Econ 312 Part A, Spring 2023

Problem Set 2

James J. Heckman

Due April 3rd, 2023 at Midnight

This draft: March 28, 2023

1. [**10 pts**] Answer the Questions embedded in the handout "Hypothesis Testing, Part I." Explain the irrelevance for likelihood and Bayesian inference of the "stopping rules" and sampling plans in the Lindley example of 9 red and 3 black balls.

2. [**5 pts**] Distinguish between the economist's thought experiment defining a regression and that of the statistician.

3. [**5 pts**] In discussing the law of demand for a good, it is frequently stated that the impact of price on demand is negative, conditioning on other factors that influence demand. When is this an accurate description? Distinguish fixing from conditioning.

Questions 4 and 5 will make use of the dataset "ps2_dt.csv," which simulates data following the framework presented in the handout "A Running Example Based on: Alternative Methods for Evaluating Interventions: An Overview." In the dataset, you will find the following variables:

- *indiv: indexes individuals*

- *time: indexes time* $t = 1, 2, 3, 4$*, where the job training program is offered at* $t = 2$ *(so that* $k = 2$*)*

- *Z: Z_i from the handout*
- *D: d_i from the handout*
- *Y_obs_q4: the observed outcome Y_{it} from the handout under the assumption that $\alpha_i = \alpha$ always. It should be used as the observed outcome used to answer parts of question 4.*
- *Y_obs_q5: the observed outcome Y_{it} from the handout under the assumption that $\alpha_i \sim N(\alpha, 1)$ iid. It should be used as the observed outcome used to answer parts of question 5.*

4. **[25 pts]** Consider the simple Generalized Roy economic framework presented in "Alternative Methods for Evaluating Interventions: An Overview." Handouts for it are posted on the class website. We will use this model for the rest of the course to illustrate a different approach for evaluationing policy. We will also use this notation in class and in the remaining problem sets as well as on the take home exam. We extend the model to account for heterogeneity in treatment response coefficients. We will consider equation (1) with and without α_i (α_i heterogenous response to treatment). Specialize (1) to $X_{1t}\beta = \beta_t$ for problems (i), (ii), (iii) below. I is sample size, i is an individual subscript, and sampling is independent across agents.

   (i) Using Data Set 1 and outcomes after $t = k$, discuss how you can identify and consistently estimate α, $\alpha(= E(\alpha_i))$, $E(\alpha_i|d_i = 1)$ comparing the empirical analog of $E(Y_{it}|d_i = 1)$ with $E(Y_{it}|d =$

0). Present estimates for $t = k+1$ and standard errors and also estimates of p with std errors. If $E(U_{it}d_i) \neq 0$, are the estimators consistent? Why?

(ii) Compare the means of Y formed before $t < k$ to the means formed after $t > k$ for each group $d_i = 1$ and $d_i = 0$. Do they help in identifying the model? Why? Under what conditions?

(iii) Compute the overall (d) mean $E(Y_t)$. What does this estimator identify? ($t > k$). What does $E(Y_t) - E(Y_{t'})$, $t > k > t'$ identify? (You know p but you don't know any single person's d_i.) Take 2 cases: (i) Time homogeneity; (ii) Time inhomogeneity.

5. **[20 pts]** Building on (4), suppose you have access to one cross section $t = k+1$ and you know Y_{it}, d_i and X_{it} in equation (1). You have access to Z. V has the stated properties.

(i) Compute $Pr(d_{i=1}|Z) = P(Z)$ assuming $V \sim N(0,1)$ and give standard errors of coefficients. Compare the probit estimates with logit estimates.

(ii) What is the subjective value of the program (up to scale σ_V)?

(iii) For $t > k$, compute IV using Z and $P(Z)$ as an instrument (two estimators) for two cases (α_i) is identical across i and $\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2)$. Compare estimates. Establish the asymptotic properties of the estimators.

(iv) For both models (α identical for all and α_i variable), derive the MTE as a function of V. What economic question does it estimate?

Compute the weights on MTE that give ATE, TOT, and PRTE.

6. **[10 pts]** Show that the OLS estimator of β for a single cross section in equation (1) for $t > k$ can be represented as (suppressing t subscripts)

$$\hat{\beta}_{OLS} = \sum_{1 \leq i \leq i' \leq I}^{I} \frac{(y_i - y_{i'})}{(d_i - di')} 1(d_i \neq d_{i'}) W_{i,i'}$$

where

$$(W_{i,i'}) = \frac{(d_{i'} - d_{i'})^2}{\sum_{1 \leq i \leq i' \leq I}^{I} (d_i - d_{i'})^2}$$

What does $\hat{\beta}_{\text{OLS}}$ estimate when α is the same for all and when α_i varies (two cases)? Compare these weights to the IV weights computed in problem (5). What is the interpretation of the denominator?

7. **[5 pts]** How do your responses differ to (4), (5), (6) when α_i is not known by agents when they make their enrollment decision but the base their decisions on $E(\alpha_i) = \bar{\alpha}$? They learn about α_i after they participate in the program.