Introduction

- Chris, WMC 3639, cmuris@sfu.ca

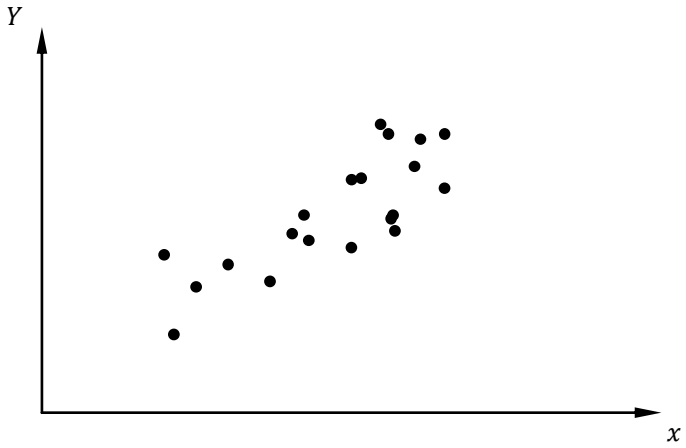
- Textbook and organization of the course
  - Part I: read it (for fun)
  - Part II: core methods
    - Ch.4. – prerequisite
    - Ch.5. – review (maximum likelihood)
    - Ch.6. – review (GMM)
    - Ch.9. – depends on you (semi-parametric)
  - Part III: bootstrap (probably skip)
  - Part IV: important!! (except 17, 18, 19)
  - Part V: panel data, 21, 22, maybe 23
  - Part VI: Ch.25 (program evaluation)

Linear Model

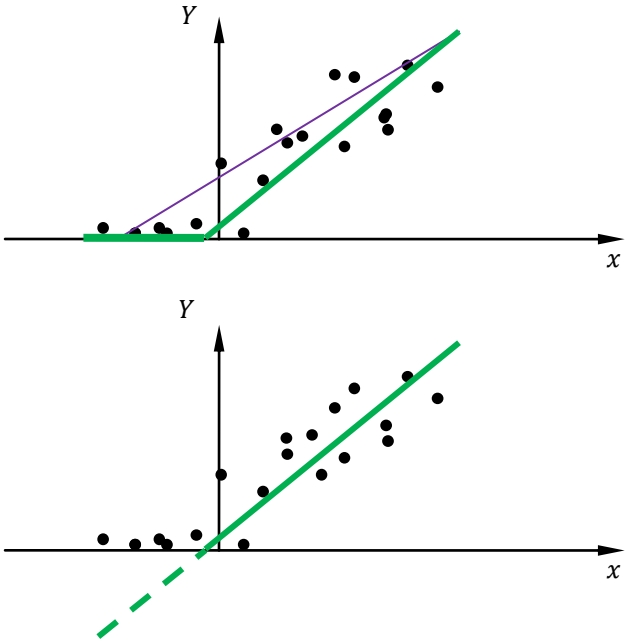
- (Y_i, X_i) iid

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \beta = (\beta_0, \beta_1), \qquad E[\epsilon_i, |X_i] = 0$$

