Problem Set 2
Empirical Demand Estimation in Insurance Markets
Due Date: Friday March 22, 2019

This problem set is designed to give you experience programming the kind of empirical demand models discussed in the unit on insurance markets in class. The material here is closely related to the paper 'Adverse Selection and Inertia in Health Insurance Markets' by Ben Handel (2013), who also generated the simulated data and questions for this problem set. Feel feel to contact him at handel@berkeley.edu with questions or comments.

As stated in the syllabus, you may work in up to groups of three on this (I suggest you do find others to work with as this is a substantial exercise and discussion will help). This is mainly for you to have a commitment device to do the work yourself, since it is in your own best interest to do so. If you have to talk to another group for advice, it's no big deal.

Note that this problem set is more like a set of instructions to implement and understand a model similar to that in Handel (2013). As such, the main output to hand in is the code, along with answers to the questions asked when there is output that is asked for apart from the code.

To answer the questions here you will need a simulated data set that can be found on the course website 'ASIN-ChoiceModelData-Final.mat.' These data contain (i) a population of families over a three year panel (ii) the results of cost model estimation that describes their ex ante out-of-pocket cost risk at the time of choice for each family and each of three plans and (iii) choices and demographics over the relevant three year time horizon. The simulated data are described in much more detail in the corresponding document 'Data-Description-Handel-ASIN.doc', see that document before starting the problem set.

The problem set begins with a series of questions designed to get you familiar with the data environment before getting to the primary model estimation. All work for this problem set should be done in Matlab or a similar environment.

1. Our first question asks you to explore the data and set up the data for choice model estimation. First, let's generate some summary statistics:

(a) For each family calculate the mean and standard deviation of their out-of-pocket expenditures for each of the three health plans in year 1.
   - What is the population mean of out-of-pocket expenditures in each plan?
   - What is the population mean of the standard deviation of out-of-pocket expenditures in each plan?
   - Provide the statistics in the two bullets above broken down by family status in year 1, given in the variable 'Tier1'.

(b) For year 1 provide tabulations for the following variables across the population of families:
   - Indicator of chronic conditions in family
   - Enrollment in Flexible Spending Account
   - Employee Manager Status
   - Family Status
   - Income
   - Ages (do this for every individual in the population, use the matrix 'Ages')[1]
   - Genders (do this for every individual in the population, use the matrix 'Genders')

(c) Now, let's transform / add to the data to get it ready for choice model estimation. To do this start a code file called 'Estimation-Code.m' in Matlab and start by loading the data.
   - Generate a 6 X Sim matrix of standard normal random draws. These will be used to index the family-plan-time specific epsilons for two plans over three years in estimation (the draws multiplied by the candidate standard deviations will give a normal distribution with that standard deviation.)
   - Generate two additional 1 X Sim matrices with standard normal draws. These will index the risk preference and PPO1200 normally distributed random coefficients.

[1]NOTE: The matrix Ages, Genders, and Famsize DO NOT match up correctly with the family status variables Tier2 and IND due to the way the simulated data were created. These matrices do reflect the number of people used to simulate each family, but the IND and Tier2 variables were created orthogonally as 'preference' variables only. In what follows, you should use the Ages and Genders matrices only where specifically referenced (in risk preferences estimation) and use Tier2 and IND to distinguish between family and single since these variables were used to simulate 'single' vs. 'family' preference effects in the model.

- The out-of-pocket variables in the simulated data are in nIs X K matrices. Create (nIs, K, Sim) three dimensional matrices that extend the data by adding a third dimension with Sim columns (this repeats the same two-dimensional matrix for each family Sim times). This will be useful to speed up your calculations.
- Take the nIs X 1 HTCi variable and put it into an (nIs, K, Sim) matrix (you repeat the HTCi matrix many times)
- Set a 'permanent' income variable equal to the income in year 2 of the data (nIs,1).
- Set a family 'age' variable equal to the maximum age in each family (nIs, 1).

