ADVANCED LABOR ECONOMICS

Sergio Urzúa

Due date: October 9, 2015

Problem Set 1

Please return the problem set on time. Use your favorite software to answer the following questions.

Question 1. Consider the following discrete choice model:

$$D = 1[\alpha_0 + \alpha_1 Z > V].$$

Assume the following parameterization of the model: $Z \sim N(0, 2)$, $V \sim N(0, 1)$, $\alpha_0 = 2$, and $\alpha_1 = -1$.

(a) Using this parameterization, generate a fake sample of 1,000 observations.

(b) Code the likelihood function associated with this problem.

(c) Estimate the parameters of the model using information on (D, Z) and your ML routine. Compare your results with the actual values.

(d) Compute standard errors for your estimates.

(e) Suppose that $V \sim N(0, \sigma^2)$ with $\sigma^2 \neq 1$. Can you estimate σ^2? Show this formally.

Question 2. Consider the following model:

$$D = 1[\alpha_0 + \alpha_1 Z > V].$$

and assume you observe $Y = a + bX + \epsilon$ if $D = 1$ and you observe $Y = 0$ otherwise.

Assume the following parameterization of the model: $Z \sim N(0, 2)$, $X \sim N(0, 1)$, $\epsilon \sim N(0, 1)$, $V \sim N(0, 1)$, $a = 0.5$, $b = -0.5$, $\alpha_0 = 2$, and $\alpha_1 = -1$.

(a) Using this parameterization, generate a fake sample of 1,000 observations.

(b) Code the likelihood function associated with this problem.

[0]For questions regarding this problem set contact: Sergio Urzua (urzua@econ.umd.edu).

(c) Estimate the parameters of the model using information on (Y, D, Z, X) and your ML routine. Compare your results with the actual values.

(d) Compute standard errors for your estimates.

Question 3. Consider the following random utility model:

$$\begin{aligned} Y_1^* &= \beta_1 Z + V_1 \\ Y_2^* &= \beta_2 Z + V_2 \\ Y_3^* &= \beta_3 Z + V_3 \end{aligned}$$

and define $D = 1$ if $(Y_1^* > Y_2^*, Y_1^* > Y_3^*)$,$D = 2$ if $(Y_2^* > Y_1^*, Y_2^* > Y_3^*)$, and $D = 3$ otherwise, with

$$V \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33}^2 \end{pmatrix}\right].$$

Consider alternative 3 as the benchmark, and assume the following parameterization: $Z \sim N(0, 3)$,

$$\begin{bmatrix} V_1 - V_3 \\ V_2 - V_3 \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

$\beta_1 = 0.25$, and $\beta_2 = 0.5$, $\beta_3 = 1$.

(a) Using this parameterization, generate a fake sample of 10,000 observations.

(b) Code the associated likelihood function.

(c) Show that you can identify $\beta_1 - \beta_3$, $\beta_2 - \beta_3$, $Cov\,(V_1 - V_3, V_2 - V_3)$, and $Var\,(V_2 - V_3)$.

(d) Estimate the parameters of the model using information on (D_1, D_2, D_3, Z) and your ML routine. Compare your results with their actual values. (Hint: Don't forget to impose the normalizations).

Question 4. Empirical Selection Models. Consider the selection model discussed in Heckman 1974.

(a) Describe the equations linking theory and practice. What are the key assumptions?

(b) Download the NLSY97. Create a data set with the variables utilized by Heckman. Present descriptive statistics for your sample.

(c) Using the NLSY97, estimate the model in Heckman 1974. (Discuss your optimization routine).

(d) Explain under what assumption the average observed wage would be larger than the unconditional average wage.

Question 5. To answer this question you need to use the following information from the NLSY79: (i) Employment status by age 30 (WORKING, WITH JOB NOT AT WORK, UNEMPLOYED, KEEPING HOUSE, GOING TO SCHOOL, UNABLE TO WORK, OTHER, IN ACTIVE FORCES), (ii) Race, (iii) Gender, (iv) Highest Grade Completed by Age 30, (v) the Armed Forces Qualifications Test score (AFQT).

(a) Present the summary statistics (number of observations, means, stdev, min and max) by race and gender. Use only the representative sample of the NLSY79.

(b) Construct the following variable: $D = 1$ if working, $D = 0$ if unemployed, and consider the following discrete choice model:

$$D = 1[\beta_0 + \beta_1 Race + \beta_2 Gender + U \geq 0]$$

where *Race* is defined as $Race = 1$ if White, $= 0$ if Black, and *Gender* is equal to if female and 0 if male. Finally, assume that $U \sim N(0, 1)$.

i. Present the confidence interval for β_1 and β_2 at the 10% significant levels. Interpret your results.

ii. Do you find significant differences across race and gender? Describe your tests in detail.

(c) Economists believe that education is an important determinant of labor market outcomes. The NLSY79 contains detailed information on years/levels of education. Construct the following 5 dummy variables: High School Dropouts, High School Graduates, Some College, Two Year College Graduates and Four Year College Graduates. Incorporate these variables in your analysis by considering the following regression model:

$$D = 1[\beta_0 + \beta_1 Race + \beta_2 Gender + \beta_3 HSGrad + \beta_5 SomeColl + \beta_6 2YColl + \beta_7 4YCollU \geq 0]$$

i. In this model, is the effect of "Some College" statistically different from the return to "Two Year College"? Explain your tests.

ii. Is the effect of "Two Year College " statistically different from the return to "Four Year College"? Explain your tests.

iii. Present the marginal effects of each schooling level by race and gender.

