hints for practice sheet and problem set 3

1. One of the classical assumptions for OLS method states $E(\epsilon|X) = 0$:

- Show that $E(\epsilon|X)$ implies $E(\epsilon) = 0$. (Hint: use the law of iterated expectation.)
- Show $E(\epsilon|X)$ implies $E(X'\epsilon) = 0$
- Finite sample properties of the OLS estimator hold for any given sample size, true or false?
- Explain what is the law of large number.
- Explain what is the central limit theorem.
- Show that the matrix $M = (I - X(X'X)^{-1}X')$ is idempotent.

unbias will hold . Efficiene will depend on sample size

4. In a regression model $y = X\beta + \epsilon$, some columns of X represent endogenous independent variables.

- Explain the problem of endogenous independent variables, what is its consequence in terms of finite and asymptotic properties of OLS estimator?

1

- If you have a matrix of exogenous variables Z where the dimension of Z is identical to the dimension of X. How do you use Z as the instrumental variables (IV) for X to estimate the parameter β. Provide the formula of IV estimator, $\hat{\beta}_{IV}$.
- If you have a matrix of exogenous variables Z where the dimension of Z is larger than the dimension of X. How do you construct the best instrumental variables (IV) for X to estimate the parameter β. Provide the formula of IV estimator, $\hat{\beta}_{IV}$.
- Is the IV estimator unbiased? Why the IV estimation result should be interpreted as "local average treatment effect"?

Generally the IV estimator is not unbiased. but consistent

Intuitively $z \neq x$

Proof. (draft)

$$E(\hat{\beta}_{IV}) = \beta + E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i^* z_i^*\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i^* u_i\right]$$

$$= \beta + E\left[E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i^* z_i^{*\prime}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i^* u_i \,\middle|\, z_i^*, x_i\right]\right]$$

$$= \beta + E\left[\left[\frac{1}{N}\sum_{i=1}^{N} x_i^* z_i^*\right]^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i^* E[u_i \mid z_i^*, x_i^*]\right]$$

$E[u_i \mid z^*, x_i^*] \neq 0$ (in general)

local average treatment effect $\neq$ Global average treatment effect

(Average treatment effect).

on average

How x effect y

The general theme of this lecture is that with heterogenous treatment effects, endogeneity creates severe problems for identification of population averages. Population average causal effects are only estimable under very strong assumptions on the effect of the instrument on the endogenous regressor (“identification at infinity”, or under the constant treatment effect assumptions). Without such assumptions we can only identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressors. We refer to such subpopulations as compliers, and to the average treatment effect that is point identifed as the local average treatment effect. This terminology stems from the canonical example of a randomized experiment with noncompliance. In this example a random subpopulation is assigned to the treatment, but some of the individuals do not comply with their assigned treatment.

cite from
http://www.cedlas-er.org/sites/default/files/cer_ien_activity_files/miami_late.pdf
Imbens, Lecture Notes 2, Local Average Treatment Effects, IEN, Miami, Oct ’10

6. Consider analysis for binary dependent variable y where $y = 1$ or 0.

- If we use a Logit model, $Pr(y = 1|\mathbf{X}) = \Lambda(\mathbf{X}\beta) = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)}$, show the marginal effect, $\frac{\partial E(y|\mathbf{X})}{\partial \mathbf{X}}$.
- Write down the log-likelihood function of the Logit model given data $y = (y_1, \cdots, y_n)'$ and $\mathbf{X} = (x_1', \cdots, x_n')'$ and explain why we need the numerical optimization method to find the MLE.
- If we study binary dependent variables by a random utility model, $U_a = \alpha_a + \beta_a x + \epsilon_a$ and $U_b = \alpha_b + \beta_b x + \epsilon_b$, and $y = 1$ if $U_a > U_b$. Can we identify structural parameters α_a,α_b, β_a and β_b from the estimation result? If not, what can be identified? Also explain why we need to normalize the variance of ϵ_a and ϵ_b to one during estimation.

#6.2

NO close form.

$$\text{Let } p(y|x) \triangleq p = \frac{\exp(x\beta)}{1+\exp(x\beta)}$$

$$L_i = p^{y_i}(1-p)^{1-y_i}$$

$$\ln L_i = l_i = y_i \ln p + (1-y_i)\ln(1-p)$$

$$= y_i \ln \frac{\exp(x\beta)}{1+\exp(x\beta)} + (1-y_i)\ln\left[1-\frac{\exp(x\beta)}{1+\exp(x\beta)}\right)$$

#6.3

Example – Binary Response Model

- The binary choice model can be motivated from the random utility theory (McFadden)
- It is assumed that an individual's binary decision depends on utilities of two alternative choices, U_a and U_b.
- $U_a = X_a\beta + \epsilon_a$ and $U_b = X_b\beta + \epsilon_b$.
- Individual chooses $y = 1$ if $U_a > U_b$ and 0 if $U_a \leq U_b$

$$\begin{aligned} Pr(y=1|X) &= Pr(U_a > U_b|X) \\ &= Pr(X_a\beta + \epsilon_a - X_b\beta - \epsilon_b > 0|X) \\ &= Pr((X_a - X_b)\beta + (\epsilon_a - \epsilon_b) > 0|X) \\ &= Pr((X_a - X_b)\beta + \epsilon > 0|X). \end{aligned}$$

- If ϵ_a and ϵ_b are both extreme Type-I distributed, then $\epsilon = \epsilon_a - \epsilon_b$ is logistic distributed.
- We impose a normalization constraint, $\text{var}(\epsilon) = 1$, on the above model for identification.