$$\Rightarrow \ \beta_{OLS} = \left(\sum_i X_i^2\right)^{-1} \sum_i X_i Y_i$$

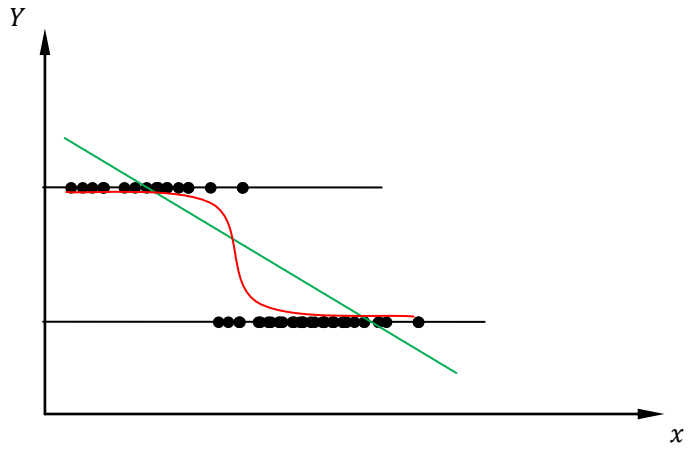


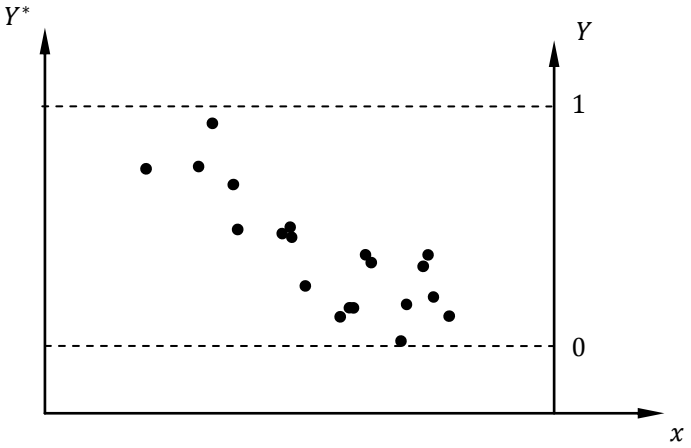
- Censored Data



  - The problem with this type of data is that they create non-linearity: there is kink at the cutoff point. Thus the linearity assumption of OLS is violated, and the estimates are going to be biased.

❖ OLS – $Y_i \in \mathbb{R}$ Cens – $Y_i \in \mathbb{R}^+ = [0.\infty)$ Binary choice – $Y_i \in \{0,1\}$
Multinomial – $X_i \in \{1, \dots, J\}$





❖ Program evaluation – how well does a program (e.g. education voucher) work?

Y_{i1} edu / wage treatment
Y_{i0} edu / wage control
$E[Y_{i1} - Y_{i0}] = E[Y_{i1}] - E[Y_{i0}]$

➢ But selection issue may bias the estimated result.

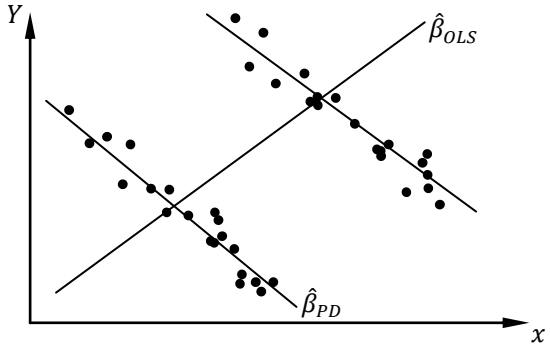
❖ Penal Data

➢ Standard linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

➢ Panel data model:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \epsilon_{it}$$



Maximum Likelihood

- Idea: What is the probability that I see the dataset that I see? Since I'm seeing this data, then I'll choose the parameter that gives the highest likelihood of generating this dataset.

- Examples.
  - Observe: 2 heads and 2 tails in a coin-flip. $\Theta = \{0.2, 0.6\}$

$$L(P) = P^2(1-P)^2$$
$$\left.\begin{array}{l} P = 0.2 \Rightarrow L(0.2) = 0.04 \times 0.64 \\ P = 0.6 \Rightarrow L(0.6) = 0.36 \times 0.16 \end{array}\right\} \Rightarrow \hat{P}_{ML} = 0.6$$

  - $\Theta = [0,1]$, Y_i. Likelihood contribution:

$$P^{\mathbf{1}\{Y_i=1\}} \cdot (1-p)^{\mathbf{1}\{Y_i=0\}}$$

Then the likelihood function:

$$L(P) = \prod_{i=1}^{N} P^{\mathbf{1}\{Y_i=1\}} \cdot (1-p)^{\mathbf{1}\{Y_i=0\}}$$

$$\Rightarrow \mathcal{L}(P) \equiv \ln L(P) = \sum_{i=1}^{N} \mathbf{1}_{\{Y_i=1\}} \ln P + \sum_{i=1}^{N} \mathbf{1}_{\{Y_i=0\}} \ln(1-P) = n_1 \ln P + n_0 \ln(1-P)$$

$$\Rightarrow \frac{\partial \mathcal{L}(P)}{\partial P} = \frac{n_1}{P} - \frac{n_0}{1-P} \overset{@\hat{P}_{ML}}{=} 0 \Rightarrow \frac{1-\hat{P}_{ML}}{\hat{P}_{ML}} = \frac{n_0}{n_1} \Rightarrow \hat{P}_{ML} = \frac{n_1}{N}$$