(d) There are two final steps to prepare the data for estimation. These steps compute quantities necessary for choice model estimation, but that are 'fixed' throughout estimation. This means it is better to compute them up front and use them as inputs into estimation.

- Define a wealth variable equal to 75,000 (just a scalar).
- Generate nine matrices (nIs,K,Sim) that describe the basic 'money at stake' for each person and each potential health state realization and each year. Here, a person-state is on the nIs,K dimensions while we repeat this Sim times for ease in estimation. The nine matrices are for each of the three plans over each of the three years. Each entry (i,k,s) in year t for plan p should be (Wealth - OOP Expense draw (i,k) (for plan p in year t) - Price (i) (for plan p year t)). NOTE: prices depend on family status and are entered accordingly into price matrices already. Label these matrices EUvector'p''t' where p is for plan and 't' is for time.
- Define six (nIs, Sim) matrices that define the actual default plans for families in non-active choice years. Label each matrix choice'p''t'. Here, year 1 is an active choice year, so t should only be 2 or 3. Use the observed choices to fill in these matrices with binary values: if a family chose plan 'p' in year 't-1' then choice'p''t' = 1. This leads to (nIs,1) matrices that you should repeat Sim times on second dimension to use in estimation.

2. Now we will move to the actual estimation. You will be estimating

the primary choice model in Handel (2013) referenced above. You will need to reference section three of the paper in order to understand the exercise. The first part of that section describes the choice model while the Estimation subsection describes specific assumptions made in estimation (specifically, distributional assumptions for unobserved heterogeneity and functions linking observed demographics to risk preferences, inertia, PPO1200 preferences, and epsilons). The discussion of the model and estimation in the paper helps to provide context for the detailed implementation that this problem set walks you through. **Note:** the cost model described in that section is already estimated in your data and will be used as an input into choice model estimation. You will also want to reference Online Appendix B on the author's website with details on the simulated maximum likelihood estimation algorithm.

The remainder of the problem set has the following major components:

- Set up the likelihood function that describes the likelihood that families make certain sequence of choices, subject to candidate model parameters
- Set up and run a non-linear optimization routine that searches over the parameter space to find the model parameters that maximize the value of this likelihood function. These are the final estimates.

First we'll set up the likelihood function then the non-linear optimization that takes that function as an input. For candidate parameters considered in non-linear optimization, this likelihood function simulates choices as if those candidate parameters are the true parameters, then matches the predicted choices to actual choices made. This function is called in the optimizer you will build in the next question and when it returns the best value, those are the parameters that simulate choices that best match the observed choices. Before you complete questions 2 and 3 setting up the likelihood function, you may want to take a quick read through question 4 to see how the likelihood function will be used. For example, section 4 defines an initial parameter value 'guess' alpha0 that will be used to start the non-linear optimization routine. The likelihood function instructions below describe what each parameter to be estimated is, and which entry in alpha that parameter corresponds to.

To start, generate a file Generate a file called 'Likelihood.m' and perform the following instructions:

(a) Use the first line to define the function in Matlab. This line should read:
function output = Likelihood(alpha,.....)

Here, alpha will be your matrix of parameters that you will estimate, and after alpha you should include all variables you will need to bring as fixed factors into the estimation, separated by commas. This includes the matrices generated above in 'EsimationCode' and most of the demographic variables in the main data.

(b) The first step is to take the candidate parameters in alpha (which will be optimized over) and create matrices describing risk preferences, PPO1200 coefficient, and epsilons as functions of those parameters. These will be used later in the file to simulate choices conditional on those parameters, which will then be matched to the data. I'll help you get started: define alpha(1,1)/100 as the intercept of the mean of the normal distribution of risk preferences, alpha(1,2)/100 as the amount this mean shifts by if income tier increases by 1, alpha(1,3)/100 is the amount the mean shifts by if family max. age increases by 1, while alpha(1,4)/100 is the standard deviation of the risk preference random coefficient distribution around the mean formed with these prior three items. [2]

