PPHA 42200: Problem Set 3

Professor: Koichiro Ito

April 22, 2022

Part I: Theory Questions (MLE, Discrete Choice)

1. Consider $y = \beta x + u$, where y and x are random variables, β is a parameter, and u is the error term. Suppose that Gauss-Markov assumptions hold, and therefore, the OLS estimator is the BLUE (Best Linear Unbiased Estimator). Because OLS is the BLUE, the maximum likelihood estimators do not improve efficiency relative to the OLS estimator in this case. Answer True/False/Uncertain and explain why.

2. Consider a random utility model, $u_{ij} = \beta x_{ij} + \epsilon_{ij}$, for multinomial choice of $j = 1, ..., J$ for individual $i = 1, .., N$. You find that the probability of individual i choosing j can be expressed by $P_{ij} = (\beta x_{ij} - \sum_{k \neq j} \beta x_{ik})/C$, where C is some constant. Show mathematically if this model has the Independence of irrelevant alternatives (IIA) property or not.

3. Consider a random utility model:
$$U_{ij} = X'_{ij}\beta + \varepsilon_{ij},$$
where we assume that ε_{ij} are independent across choices and individuals, and have type I extreme value distributions.

   (a) Show that the probability that person i chooses j (P_{ij}) can be written by:
$$\begin{aligned} P_{ij} &= Pr(X'_{ij}\beta + \varepsilon_{ij} > X'_{im}\beta + \varepsilon_{im}), \text{ for all } m \neq j \\ &= Pr(\varepsilon_{im} - \varepsilon_{ij} < X'_{ij}\beta - X'_{im}\beta), \text{ for all } m \neq j \\ &= \frac{\exp(X'_{ij}\beta)}{\sum_{m=1}^{J} \exp(X'_{im}\beta)}. \end{aligned}$$

   (b) Suppose that we observe data on choices based on this random utility model for person $i = 1, ..., N$ for choice $j = 1, ..., J$. Denote the observed choice by $y_{ij} = 1(\text{person } i \text{ chooses } j)$. Derive the log likelihood function.

   (c) Derive the first order condition for the log likelihood function. You do not need to 'solve' the FOC.

   (d) What is the "Independence of Irrelevant Alternatives" assumption? Explain the assumption in words and give an example other than the classic red bus/ blue bus problem. How exactly does the assumption relate to the random utility model you worked with above?

Part II: Empirical Analysis (Matching)

Matching, Reweighting, and the Effects of Maternal Smoking on Infant Health

The questions below are based heavily on the paper Almond et al. 2005, and problem sets from Ken Chay and John DiNardo based on some of the data used in the paper. The goal of this assignment is to examine the research question: what is the causal effect of maternal smoking during pregnancy on infant birthweight and other infant health outcomes. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match.

The data in Stata format can be downloaded from our course website. There should be 48 variables in the data and, after you are finished with the cleaning steps described below, 82,466 observations.

The data here are "real" and quite imperfect, which will help simulate the unpleasantness of real world data work. Unlike the real world where you will confront this bleak situation largely alone, I will provide you with some hints for working your way through the raw data. You can download part of the codebook for the data to help you figure out the relevant variables.

1. The first order of business is to go through the code book, decide on the relevant variables, and process the data. This involves several steps:

   (a) Fix missing values. In the the data set several variables take on a value of, say, 9999 if missing. We have already checked for missing observations for about 2/3 of the variables. The remaining variables need to be checked and are the last 15 in the variable list (i.e. from 'cardiac' to 'wgain'). Refer to the codebook for missing value codes. Produce an analysis data set that drops any observation with missing values.

   (b) If this were a real research project you would want to consider other approaches to missing data besides termination with extreme prejudice. What observations do you have to drop because of missing data. Might this affect your results? Do the data appear to be missing completely at random? How might you assess whether the data appear to be missing at random?

   (c) Produce a summary table describing the final analysis data set.

2. The next part of the assignment is to try to estimate the "causal" effect of maternal smoking during pregnancy on infant birth weight. For the next questions use the clean dataset that you produced before. Let's start out using techniques that are familiar, and think about whether they are likely to work in this context. Answer the following questions.

   (a) Compute the mean difference in APGAR scores (both five and one minute ver- sions) as well as birthweight by smoking status.

   (b) Under what circumstances can one identify the average treatment effect of ma- ternal smoking by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers? Estimate its impact under this assumption. Provide and comment on some evidence for or against the validity of the assumption (A useful "Table 1" of any paper is one that describes the overall averages of the observations, and then describes the subsets of people who do and do not receive the treatment (when it is binary)).

   (c) Suppose that maternal smoking is randomly assigned conditional on the other observable "predetermined" determinants of infant birth weight. First discuss which (if any) of the variables contained in the data set can clearly be considered to be predetermined. In general, what kinds of variables can be considered predetermined and what kinds of variables cannot?

(d) What does "selection on observables" imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables? Use a basic linear regression model, in conjunction with your answer to part (c), to estimate the impact of smoking and report your estimates. Under what circumstances is the average treatment effect correctly identified by this linear regression with covariates?

3. Describe the propensity score approach to the problem of estimating the average causal effect of smoking when the treatment is randomly assigned conditional on the observ- ables. How does it reduce the dimensionality problem of multivariate matching? Try a few ways to estimate the effects of maternal smoking on birthweight:

   (a) First create the propensity score. For our purposes let's use a logit specification. First specify the logit using all of the "predetermined" covariates (don't include interactions). Next, include only those "predetermined" covariates that enter significantly in the first logit specification. How comparable are the propensity scores? If they are similar does this imply that we have the "correct" set of covariates in the logit specification used for our propensity score?

   (b) (Including propensity score as a covariate) Control directly for the estimated propensity scores using a regression analysis, and estimate an average treatment effect. State clearly the assumptions under which your estimate is correct.

   (c) (Weighting with propensity score) As discussed in class, one can use the estimated propensity scores to reweight the outcomes of non- smokers and estimate the average treatment effect. Compute an estimate of the average treatment effect and the "effect of the treatment on the treated" by appropriate reweighting of the data.

4. (Blocking on propensity score) A potentially more informative way to describe how birth weight affects smoking is to estimate the "non-parametric" conditional mean of birth weight as a function of the estimated probability of smoking, separately for smokers and non-smokers on the same graph. To do so, divide the data from smokers into 100 approximately equally spaced bins based on the estimated propensity score. Do the same for nonsmokers. Use the blocking estimator we discussed in class. Interpret your findings and relate them to the results in (3b).

5. Low birth weight births (less than 2500 grams) are considered particularly undesirable since they comprise a large share of infant deaths. Redo the last question using an indicator for low birth weight birth as the outcome of interest. Interpret your findings.

6. Concisely and coherently summarize your results above providing some intuition. Write it like you would the conclusion of a paper. In this summary, describe whether you think your best estimate of the effects of smoking is credibly identified. State why or why not.