  - $Y_i = \mu + \epsilon_i$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $Y_i \sim \mathcal{N}(\mu, \sigma^2)$

$$f(Y_i) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot \exp\left\{-\frac{(Y_i-\mu)^2}{\sigma^2}\right\}$$

$$L(\mu, \sigma) = (2\pi)^{-\frac{n}{2}} \cdot \sigma^{-n} \cdot \prod_{i=1}^{n} \exp\left\{-\frac{(Y_i-\mu)^2}{\sigma^2}\right\}$$

$$\mathcal{L}(\mu) = -\frac{n}{2}\ln 2\pi - n\ln\sigma + \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\Rightarrow \frac{\partial \mathcal{L}(\mu)}{\partial \mu} = -\frac{2}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu) \overset{@\hat{\mu}_{ML}}{=} 0 \Rightarrow \hat{\mu}_{ML} = \frac{\sum_i Y_i}{n}$$

  - Suppose now $\mu = X_i\beta$, where $X_i, \beta \in \mathbb{R}$. Then the likelihood function is

$$L(\beta) = (2\pi)^{-\frac{n}{2}} \cdot \sigma^{-n} \cdot \prod_{i=1}^{n} \exp\left\{-\frac{(Y_i - X_i\beta)^2}{\sigma^2}\right\}$$

$$\mathcal{L}(\beta) = -\frac{n}{2}\ln 2\pi - n\ln\sigma + \frac{1}{\sigma^2}\sum_{i}(Y_i - X_i\beta)$$

$$\Rightarrow \frac{\partial \mathcal{L}(\beta)}{\partial \beta} = -\frac{2}{\sigma^2} \cdot \sum_{i}(Y_i - X_i\beta)X_i \overset{@\hat{\beta}_{ML}}{=} 0 \Rightarrow \hat{\beta}_{ML} = \left(\sum_{i=1}^{n} X_i^2\right)^{-1} \sum_{i=1}^{n} X_i Y_i$$

Properties of Maximum Likelihood Estimator

Cameron & Trivedi, Chapter 5
Background Appendix A
CH. 5: 5.1, 5.2, 5.3, 5.6
More. Newey, McFadden, *Handbook of Econometrics IV*, Chpt 36

❖ N iid observations, $\{(Y_i, X_i), i = 1, \dots, N\}$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \qquad X = \begin{pmatrix} X_1' \\ \vdots \\ X_N' \end{pmatrix}, \qquad \theta \in \Theta \subset \mathbb{R}^q, \qquad \theta_0 \text{ is the true value of } \theta$$

❖ Extremum Estimators

➢ Consider an objective function

$$Q_N(\theta, Y, X) = Q_N(\theta)$$

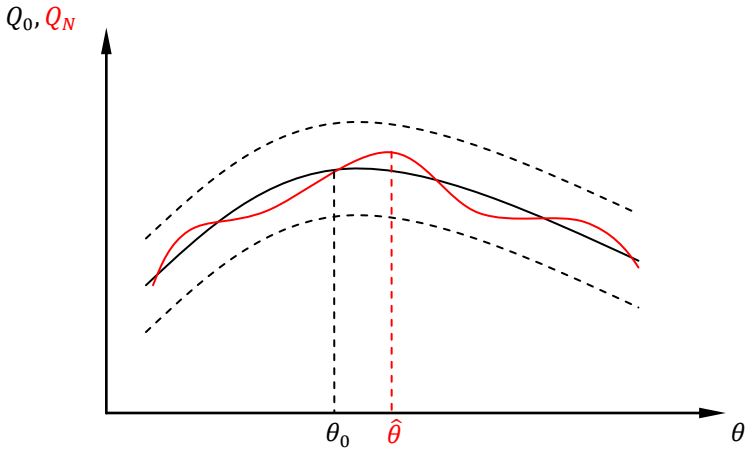
with

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_N(\theta)$$

➢ ***Consistency***: $\hat{\theta} \xrightarrow{p} \theta_0$ (want: estimator converges in probability to true parameter)