12 / 28

No. 2 equations can't identify 4 unknow

#6.3

$$P(y=1) = P(U_a > U_b)$$

$$= P(\alpha_a + \alpha_\beta + \varepsilon_a > \alpha_a + \beta_b + \varepsilon_b)$$

$$= P\left(\varepsilon_a - \varepsilon_b > \underbrace{(\alpha_b - \alpha_a)}_{A} + \underbrace{(\beta_b - \beta_a)}_{B} x\right)$$

can be identified

- Because the distribution characteristics we need to normalize the variance of ε_a and ε_b to one during estimation.
- Extreme value type I distribution

  $X - Y \sim$ logistic but only if

  X and Y have the same variance

7. Suppose that n i.i.d. trials of x from a binomial distribution, i.e., $x_i \sim Bin(N, p)$, where N is known and p is to be estimated. The probability function is

$$f(x_i|p) = \frac{N!}{x_i!(N-x_i)!} p^{x_i}(1-p)^{N-x_i}$$

- Please write down the loglikelihood function for estimating p based on $x = (x_1, \cdots, x_n)$.
- Derive the maximum likelihood estimator (MLE) $\hat{p}_{mle}$.
- Derive the variance of MLE, $\mathrm{var}(\hat{p}_{mle})$.
- Explain what is Cramér-Rao lower bound and answer whether your answer to the previous question reaches it or not.

Asymptotic Efficiency

- Cramér-Rao Lower bound: under certain regularity conditions (i.e., the score function has mean zero and the information matrix equality holds), the variance of an unbiased estimator of the parameter vector θ estimated from n independent sample observations will always be at least as large as $\frac{1}{n}\mathcal{I}(\theta_0)^{-1}$.
- The asymptotic variance of the MLE estimator $\hat{\theta}$ exactly equals to the Cramér-Rao bound and is therefore efficient.
- This is similar to the Gauss-Markov theorem which establishes a lower bound for unbiased estimator in homoskedastic linear regression.

Beamer 23. 24

24 / 24

Conditional information matrix equality (CIME)

$-E[H_i(\theta_0)|x]$

$= E[S_i(\theta_0)S_i(\theta_0)'|x]$

Unconditional information matrix equality (UIME)

$-E[H_i(\theta_0|x]$

$= E[S_i(\theta_0)S_i(\theta_0)']$

Under certain regularity condition

① Mean zero ② CIME & UIME holds

↓

Cramer-Rao Inequality

↓

Cramer-Rao lower bond

↓

Find efficient MLE estimator

Cramer-Rao Inequality

Let $X_1 \dots X_n$ be an iid sample with pdf $f(x, \theta)$.

Let $\hat{\theta}$ be an unbiased estimator of θ; i.e., $E(\hat{\theta}) = \theta$.

If $f(x, \theta)$ is regular then $\mathrm{var}(\hat{\theta}) \geq I(\theta|x)^{-1}$

where $I(\theta|x) = -E[H(\theta|x)]$ denote sample information matrix

I. Let $x_1, \cdots, x_n$ be iid Bernoulli(p).

1. Write down the likelihood function of p
2. Derive the MLE estimator of p, $\hat{p}_{mle}$
3. Derive the variance of MLE, $var(\hat{p}_{mle})$.

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{x_i}$$

$$l(p) = \log p \sum_{i=1}^{n} x_i + \log(1-p) \sum_{i=1}^{n} (1-x_i)$$

$$\frac{\partial l(p)}{\partial p} = \cdots \overset{set}{=} 0$$

$$\hat{p}_{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\vdots$$

$$\hat{p}_{mle} \overset{A}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

II. The random variable Y comes from i.i.d. Poisson distribution, i.e.,

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

1. Obtain the log-likelihood function for a random sample of n observations, i.e., $(y_1, y_2, \cdots, y_n)$.
2. Obtain the maximum likelihood estimator of λ.
3. Explain how to estimate the variance of the MLE estimator for λ.
4. If we formulate λ by $\exp(X\beta)$, then we obtain $\log(E(Y|X)) = X\beta$, the so-called Poisson regression. Please write down the log-likelihood function of β and discuss why a numerical method such as Newton method is needed to obtain the MLE of β?

$$L(\lambda; y_1 \dots y_n) = \prod_{j=1}^{n} \frac{\lambda^{y_j} e^{-\lambda}}{y_j!}$$

$$l(\lambda; y_1 \dots y_n) = \sum_{j=1}^{n} \ln\left(\frac{\lambda^{y_j} e^{-\lambda}}{y_j!}\right) = \sum_{j=1}^{n} \left[\ln(\lambda^{y_j}) + \ln(e^{-\lambda}) - \ln(y_j!)\right]$$

$$= \dots$$

$$= -n\lambda - \ln(\lambda) \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \ln(y_j!)$$

$$\frac{d}{dl}(\lambda; y_1 \dots y_n) = -n + \frac{1}{\lambda} \sum_{j=1}^{n} y_j$$

$$\hat{\lambda} = \frac{1}{n} \sum_{j=1}^{n} y_j$$

$$\hat{\lambda} \overset{A}{\sim} N\left(\lambda, \frac{\lambda}{n}\right)$$

$$\vdots$$