Economics 628 **Term 1 2017/2018**

Topics in Applied Econometrics I **Hiro Kasahara**

Homework 1

(Due: Wednesday, Sept 20 at the start of the class)

Note: Study groups *discussing* the problems are strongly encouraged. But please write your own answers and submit your own programs (no copy and paste from your classmate's program!).

1 College Decision

Read "HW1_data" into Matlab. In this exercise, we use the first 13 variables: 'college,' 'nearc4,' 'IQ,' 'motheduc,' 'fatheduc,' 'reg662,'...'reg669'. The variable 'college' is a dummy variable that takes a value of one if the years of education is no less than 13. For definition of other variables, see David Card (1993) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" NBER Working Paper No. 4483. The "HW1_data" contains the observations that met the following sample selection criteria: (i) non-black, (ii) IQ, motheduc, and fatheduc are reported.

We denote 'college' variable for the i-th observation by y_i while other 12 variables by $x_{1,i}$,...,$x_{12,i}$ according to the order in the data set so that $x_{1,i}$ is the value of 'nearc4' for the i-th observation. Let $x_i = (1, x_{1,i}, ..., x_{12,i})'$ be a $1 \times K$ row vector with $K = 13$, where the first element represents a constant term. The number of observations is $N = 1457$.

We consider a latent variable model

$$y_i^* = x_i'\theta + \epsilon_i, \quad y_i = 1(y_i^* > 0) \tag{1}$$

where ϵ_i is a iid draw from standard normal distribution so that the likelihood function of an individual observation is $L_i(\theta) = \Phi(x_i\theta)^{y_i}[1 - \Phi(x_i\theta)]^{1-y_i}$. We estimate the parameter vector $\theta = (\theta_0, \theta_1, ..., \theta_{12})'$ by the maximum likelihood estimator $\hat{\theta}$ that maximizes the log-likelihood function $\sum_{i=1}^{N} \ln(L_i(\theta))$. Denote the score of the log likelihood for observation i by $s_i(\theta) = ((\partial/\partial\theta) \ln L_i(\theta))$.

This exercise asks you to estimate θ with standard errors by writing a code in Matlab. For your reference, I've written a matlab program "Card_HW1.m" with function m-file "g_ols.m," which estimate the linear probability model.

Please submit your Matlab/STATA codes and the output from your code.

1. Before apply your code to the actual data, you would like to check if your code is correctly working. For this purpose, we generate pseudo data set to check your code as follows.

   (a) Draw e_i from $N(0,1)$ for $i = 1, ..., N$ using 'randn'. (Type 'help randn' in command line or go to 'Help'→'Product Help' and then type 'randn'.) Compute $\tilde{y}_i^* = x_i\theta + e_i$ with

   $$\theta = (-6.76, 0.34, 0.04, 0.07, 0.11, 0.3, 0.36, 0.51, 0.67, 0.88, 0.83, 1.18, 0.56)'$$

   and then compute $\tilde{y}_i = 1(\tilde{y}_i^* > 0)$.

   (b) Write a function m-file that computes the *negative* value of log likelihood, $-\sum_{i=1}^{N} \ln(L_i(\theta))$, as its output given the value of θ, the $(N \times K)$ data matrix for explanatory variables $X = [x_1; x_2; ...; x_N]$, and the data vector for outcome variable $y = (y_1, ..., y_N)'$. (Find out the name of Matlab command to compute the standard normal cdf by yourself.)

   (c) Apply the matlab's optimization routine 'fminunc()' to the data X and $\tilde{y} = (\tilde{y}_1, ...\tilde{y}_N)'$ to estimate θ. Denote the ML estiamte by $\hat{\theta}$. To find out how to use fminunc(), type 'help fminunc' in command line or go to 'Help' → 'Product Help', and then type 'fminunc.'[1]

   (d) Prove the information matrix equality

   $$E[s_i(\theta)s_i(\theta)'] = -E[(\partial^2/\partial\theta\partial\theta')\ln L_i(\theta)]$$

   when θ is evaluated at the true value of θ.