1) Θ is compact
2) There is a *true* model, Q_0, which has a minimum at θ_0
3) Q_N is continuous and measureable
4) $Q_0(\theta)$ is the true criterion function, and $Q_N(\theta)$ is the sample criterion function.
   - Q_N converges uniformly in probability to Q_0:

$$\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \xrightarrow{p} 0$$



▪ *Proof*. Suppose $\hat{\theta} \neq \theta_0$

$$\begin{aligned} 0 &< Q_0(\theta_0) - Q_0(\hat{\theta}) \\ &= Q_0(\theta_0) - Q_N(\hat{\theta}) + Q_N(\hat{\theta}) - Q_0(\hat{\theta}) \\ &< Q_0(\theta_0) - Q_N(\theta_0) + Q_N(\hat{\theta}) - Q_0(\hat{\theta}) \\ &< 2 \cdot \sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \xrightarrow{p} 0 \end{aligned}$$

- Add identification assumption: Q_0 has a unique maximum at θ_0; that is, for any $\theta \neq \theta_0$, $Q_0(\theta) < Q_0(\theta_0)$

$$P\left(\left|\hat{\theta}-\theta_0\right|>\delta\right)\le P\left(\left|Q_0\left(\hat{\theta}\right)-Q_0(\theta_0)\right|>\epsilon\right)\ \Rightarrow\ \hat{\theta}\xrightarrow{p}\theta_0$$

- ***Asymptotic Normality***.
  - $\hat{\theta}=\arg\max Q_N(\theta)$
  - $D_N(\theta)=\left.\frac{\partial Q_N}{\partial\theta}\right|_\theta$ with $D_N\left(\hat{\theta}\right)=0$
  - Use a Taylor expansion:

$$D_N\left(\hat{\theta}\right)=D_N(\theta_0)+H_N(\theta^+)\left(\hat{\theta}-\theta_0\right),\qquad H_N(\theta)=\left.\frac{\partial^2 Q_N(\theta)}{\partial\theta\,\partial\theta'}\right|_\theta$$

$$0=D_N(\theta_0)+H_N(\theta^+)\left(\hat{\theta}-\theta_0\right)$$

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=-\underbrace{\left[H_N(\theta^+)\right]^{-1}}_{\lim_{N\to\infty}H_N(\theta^+)=H_0}\sqrt{N}D_N(\theta_0)$$

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}-H_0^{-1}\underbrace{\boxed{\sqrt{N}D_N(\theta_0)}}_{\text{subject to CLT}}=-H_0^{-1}\mathcal{N}(0,B_0)=\mathcal{N}\underbrace{\left(0,H_0^{-1}B_0H_0^{-1}\right)}_{\substack{H_0\text{ is invertible}\\ \text{and }H_0<\infty}}$$

❖ Maximum Likelihood Estimator (MLE) is an extremum estimator

$$Q_N=\frac{1}{N}\sum_{i=1}^{N}\ln L(\theta)=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\ln f(Y_i,X_i|\theta)\\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\ln f(Y_i|X_i,\theta)f(X_i|\theta)\end{cases}$$

- LLN

$$Q_N(\theta)\xrightarrow{p}Q_0(\theta)=E[\ln f(Y_i|X_i,\theta)]$$

Evaluate the criterion function $Q_0(\cdot)$ at θ_0 and $\tilde{\theta}\neq\theta_0$

$$\begin{aligned}E\left[\ln f\left(Y_i\middle|X_i,\tilde{\theta}\right)\right]-E[\ln f(Y_i|X_i,\theta_0)]&=E\left[\ln\left(\frac{f\left(Y_i\middle|X_i,\tilde{\theta}\right)}{f(Y_i|X_i,\theta_0)}\right)\right]\\&<\ln E\left[\frac{f\left(Y_i\middle|X_i,\tilde{\theta}\right)}{f(Y_i|X_i,\theta_0)}\right]\\&=\ln\int\frac{f\left(Y_i\middle|X_i,\tilde{\theta}\right)}{f(Y_i|X_i,\theta_0)}\,f(Y_i|X_i,\theta_0)\,dy\\&=\ln\int f\left(Y_i\middle|X_i,\tilde{\theta}\right)dy\\&=0\end{aligned}$$

