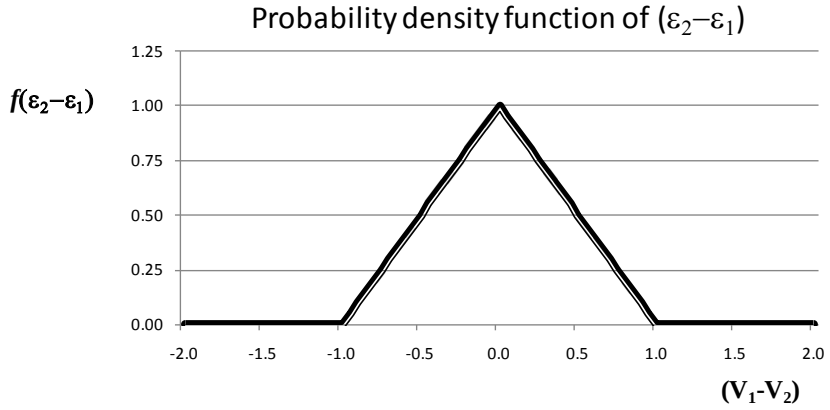
You are working on a model of air travel itinerary choices. You are exploring different specifications of fare. Several alternative model specifications are shown below. To keep things simple, we've shown only the portion of the model results focusing on the attributes of fare and the total trip time and the decision-maker characteristic of the purpose of the trip (business versus leisure). Based on these estimation results, **which model would you select as the best model and why?** In other words, explain why your chosen model is preferred to each of the other three model specifications.

	Model 1 MNL		Model 2 MNL		Model 3 MNL		Model 4 MNL	
Parameter	est	t-stat	est	t-stat	est	t-stat	est	t-stat
Fare ($) specification								
Fare	0.08	(3.2)	-0.01	(2.1)	-0.09	(1.7)		
Fare – business travelers							-0.08	(1.5)
Fare – leisure travelers							-0.13	(2.1)
Total trip time (hours)	-4.70	(4.0)	-6.50	(3.5)	-5.05	(2.8)	-5.10	(3.1)
Log-likelihood		-210.1		-213.7		-220.0		-219.9
Rho-squared		0.302		0.290		0.269		0.270

Table 2: Estimation results

Part 3 B) (10 Points): Supplemental Problem

Say you are working with a random utility binary choice model (utility maximizing) where the assumed distribution of the error terms is as shown below. Given this model, if the systematic utility of alternative 2 is greater than the systematic utility of alternative 1 by a value of 0.5, what does this model say is the probability of choosing alternative 1?



Appendix A: Theoretical Reminder

Recall that the random utility maximization model is based on the assumption that a decision-maker n $(n = 1, \dots, N)$, faced with a finite set C_n of mutually exclusive alternatives i $(i = 1, \dots, I_n)$, chooses the alternative j which provides the greatest utility U_{nj}. The utility of each alternative U_{ni} is described as some function of explanatory variables that comprises the systematic part of the utility function, V_{ni}, and some stochastic component, represented by the disturbances ε_{ni}:

$$U_{ni} = V_{ni} + \varepsilon_{ni} = \mathbf{x_{ni}}\boldsymbol{\beta} + \varepsilon_{ni}$$

, where $\mathbf{x_{ni}}$ is a $(1\times K)$ vector of attributes of alternative i for decision-maker n, and $\boldsymbol{\beta}$ is a $(K\times 1)$ vector of parameters to be estimated. Assuming that ε_{ni} is distributed Gumbel i.i.d. across alternatives and decision-makers in the sample population yields the multinomial logit expression that you are all familiar with by now:

$$P_n(i|\mathbf{x_n};\boldsymbol{\beta}) = \frac{\exp(\mathbf{x_{ni}}\boldsymbol{\beta})}{\sum_{j\epsilon C_n}\exp(\mathbf{x_{nj}}\boldsymbol{\beta})}$$

, where $P_n(i|\mathbf{x_n};\boldsymbol{\beta})$ is the probability that decision-maker n chose alternative i. Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimate for $\boldsymbol{\beta}$ that you obtain for the above MNL model. You will be using these estimates for both parts to this problem set.

Sample Enumeration

Once we have the parameter estimates $\widehat{\boldsymbol{\beta}}$ for a model, we can use these to calculate aggregate demands for different alternatives as follows:

$$D(i) = \sum_{n=1}^{N} w_n P_n(i|\mathbf{x_n};\widehat{\boldsymbol{\beta}})$$

, where $D(i)$ denotes the aggregate demand for alternative i, and w_n is the weight for individual n when dealing with a stratified sample. To get a share, divide the demand by the sum of the individual weights. Different forecasts are generated by varying the vector of observable attributes $\mathbf{x_n}$, the parameter estimates $\widehat{\boldsymbol{\beta}}$ stay unchanged.

You could use a spreadsheet (or some other database software) to do the forecasts. However, we'd like you to use the python script (MNL-Forecasting.ipynb) provided with the homework to generate the probabilities.

Upon execution, the code enumerates for each observation in the dataset the requested

choice probabilities for the various alternatives. It also caters for the individual weights in calculating market shares.You will need to make changes to the script to, among other things, account for the effect of the congestion toll.