   (e) Compute the estimate for the asymptotic variance matrix for $\hat{\theta}$ by applying the three different formula: (i) the negative of Hessian, (ii) the outer product of gradients, and (iii) "sandwich formula". Report the standard error for $\hat{\theta}$ based on these three different estimators of the asymptotic variance. Is the true value of θ in (a) inside the 90 percent confidence interval?

[1] If you do not have an access to optimization toolbox, you can download free optimization software, "csminwel", written by Chris Sims from http://sims.princeton.edu/yftp/optimize/

(f) Numerically compute the estimate for the asymptotic variance matrix for $\hat{\theta}$ by using the outer product of gradients estimation for MLE. Use numerical derivatives to compute the score. You can write your own code for this but "OPG.m," which numerically compute the outer product of gradients given the vector of individual observation's loglikelihood is also provided. If you use "OPG.m," please go through the code inside of OPG.m and understand what's going on.

2. Generate a pseudo data set $\tilde{y}$ by drawing e_i from $N(0,2)$ instead of $N(0,1)$. Estimate the value of θ using the same program you used in Question 1. Explain why the estimate is now very different from the previous one.

3. Apply your Matlab code to the actual data X and $y = (y_1, ..., y_N)$. Report the estimate of θ with standard errors in a table. Also, write STATA code to estimate θ and report the estimate with standard errors.

4. Compute the effect of an increase in IQ by one unit on the probability of going to college for each individual and then take its average across individuals. This is the average partial effect of IQ on college attendance probability. Compute the standard error for this average partial effect of IQ using the delta method. How are they different from the estimate of the linear probability model?

5. One possible policy to increase the college attendance rate is to improve the geographic access to colleges. In the data, 424 out of $N = 1457$ individuals were not living near 4 year colleges when graduated from high school; among them, only 243 individuals attended college. This implies that the average college attendance rate among these 424 individuals is 57.3 percent (as opposed to 64.2 percent of the overall average attendance rate among 1457 individuals). What is the partial effect of building 4 year colleges near the living place of these 424 individuals on the average college attendance rate among them? How are they different from the estimate of the linear probability model?

   [Hint: Compute the counterfactual probability for those 424 individuals to go to colleges and then take their average. Compute the difference between the predicted probability under the actual value of 'nearc4=0' and the 'counterfactual' predicted probability under 'nearc4=1'. The standard error can be computed by applying the delta method.]

2 MLE

Consider a discrete choice model:

$$y_i = 1(x_i'\beta + \epsilon_i > 0) \tag{2}$$

where $\epsilon_i | x_i \sim_{iid} N(0,1)$. $\{Y_i, X_i\}_{i=1}^n$ are independently drawn from the model (2).

1. Derive the asymptotic distribution of the maximum likelihood estimator (MLE) for β.

2. Suppose that we are interested in the average partial effect of β on the probability of $Y_i = 1$.

   (a) Derive the sample analog estimator of $E_X\left[\frac{\partial \Pr(Y=1|X)}{\partial X}\right]$.

   (b) Derive the asymptotic variance of the sample analog estimator of $E_X\left[\frac{\partial \Pr(Y=1|X)}{\partial X}\right]$.

Answer: This is done in class.

3 MLE for an endogenous regression model

Consider the following model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where x_i and ϵ_i are jointly normally distributed as:

$$\begin{pmatrix} \epsilon_i \\ x_i \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\begin{pmatrix} 0 \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon x} \\ \sigma_{\epsilon x} & \sigma_x^2 \end{pmatrix}\right).$$

We observe $\{y_i, x_i\}_{i=1}^n$. Define $\rho := \frac{\sigma_{\epsilon x}}{\sigma_\epsilon \sigma_x}$ and let $\theta = (\mu_x, \beta_0, \beta_1, \sigma_\epsilon, \sigma_x, \rho)'$.