Use one of the 1 X Sim matrices generated in 1c and multiply by alpha(1,4)/100. Fill in an (nIs,Sim) matrix with this same row for each family. This is the standard deviation of risk preference random coefficients for candidate parameter alpha(1,4)/100. Next, generate an (nIs,1) matrix of family specific means where each entry is the mean of the risk preference random coefficient distribution for each family. This should equal alpha(1,1)/100 + alpha(1,2)/100*Income + alpha(1,3)/100*MaxAge. Transform into an (nIs,Sim) matrix which repeats this columns vector Sim

[2]Note: here we are dividing the parameters by 100 so that the numbers being optimized over by the non-linear optimizers are 'bigger' than the actual CARA coefficients. This helps performance / finding the optimum of the likelihood function since the CARA numbers un-normalized are quite small. Of course, in the results, we still care about the true CARA coefficients, which are alpha(1,x)/100.

times. Now, add the 'means' matrix to the 'standard deviation' matrix: you've just simulated a distribution of risk preference for each family. Take this (nIs,Sim) matrix and transform it into an (nIs,K,Sim) matrix by repeating the (nIs,Sim) matrix K times. The reason for this will make sense later when you simulate expected utilities: it helps speed up those calculations. Finally, make sure to set this matrix equal to the maximum of itself and 0.000001, truncating risk preference coefficients above 0 as the theory specifics.

(c) I'm going to be less helpful for the next few steps. First, generate the random coefficient distributions for PPO1200 using the other vector of 1 X Sim draws in 1c and parameters alpha(2,1), alpha(2,2) (mean and standard deviation for singles) and alpha(2,3), alpha(2,4) (mean and standard deviation for families). Create a (nIs,K,Sim) matrix similar to what you did for risk preferences for these draws. Note: YOU have to make sure to fill in this matrix appropriately as a function of family status, use the IND indicator variable to determine 'single' or 'family' (or Tier2 variable) NOT the size of Ages and Genders, which doesn't map correctly.

(d) Create six (nIs,K,Sim) matrices for the mean 0 normal epsilon draws for PPO500 and PPO1200 over the three years. Let alpha(3,1) and(3,2) be the standard deviations for individuals and alpha(3,3) and alpha(3,4) for families. Use the 6 X Sim matrix generated in 1c and procedure similar to that for the PPO1200 random coefficients to generate the final six matrices. Each matrix should correspond to either PPO1200 or PPO500, and to either year 1,2,or 3 (since these draws vary for an individual over time).

(e) Create a matrix that describes preferences for individuals with high total costs, as described by the matrix you transformed in 1c. This preference is constant for all people with HTCi = 1: use parameter (4,1) and create an (nIs,K,Sim) matrix = 0 from people with HTCi=0 and = alpha(4,1) with HTCi=1. This will be added to PPO500 and PPO1250 money at stake up front later (so reflects the disutility of high-cost people for those plans).

(f) Finally, before simulating choices, you need to fill in inertia for each family. Use alpha(4,2) and alpha(4,3) to be individual and family inertia intercepts. Then, using observables demographics, use alpha(4,4), alpha(5,1), alpha(5,2), alpha(5,3), alpha(5,4), and

alpha(6,1) to index inertia as a function of observed demographics:

- FSA enrollment (1 or 0)
- Income (1-5)
- Quantitative Sophistication Indicator (1 or 0)
- Manager Indicator (1 or 0)
- Chronic Conditions Indicator (1 or 0)
- Large Expenditure Change in Prior Year (1 or 0)

Generate inertia preference matrices for both year two and year three (some of these indicators changes over time!!) that are (nIs,Sim) using the above variables and parameters (use the linear format described in the Estimation section in the paper). Then, you will create 6 (nIs,K,Sim) matrices describing inertia preferences for each plan in years 2 and 3 (the non-active choice years). Use the choice'p''t' matrices created in 1d (inputs into the likelihood function!) to multiply these matrices such that inertia benefits the incumbent plan but has no benefit for plans which are alternative options in the observed data. Thus, only one plan for each family can get a benefit from incumbency in each year, and you should take that into account in these matrices, which will be added to utility.