(d) The AFQT is a general measure of trainability and a primary criterion of enlistment eligibility for the Armed Forces. It has been used in applied papers as a measure of abilities. Thus, consider the model:

$$D = 1[\beta_0 + \beta_1 R + \beta_2 G + \beta_3 HSGrad + \beta_5 SomeColl + \beta_6 2YColl + \beta_7 4YColl + \gamma AFQT + U \geq 0]$$

where R denotes race and G gender.

i. Is the parameter γ statistically significant? Is the marginal effect of AFQT statistically significant?

ii. Plot the marginal effect of AFQT on the probability of being working by schooling level and race. In each case, evaluate the variables *Gender* at its mean in the overall population. Interpret your results.

(e) In the previous questions you analyzed the determinants of a binary variable D (working vs. unemployed). Consider the following extension: $D = 1$ if working, $D = 2$ if unemployed, $D = 3$ if inactive (out of the labor force).

i. Use a multinomial logit model to analyze the effects of race, gender, schooling levels, and ability on D.

ii. Present the estimated parameters and marginal effects. Are the variables included in the model significant determinants of the employment status of the individuals in your sample.

iii. Compare your previous results with those obtained using a multinomial probit.

iv. Briefly describe the differences between a multinomial logit and a multinomial probit.

Question 6. Race, Gender and Earnings. To answer this question you need to use the following information from the NLSY79: (i) Employment status by age 30 (WORKING, WITH JOB NOT AT WORK, UNEMPLOYED, KEEPING HOUSE, GOING TO SCHOOL, UNABLE TO WORK, OTHER, IN ACTIVE FORCES), (ii) Race, (iii) Gender, (iv) Highest Grade Completed by Age 30, (v) the Armed Forces Qualifications Test score (AFQT), (vi) Annual earnings by age 30, (vii) The Rosenberg Self-Esteem Scale (administered during 1980), (viii) labor market experience[1].

[1] You can also use potential experience: age-years of education-6

Consider the following model for the analysis of the returns to schooling:

$$\ln Y = \alpha_0 + \alpha_1 Race + \alpha_2 Gender + \alpha_3 \mathrm{S} + \alpha_4 Exp + \alpha_5 Exp^2 + \epsilon \tag{1}$$

where Y =earnings, S =Years of Education.

(a) Using equation (1), explain the tests analyzed by Heckman, Lochner and Todd (HLT, 2003).

(b) Estimate the model using OLS.

- Do you find significant differences across race and gender?
- Interpret the coefficient associated with S.
- What is the return to experience?
- Following HLT (2003), test the specification of the model. In your answer you need to consider at least two of the tests analyzed by HLT (2003)

(c) Based on the previous model, present an Oaxaca decomposition for the wage differential between males and females. What fraction can be interpreted as discrimination? endowments?

(d) Based on the previous model, present an Oaxaca decomposition for the wage differential across races (White vs Non-white). What fraction can be interpreted as discrimination? endowments?

(e) Neal and Johnson (1994) discussed the importance of abilities in the context of racial discrimination. Add IQ(=AFQT) to your regression model.

- Justify your specification (additive separability, interactions, etc).
- Discuss the impact of incorporating this variable over your results. Can you confirm the finding of Neal and Johnson (1994)?
- Based on the previous model, present an Oaxaca decomposition for the wage differential between males and females. What fraction can be interpreted as discrimination? endowments?
- Based on the previous model, present an Oaxaca decomposition for the wage differential across races (White vs Non-white). What fraction can be interpreted as discrimination? endowments?

(f) Add "Self-esteem" to your model.

- Justify your specification (additive separability, interactions, etc).

- What are the effects on the coefficients associated with S, gender and race?
- Based on the previous model, present an Oaxaca decomposition for the wage differential between males and females. What fraction can be interpreted as discrimination? endowments?
- Based on the previous model, present an Oaxaca decomposition for the wage differential across races (White vs Non-white). What fraction can be interpreted as discrimination? endowments?

(g) Summarize the main findings from your Oaxaca decompositions. Discuss your results for gender and race.

(h) Urzua (2008) re-analyzed Neal and Johnson (1994). Briefly describe the main argument of this paper.

(i) A limitation of the previous linear regression model is due to the fact that earnings are observed only for those individuals working. Thus, we can write the following non-linear model for (log) earnings:

$$\begin{aligned} \ln Y &= \alpha_0 + \alpha X + U && \text{if working} \\ \ln Y &= . && \text{otherwise} \end{aligned}$$

where X includes the set of controls previously considered. We assume that employment can be modeled as:

$$D = 1[\beta_0 + \beta_1 S + \beta_2 IQ + \beta_3 Race + \beta_4 Race + V \geq 0]$$

where $V \sim N(0,1)$, and $D = 1$ if working and $D = 0$ if unemployed. Don't consider the inactive people in your analysis. Discuss whether the OLS estimate of α associated with schooling would be downward or upward biased in this set-up. Provide a simple intuition.

(j) Using the previous set-up, implement a two-stage procedure to estimate the effect of race and gender on earnings controlling for selection into the labor market. Do you find changes in the coefficients associated with race and gender?

(k) Would you argue that the racial differentials in earnings are due to "taste discrimination" or "statistical discrimination"? What is the experimental evidence telling us?

Question 7. Theory

(a) Imbens and Angrist 1994: Define LATE and show how you can estimate it. Show the importance of monotonocity. (Hint: Derive the identification argument of LATE as a causal parameter).

(b) Present a brief analysis of the differences between the schooling decision models in Mincer (1958), Willis and Rosen (1979), Card (2001) and Heckman, Lochner and Todd (2003). In your answer, describe each of the questions analyzed by the authors and explain how they combine theory and empirics to provide meaningful answers.