Problem Set 2

MIT 14.385, Fall 2007

Due: Wednesday, 10 October 2006 5pm

This problem set gives you a choice between (a) solving a number of applied/empirical problems and/or (b) proving some of the theoretical results which were stated but not proven during the lecture. Doing both the applied part (a) and the theoretical part (b) means that you don't have to take the midterm exam.

For the empirical part, you may use any software package you like. If you use a canned package such as Stata, you *must* describe exactly what it is that your nonlinear estimation command is doing (e.g. how does it iterate, what is its convergence criterion, how does it calculate the Hessian and outer product of the score, etc. Please hand in *clean, well-written, thoroughly annotated* code and *clearly formatted, readable* summaries of your empirical results. Bonus points will be given for results tables created within your code. Negative points will be awarded for unreadable, confusing or sloppy code.

1 Applied/Empirical Problems

The purpose of this exercise is to acquaint you with a few commonly applied nonlinear estimators. The data set for the empirical exercises comes from a subset of the NLSY used by Herrnstein and Murray in their book, *The Bell Curve* (no endorsement implied).

1. Smoking and Lost Workdays [Problem 15.6 from Wooldridge]

   Consider taking a large random sample of workers at a given point in time. Let $sick_i = 1$ if person i called in sick during the last 90 days, and zero otherwise. Let $\mathbf{z}_i$ be a vector of individual and employer characteristics. Let $cigs_i$ be the number of cigarettes individual i smokes per day (on average).

   (a) Explain the underlying experiment of interest when we want to examine the effects of cigarette smoking on workdays lost.

(b) Why might $cigs_i$ be correlated with unobservables affecting $sick_i$?

(c) One way to write the model of interest is

$$P(sick = 1|\mathbf{z}, cigs, q_1) = \Phi(\mathbf{z}_1\delta_1 + \gamma_1 cigs + q_1)$$

where $\mathbf{z}_1$ is a subset of $\mathbf{z}$ and q_1 is an unobservable variable that is possibly correlated with $cigs$. What happens if q_1 is ignored and you estimate the probit of $sick$ on $\mathbf{z}_1$, $cigs$?

(d) Can $cigs$ have a conditional normal distribution in the population? Explain.

(e) Explain how to test whether $cigs$ is exogenous. Does this test rely on $cigs$ having a conditional normal distribution?

(f) Suppose that some of the workers live in states that recently implemented no-smoking laws in the workplace. Does the presence of the new laws suggest a good IV candidate for $cigs$?

The following parts are not from Wooldridge:

(g) What happens if Jerry catches you running The Forbidden Regression? What is The Forbidden Regression and why is it so Forbidden? Try to answer in words and from memory rather than in equations from a textbook. If you don't remember, talk to your friends before consulting a book. Is Wooldridge asking you to run a Forbidden Regression in part (f)?

(h) In part (f), Wooldridge asks about IV in a general sense. Be more specific. What would you *actually do* to run IV in this setting? Write down the equation(s).

2. Women and Work

Consider the following model:

A woman's weekly wage rate is determined by the equation

$$\log w = x'_W\beta_W + u_W \tag{1}$$

where w is the woman's weekly wage rate and the vector x contains an intercept, education

E, age A, $AFQT$, race, A^2, $A * E$ and $AFQT * E$

Weeks of work are determined by the equation

$$\begin{aligned} Weeks &= x'_{WK}\gamma_{WK} + \gamma_W \log w + v \\ &= x'\beta_{WK} + u_{wk} \end{aligned} \tag{2}$$

where $Weeks$ is weeks worked, and x_{WK} contains an intercept, family income, and a dummy as to whether the woman is currently married or not. The vector x contains all of the unique elements of x_W and x_{WK}. The discrete variable δ is an indicator for a positive number of weeks worked, i.e. $\delta = \mathbf{1}\{Weeks > 0\}$.

For the following, use the data set women.txt.

(a) Probit and multinomial logit

i. Estimate the parameters determining labor force participation assuming that the error $u_{wk} \sim N\left(0, \sigma^2\right)$ in model (**??**). Test the hypothesis that marriage influences participation using both Wald and likelihood ratio test statistics. Construct bootstrap standard errors and compare these to the ML standard errors.

ii. Consider a three-state classification of a woman's hours of work: she doesn't work at all (designate by setting the discrete variable $b = 1$); she works part of the year which implies that $0 < Weeks < 20$ (designated by $b = 1$); and she works most of the year with $Weeks \geq 20$ ($b = 3$). Discuss whether a multinomial logit model is appropriate here. Estimate a multinomial logit model for $P(b = j \mid x)$. Discuss how to compute conditional probabilities at a given ("interesting") $\tilde{x}$. Using your estimation results, test whether marriage influences the likelihood that a woman works part of the year instead of most of the year. Finally, construct bootstrap standard errors for your estimated coefficients and compare these to the ML standard errors.

(b) Censored Samples

Assume that the errors (u_W, u_{WK}) follow a bivariate normal distribution:

$$\begin{pmatrix} u_W \\ u_{WK} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right)$$

Data are available on wages only for those who work a positive number of weeks.

i. Where does this model fit in Amemiya's classification of Tobit models (Tobit I-V)? Where did the censoring question on last week's problem set (Q5) fit in the classification? What is the essential difference between the two models?

ii. Discuss briefly why OLS will not be a satisfactory technique for estimating the coefficients β_W of model (**??**).