3. Now, we will generate choice predictions conditional on candidate parameters within the likelihood function, and match them to observed choices.

   (a) Create state by state (of health draws) v-NM utility values. To do this compute $u_k(x)$ as specified in the paper:

$$u_k(x) = -\frac{1}{\gamma_k(X_k^A)} e^{-\gamma(X_k^A)x}$$

$$x = W_k - P_{kjt} - OOP + \eta(X_{kt}^B, Y_k)\mathbf{1}_{kj,t-1} + \delta_k(Y_k)\mathbf{1}_{1200} + \alpha H_{k,t-1}\mathbf{1}_{250} + \epsilon_{kjt}(Y_k)$$

   See page 15 of the paper for an explanation of notation and the estimation section for further implementation details.[3]

[3]Note: In this problem set, α should enter as a coefficient on PPO500 and 1250 (same for both plans) rather than on PPO250, since this it how it enters in the sample code.

You should use your EUvector'p''t' matrices from 1d, which are inputs into the likelihood function, here. Those (nIs,K,Sim) matrices describe W-OOP-P for each state / health realization and health plan, which is part of x above that is fixed regardless of choice model parameters. You should use these as inputs to computing the (nIs,K,Sim) matrices of v-NM utility values above. Try to do this without loops, which slow down programs substantially!

(b) Take the matrices produced in the last subsection, and create three expected utility matrices (one for each *year*) with dimensions (nIs,nPlans,Sim). These matrices should give the expected utility over the K simulated health draws for each person, plan, and simulation of random coefficients. Once computed, normalize these matrices by dividing by all entries by the expected utility for PPO1200 in each year *and then taking the reciprocals of these normalized values.* Taking the reciprocals is necessary because the CARA utility function is negative, so when you normalize, making all values positive, you have to take the reciprocal. The result is that expected utilities for PPO1200 will be 1 and those of the other plans will be defined in relation to that.

(c) Now the 'natural' thing to do would be, for each simulated draw of random coefficients and epsilon, to pick the sequence of plans which maximizes expected utility for years one to three for each family, and match those to the actual choices made. If you try to do this, the non-linear optimization won't work because the function is not continuous when binary choices are made as a function of the underlying parameters.

Create a 'smoothed' matrix of expected utilities that first raises the expected utilities you just calculated to some even power (say 6) and second divides these transformed utilities for each family and each simulation in each year by the sum of their transformed utilities. Thus, for each year, simulated draw of preferences, and family, the numbers in your matrix should summed over the second dimension (plans) should add up to one. This 'smoothed' matrix creates a continuous probabilistic function where the choice of the plan yielding highest expected utility approaches one and is increasing as the simulated utility gap between that and the other

plans becomes larger. Your resulting matrices can thus be seen as a discrete probability distribution over all three choices for each simulated preference draw, person, and year. See Estimation Appendix B for more details. *Note: The description here is slightly different than that in Appendix B, but the resulting implementation is the same. The instructions here implement the algorithm in this Appendix after some additional formula simplifications.*

(d) Finally, it's time to compute the log-likelihood function value, to do this you have to generate the probability for a given sequence of three *observed* choices given the simulated choices for the candidate parameters. It is crucial to compute probabilities over the entire sequence of choices since inertia links choices over time for a given preference parameter simulation. This is the part where you match simulated choices to actual observed data. You should:

- For each person, year, and simulated preference draw, multiply the probability of having made the *observed* year 1 choice by the *observed* year 2 choice by the *observed* year 3 choice. NOTE: In the limit, as the smoothing factor becomes large, this equals 1 for the sequence that contingently maximizes expected utility through the years, and 0 otherwise.
- Find the average of the probabilities of observing the actual sequence of choices across all simulated draws (so along the Sim dimension). This should leave you with a (nIs,1) matrix that describes the probability of a given family making a given sequence of choices over time if the candidate parameters, including those governing the distributions of random coefficients, are the true parameters.
- Create the final log-likelihood function value equal to the sum of the log of these family choice sequence probabilities across all families. This is the log-likelihood function value. Make the output of the function the negative of this value, because non-linear optimization will search for the minimum of the likelihood function rather than the maximum (this convention is standard).