1. What is the conditional density function of y_i given x_i?

2. What is the marginal density function of x_i?

3. Can we estimate θ by MLE? Discuss identification of θ, i.e., when the sample size n is ∞. In particular, what can we say about the identification of β_1?

Answer: $E[\epsilon_i|X_i] = \rho\frac{\sigma_\epsilon}{\sigma_x}X_i$ and $\text{Var}(Y_i|X_i) = \text{Var}(\epsilon_i|X_i) = (1-\rho^2)\sigma_\epsilon^2$. Therefore, the conditional density function of y_i given x_i and the marginal density function, respectively, are

$$f(y_i|x_i;\theta) = \frac{1}{\sqrt{(1-\rho^2)}\sigma_\epsilon}\phi\left(\frac{y_i - \beta_0 - (\beta_1 + \rho(\sigma_\epsilon/\sigma_x))x_i}{\sqrt{(1-\rho^2)}\sigma_\epsilon}\right) \quad \text{and} \quad f(x_i;\theta) = \frac{1}{\sigma_x}\phi\left(\frac{x_i - \mu_x}{\sigma_x}\right).$$

Let $\tilde{\beta}_1 := \beta_1 + \rho(\sigma_\epsilon/\sigma_x)$ and $\tilde{\sigma}_{y|x} := \sqrt{(1-\rho^2)}\sigma_\epsilon$. μ_x and σ_x are identified from the MLE using the marginal density function of x while $(\beta_0, \tilde{\beta}_1, \tilde{\sigma}_{y|x})'$ is identified from the MLE using the conditional density function of y given x. Therefore, we may point identify $(\mu_x, \beta_0, \sigma_x)$. On the other hand, we may not uniquely determine the value of β_1, σ_ϵ, and ρ from the value of $\tilde{\beta}_1$, $\tilde{\sigma}_{y|x}$, and σ_x. Any way to restrict the possible value of β_1 (i.e. partial identification)? From $\tilde{\beta}_1 := \beta_1 + \rho(\sigma_\epsilon/\sigma_x)$, we have

$$\tilde{\beta}_1 - \max_{\rho\sigma_\epsilon}(\rho\sigma_\epsilon/\sigma_x) \le \beta_1 \le \tilde{\beta}_1 - \min_{\rho\sigma_\epsilon}(\rho\sigma_\epsilon/\sigma_x), \tag{3}$$

where the value of $\tilde{\beta}_1$ and σ_x is known. It is possible that $\tilde{\sigma}_{y|x} := \sqrt{(1-\rho^2)}\sigma_\epsilon$ provides some restriction on the possible value of $\rho\sigma_\epsilon$. However, this turns out to be not the case because $\tilde{\sigma}_{y|x} := \sqrt{(1-\rho^2)}\sigma_\epsilon$ implies that $\tilde{\sigma}_{y|x} := \sqrt{(1-\rho^2)/\rho^s}\text{sign}(\rho)\rho\sigma_\epsilon$ so that

$$\min_{\rho\in[-1,1]} \frac{\tilde{\sigma}_{y|x}}{\sqrt{(1-\rho^2)/\rho^s}\text{sign}(\rho)} \le \rho\sigma_\epsilon \le \max_{\rho\in[-1,1]} \frac{\tilde{\sigma}_{y|x}}{\sqrt{(1-\rho^2)/\rho^s}\text{sign}(\rho)}$$

which implies that $-\infty \le \rho\sigma_\epsilon \le \infty$. It follows from (3) with $-\infty \le \rho\sigma_\epsilon \le \infty$ that $-\infty \le \beta_1 \le \infty$. Therefore, the data does not provide any information on β_1.

4 Memo on Question 1.5

We may derive the variance for the average partial effect in Question 1.5 as follows.