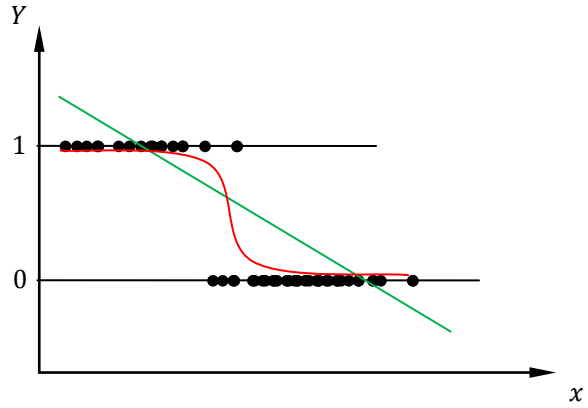
Binary Choice

C & T Chapter 14, study parts 1, 2, 3, 4; read 8

- Introduction
- Set up Model
- Estimation
- Identification
- Marginal effects, interpretation of parameters
- Properties of estimator

❖ Intro: choice of fishing mode

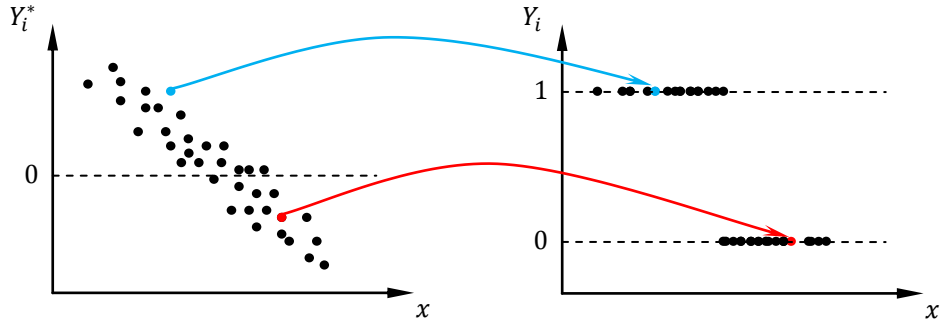
➢ $Y_i = \begin{cases} 1 = \text{fish from boat} \\ 0 = \text{fish from pier} \end{cases}$ and X_i = relative price of boat v.s. pier



- Why not OLS?
  - $E[Y_i|X_i] = \Pr(Y_i = 1|X_i) = x_i\beta \notin \{0,1\}$ but $\Pr(Y_i = 1|X_i) = x_i'\beta \in [0,1]$ → this leads to interpretational issues
  - $V[Y_i|X_i] = E[Y_i^2|X_i] - (E[Y_i|X_i])^2 = \underbrace{p(1^2) + (1-p)0^2}_{E[Y_i^2|X_i]} - p^2 = p(1-p) = X_i'\beta(1 - X_i'\beta)$ → this leads to heteroscedasticity

➢ Interpret Y_i^* as the "attractiveness of boat fishing to person i

$$Y_i^* = X_i'\beta + u_i, \qquad E[u_i|X_i] = 0$$
$$Y_i^* \geq 0 \;\Rightarrow\; Y_i = 1$$
$$Y_i^* \leq 0 \;\Rightarrow\; Y_i = 0$$

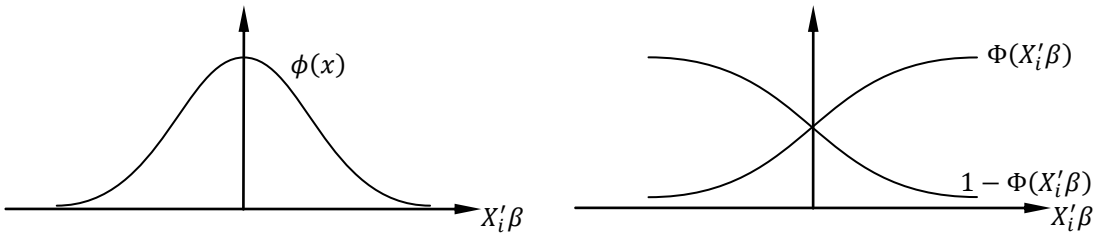


Let $\Pr(u_i \leq u|X_i) = F(u)$.

$$\Pr(Y_i = 1|X_i) = \Pr(Y_i^* \geq 0|X_i)$$

$$= \Pr(X_i'\beta + u_i \geq 0 | X_i)$$
$$= \Pr(u_i \geq -X_i'\beta | X_i)$$
$$= 1 - F(-X_i'\beta)$$

Assume $F = \Phi$.