4. Now, you have your likelihood function all set up, which is the hard part. To estimate the model parameters, all you have to do is run a non-linear optimizer to find the parameter values that yield that

highest likelihood function value. This question takes you through that process:

(a) You will do constrained minimization with either fmincon or ktrlink from the KNITRO optimization package. To do this you need to define the starting values and bounds for parameter matrix alpha which goes into the likelihood function. Define your bounds as follows (the bounds for the parameters alpha must be the same dimension as alpha):

$lb = [0, -.5, -.5, 0; -Inf, 0.1, -Inf, 0.1; 1, 1, 1, 1; -Inf, 0, 0, -5000; -5000, -5000, -5000, -5000; -5000, 0, 0, 0;]$;

$ub = [2, 1, 1, 1; Inf, Inf, Inf, Inf; 20000, 20000, 20000, 20000; Inf, 20000, 20000, 5000; 5000, 5000, 5000, 5000; 5000, 0, 0, 0;]$;

Also, you have to define starting values for the parameter matrix alpha. Use the following matrix to begin with for this:

$alpha0 = [0.06, 0.003, 0, 0.04; -2500, 700, -2200, 800; 300, 800, 300, 800; -500, 1250, 1750, -500; 0, 0, 0, 0; 0, 0, 0, 0;]$;[4]

(b) Before beginning non-linear optimization compute your likelihood function value at alpha0. **What value do you get?** If your value is in the neighborhood of 2000-3000, you are in good shape (the actual value is the negative of this, but the output will be positive since you have defined your likelihood output as the negative of the true output). *Make sure when evaluate your likelihood function that you pass in not only alpha0, but also all fixed variables that go into that function, including demographics, the normal draws and objects from 1c, and the matrices from 1d which help speed things up.* **Interpret this value of the likelihood function, what does it mean?**

(c) Set up your non-linear optimization routine with fmincon or ktrlink. This should take in starting value alpha0, but optimize over

[4]Note: Remember that alpha(1,1) to alpha(1,4) are divided by 100 in the likelihood function. This means that the starting values and output are 100 times bigger than the actual coefficients. You should report the results as the actual CARA coefficients, so alpha(1,x)/100. This normalization is done to avoid optimization errors with small numbers for some non-linear optimizers.

alpha starting from that point to find the alpha that minimizes the negative likelihood function (maximizes the true function) subject to the bounds on alpha passed in. **If you have trouble with this part, after struggling for a while you can email me for some hints.** You will want to setup some options to be passed into the optimization such as:

$options = optimset('MaxFunEvals', 40000,' MaxIter', 40000,' TolFun', 10^-4);$

What do these options do for the optimization? How does varying them impact the results if at all? What other options could be used and how would they impact estimation?

Run your non-linear optimization, and report out the values for all 21 parameters in alpha that you've estimated (alpha(6,2), alpha(6,3) and alpha(6,4) are zeros filling out the matrix.). Tell me what you get for:

- Risk preference estimates
- Inertia estimates
- PPO1200 random coefficient estimates
- Epsilon variance estimates
- Estimated preference of high total cost people for PPO250
- The final likelihood function value at the estimated parameters (i.e. the 'maximum' likelihood function value)

(d) Incorporating demographic heterogeneity and the parameter estimates, what is the population distribution of inertial costs?

(e) Consider the mean of the CARA risk preference distribution of random coefficients. Translate this coefficient into the value of X that makes a family indifferent between no gamble and a gamble where they win \$100 with 50% probability and lose \$X with 50% probability.

(f) Finally, re-run the optimization with a few different starting values that you choose. Do the parameter estimates change? If so, which ones are 'better' i.e. which ones would you select to report as estimates? Why?