Define $D_i = 1$ if near4c=0 and 0 otherwise. Let $q(x_i, \beta) = [G(x_{1i}'\beta) - G(x_i'\beta)]D_i$, where x_{1i} equal to the value of x_i except that the value of "near4c" is replaced with 1 if it is equal to 0. Then, the average partial effect in the population is written as

$$APE = E[G(x_{1i}'\beta_0) - G(x_i'\beta_0)|D_i = 1] = \frac{E[q(x_i, \beta_0)]}{E[D_i]}$$

while its estimate is

$$\widehat{APE} = \frac{(1/n)\sum_{i=1}^n q(x_i, \hat{\beta})}{(1/n)\sum_{i=1}^n D_i}.$$

Then, by adding and subtracting terms, and applying $(1/n)\sum_{i=1}^n D_i = E[D_i] + o_p(1)$, $(1/n)\sum_{i=1}^n \nabla_{\beta'} q(x_i, \beta_0) = E[\nabla_{\beta'} q(x_i, \beta_0)] + o_p(1)$, $\sqrt{n}(\hat{\beta} - \beta_0) = -E[H(w_i, \beta_0)]^{-1}(1/\sqrt{n})\sum_{i=1}^n s(w_i, \beta_0) + o_p(1)$ with the delta method, we get

$$\begin{aligned}\sqrt{n}(\widehat{APE} - APE) &= \frac{(1/\sqrt{n})\sum_{i=1}^n q(x_i, \hat{\beta}) - q(x_i, \beta_0)}{(1/n)\sum_{i=1}^n D_i} + \frac{(1/\sqrt{n})\sum_{i=1}^n q(x_i, \beta_0) - E[q(x_i, \beta_0)]}{(1/n)\sum_{i=1}^n D_i} \\ &\quad + \frac{E[q(x_i, \beta_0)]}{(1/n)\sum_{i=1}^n D_i E[D_i]}(1/\sqrt{n})\sum_{i=1}^n (D_i - E[D_i]) \\ &= -\frac{E[\nabla_{\beta'} q(x_i, \beta_0)]E[H(w_i, \beta_0)]^{-1}}{E[D_i]}\frac{1}{\sqrt{n}}\sum_{i=1}^n s(w_i, \beta_0) + \frac{1}{E[D_i]}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{q(x_i, \beta_0) - E[q(x_i, \beta_0)]\} \\ &\quad + \frac{E[q(x_i, \beta_0)]}{E[D_i]^2}\frac{1}{\sqrt{n}}\sum_{i=1}^n (D_i - E[D_i]) + o_p(1)\end{aligned}$$

so that

$$\sqrt{n}(\widehat{APE} - APE) = B'\frac{1}{\sqrt{n}}\sum_{i=1}^n T_i + o_p(1)$$

with

$$B = \begin{pmatrix} -E[\nabla_{\beta'} q(x_i, \beta_0)]E[H(w_i, \beta_0)]^{-1}/E[D_i] \\ 1/E[D_i] \\ E[q(x_i, \beta_0)]/E[D_i]^2 \end{pmatrix} \quad \text{and} \quad T_i = \begin{pmatrix} s(w_i, \beta_0) \\ q(x_i, \beta_0) - E[q(x_i, \beta_0)] \\ D_i - E[D_i] \end{pmatrix}.$$

Thus, the asymptotic variance of $\sqrt{n}(\widehat{APE} - APE)$ is given by $B'\text{Var}(T_i)B$, where we can consistently estimate B and $\text{Var}(T_i)$ by their sample counterpart evaluated at $\hat{\beta}$; for example, $\hat{\text{Var}}(T_i) = (1/n)\sum_{i=1}^n \hat{T}_i\hat{T}_i'$ with $\hat{T}_i = (s(w_i, \hat{\beta}), q(x_i, \hat{\beta}) - (1/n)\sum_{i=1}^n q(x_i, \hat{\beta}), D_i - (1/n)\sum_{i=1}^n D_i)'$.