Since OLS is not satisfactory, we will explore more sophisticated techniques. Estimate the coefficients β_W of model (**??**) accounting for sample selection using the following three methods:

iii. Two-step estimation of the censored regression model ("censored regression model" refers to the fact that there is a selection problem here and you are observing only a (censored) subset of the whole data). Hint: see Wooldridge 563-564.

iv. Nonlinear least squares estimation of the censored regression model.

v. Maximum likelihood using the generalized tobit model ("generalized tobit" refers to the fact that the censoring point for each observation is not fixed, but depends on X's, while "usual tobit" has a fixed censoring point for every observation)

(c) Simultaneous Equations with Dummy Endogenous Variables

In place of equation (**??**), consider the wage equation:

$$\log w = x'_W \beta_W + d\alpha + u_W \tag{3}$$

where $d = 0$ if $b = 2$ and $d = 1$ if $b = 3$ (we do not observe wages for those who do not work, so we do not need to specify a value of d when $b = 1$). The coefficient α

may be thought of as representing a wage premium for full-time work. Assume that $E[u_W \mid d = 1, x] = 0$ in (**??**).

i. Briefly discuss why OLS will be biased

ii. Estimate the parameters of model (**??**) using instrumental variables.

iii. Test for the endogeneity of d.

3. Ability and Poverty/Work – use the combined data sets women.txt and men.txt for this exercise.

Consider the following logit model:

$$P = \alpha_0 + \alpha_1 Educ + \alpha_2 SES + \alpha_3 Age + \alpha_4 Sex + \alpha_5 Race + \alpha_6 AFQT + \varepsilon_1 \qquad (4)$$

(a) In model (**??**) let $P = 1$ if the individual was in poverty in 1989 and $P = 1$ otherwise.

i. Estimate (**??**) using the logit model both unrestricted and restricting $\alpha_6 = 0$.

ii. Test whether including ability as measured by the AFQT reduces the magnitude of α_5 (in absolute value).

iii. Test whether $\alpha_5 = 0$ in both the unrestricted model and the restricted model.

iv. Evaluate the following claim: "Black-white differences in poverty rates disappear once ability is accounted for."

(b) In model (**??**) let $P = 1$ if the individual was out of the labor force for more than a month in 1989 and $P = 0$ otherwise. Repeat the exercises in (i) and evaluate the following claim: "Black-white differences in labor force attachment disappear once ability is accounted for."

1.1 Overall Tips:

- We have covered the general theory of nonlinear / extremum estimators in class; this problem set will give you the opportunity to implement this knowledge. You may need to do some

background reading on your own to familiarize yourself with the specific estimators used: Wooldridge 16.1-16.7 and 17.1-17.6 is a good place to start.

- Write down the model before you start running regressions. For example, in 2(b) you should write down the correct conditional mean, adjusting for the selection problem. You have done this in previous theoretical problem sets so it should not be a problem. It will also make clear to you what to tell the computer to do.

- You should mention clearly your method of estimation–if running OLS/probit etc, write your equation clearly and say which variables are on the LHS and on the RHS. For non-linear estimators (MLE, NLS etc) write down explicitly the Q function you are maximizing and what formula you are using to get the standard errors.

- Turn in the relevant computer output and written materials to answer the following questions. Again, if you use a canned package be sure you say exactly what it is doing.

- Useful Stata commands include *heckman, nl, tobit*

- If you want a warmup / small extra-credit exercise, replicate the Keane and Wolpin example from W, Section 15.9 and / or do W 15.7.

- Question 2(b) asks you about Amemiya's classification of Tobit models. See Amemiya (1985), Chapter 10 or Wooldridge Chapters 16-17 for discussion of this classification.

2 Theory Problems

In this section we ask you to prove some results given in the lecture notes. In some cases there are a few hints in the lecture notes. Also, answers are typically available in research articles, but give a try to each problem before looking up the answers.

Some problems are hard, so we do not expect you to solve them. However, in order to earn an A grade, we expect that you will attempt to solve each problem (i.e. write down you thinking about possible lines of attack), and that a half or more of the questions will be solved correctly.

1. Prove the uniform law of large numbers (ULLN) for dominated moment functions as stated in Lemma 1 on the lecture 2 handout.

2. Prove uniform convergence of the objective function under the stochastic equicontinuity assumption (Lemma 2, lecture 2 handout). Note the hint below the statement of the Lemma in the notes.

3. Prove uniform convergence of the objective function under the Hölder continuity condition (Lemma 3, lecture 2 handout).

4. Prove uniform convergence of the objective function under convexity of the sample objective function (Lemma 4, lecture 2 handout).

5. Prove Theorem 2 (consistency under convexity) in lecture 2 notes.

6. Prove the technical result on the exchange in integration and differentiation in lecture 5 notes. Comment on the application of this lemma to information matrix inequality in "regular" likelihood models. State an example of a non-regular model where information matrix equality does not hold and hence Cramer-Rao efficiency bound does not apply. Explain how this example fails to satisfy the exchangeability of differentiation and integration.

7. Supply the details of the proof of the bootstrap consistency for the sample mean (cf. Lecture 7). You can find answers in published articles, but give it a try before you look up these answers.

8. Solve the question on the bootstrap refinement listed in lecture 7. The idea of this question is for you to work through the answer already provided in Horowitz handbook chapter, and really learn the details of this result. The details of Edgeworth expansions are mentioned in Amemiya's text and in Van der Vaart.

9. Write down and explain the details of subsampling consistency, as requested in lecture note 7. You can find a good discussion in Horowitz and Romano, Politis, Wolf's book.