- Contribution of observation to likelihood function (Probit, $F = \Phi$)
$$[\Phi(X_i'\beta)]^{Y_i}[1 - \Phi(X_i'\beta)]^{1-Y_i}$$
- Contribution of observation to log-likelihood function
$$Y_i \ln[\Phi(X_i'\beta)] + (1 - Y_i)\ln[1 - \Phi(X_i'\beta)]$$
Independence over i gives
$$\mathcal{L}_N(\beta) = \sum_{i=1}^{N} \{Y_i \ln[\Phi(X_i'\beta)] + (1 - Y_i)\ln[1 - \Phi(X_i'\beta)]\}$$
- First order condition:
$$\frac{\partial \mathcal{L}_N(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} \left\{ \underbrace{\frac{Y_i}{\Phi(X_i'\hat{\beta}_{ML})}}_{m_{1i}} \phi(X_i'\hat{\beta}_{ML}) X_{ik} - \underbrace{\frac{1 - Y_i}{1 - \Phi(X_i'\hat{\beta}_{ML})}}_{m_{0i}} \phi(X_i'\hat{\beta}_{ML}) X_{ik} \right\} = 0$$
$$= \sum_{i=1}^{N} \{(m_{1i} - m_{0i})\phi(X_i'\hat{\beta}_{ML}) X_{ik}\} = 0$$
  - m_{1i} is higher if Y_i turns out to be an expected 1, and m_{i0} if Y_i turns out to be an expected 0. So we can think of the MLE as weighing (by $\phi(X_i'\hat{\beta}_{ML})X_{ik}$) the difference between expected and unexpected Y_i's.

If we don't assume Probit, then $1 - F(-X_i'\beta) \nRightarrow F(-X_i'\beta)$. So we have
$$\mathcal{L}_N(\beta) = \sum_{i=1}^{N} \left\{ \left(1 - F(-X_i'\beta)\right)^{Y_i} \left(F(-X_i'\beta)\right)^{1-Y_i} \right\}$$

➢ Can we estimate the cutoff (i.e. the point at which Y_i switches from 1 to 0)?
$$Y_i^* = X_i'\beta + u_i = \beta_0 + \tilde{X}_i'\tilde{\beta} + u_i, \qquad u_i | X_i \sim \mathcal{N}(0,1)$$
$$Y_i^* \geq \gamma \;\Rightarrow\; Y_i = 1, \qquad Y_i^* < \gamma \;\Rightarrow\; Y_i = 0$$
Then,
$$\Pr(Y_i = 1 | X_i) = \Pr(Y_i^* \geq \gamma | X_i) = \Pr\left(\beta_0 + \tilde{X}_i'\tilde{\beta} + u_i \geq \gamma \middle| X_i\right) = \Phi\left([\beta_0 - \gamma] + \tilde{X}_i'\beta\right)$$
Since the computer cannot distinguish β_0 and γ, hence it's unable to identify γ.
- If there's no β_0, then we can identify γ

❖ Marginal Effects

- Linear model: $Y_i = X_i'\beta + u_i$ with $E[u_i|X_i] = 0$

$$E[Y_i|X_i] = X_i'\beta = \sum_{k=1}^{K} x_{ik}\beta_k\,, \qquad \frac{\partial E[Y_i|X_i]}{\partial x_{ik}} = \beta_k$$

  - In the linear case, the marginal effect is always flat.

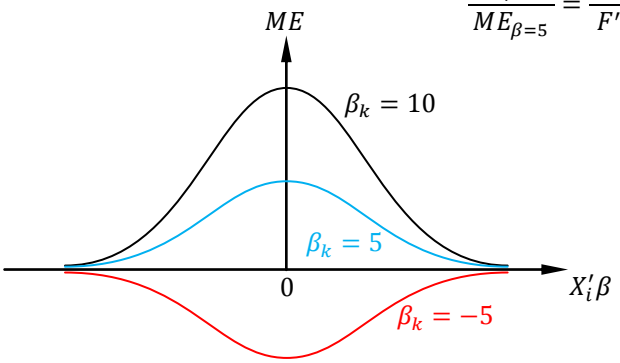
- Binary choice:

$$\begin{aligned} E[Y_i|X_i] &= \Pr(Y_i = 1|X_i)\,E[Y_i|X_i, Y_i = 1] + \Pr(Y_i = 0|X_i)\,E[Y_i|X_i, Y_i = 0] \\ &= F(X_i'\beta), \qquad \text{[assuming symmetry]} \end{aligned}$$

$$\frac{\partial E[Y_i|X_i]}{\partial x_{ik}} = \frac{\partial F(X_i'\beta)}{\partial x_{ik}} = \underbrace{F'(X_i'\beta)}_{\oplus}\beta_k$$

  - In the binary choice case…
    - the sign of β_k determines the sign of marginal effect
    - ME depends on $X_i'\beta$

$$\frac{ME_{\beta=10}}{ME_{\beta=5}} = \frac{F'(X_i'\beta)\cdot 10}{F'(X_i'\beta)\cdot 5}$$



- Logit – choice of the distribution of u_i

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u}, \qquad \begin{cases}\Lambda(u) \to 0 & as\ u \to -\infty \\ \Lambda(u) \to 1 & as\ u \to +\infty\end{cases}, \qquad \Lambda'(u) > 0\ \ \forall u \in \mathbb{R}$$

  - Recall the model we've been working on:

$$Y_i^* = X_i'\beta + u_i, \qquad u\_i|X_i \sim \Lambda(\cdot)$$
$$Y_i^* \geq 0 \ \Rightarrow\ Y_i = 1, \qquad Y_i^* < 0 \ \Rightarrow\ Y_i = 0$$

  - Some calculus:

$$\begin{aligned} \Lambda(z) &= \frac{e^z}{1+e^z} \\ \Lambda'(z) = \frac{\partial \Lambda(z)}{\partial z} &= \frac{(1+e^z)e^z - (e^z)^2}{(1+e^z)^2} \\ &= \frac{e^z}{1+e^z} - \left(\frac{e^z}{1+e^z}\right)^2 \\ &= \frac{e^z}{1+e^z}\left(1 - \frac{e^z}{1+e^z}\right) \\ &= \Lambda(z)\big(1 - \Lambda(z)\big) \end{aligned}$$

$$\frac{\partial E[Y_i|X_i]}{\partial x_{ik}} = \Lambda'(z)\beta_k = \underbrace{\Lambda(z)\big(1-\Lambda(z)\big)}_{\leq 0.25}\beta_k \leq 0.25\beta_k$$

Let $p = \Pr(Y_i = 1|X_i)$. Then,

$$p = \frac{e^{X_i'\beta}}{1+e^{X_i'\beta}}, \qquad 1-p = \frac{1}{1+e^{X_i'\beta}}$$

We thus have

- ***Odds ratio*** (ratio between two probabilities):
$$\frac{p}{1-p} = e^{X_i'\beta}$$
- ***Log odds ratio***:
$$\ln\frac{p}{1-p} = X_i'\beta$$
If x_{ik} increases by 1, log odds increase by β_k

❖ Consistency for binary choice

Theorem. 5.1. Under some conditions, $\hat{\beta}_{ML} \xrightarrow{p} \beta_0$ as $N \to \infty$ (where $\beta_0 = \arg\max_\theta Q_0(\theta)$). The conditions are

- Θ is compact
  - Just pick $\Theta = [-1e^9, 1e^9]^q$
- Q_N is measurable and continuous for all $\theta \in \Theta$
  - Just assume measurability is satisfied. For continuity, in the case of logit
$$Q_N(\beta) = \sum_{i=1}^{N} Y_i \ln(X_i'\beta) + (1-Y_i)\ln\big(1-\Lambda(X_i'\beta)\big)$$
is continuous.
- Q_N converges uniformly to Q_0, and Q_0 has a unique maximum at θ_0
  - Q_N converges pointwise to Q_0, since it's also bounded, uniform convergence is guaranteed
  - For the existence of unique global maximum, need Q_N to be globally concave
$$\frac{\partial Q_N(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} \frac{Y_i}{\Lambda(X_i'\beta)}\cdot\Lambda'(X_i'\beta)x_{ik} - \frac{1-Y_i}{1-\Lambda(X_i'\beta)}\Lambda'(X_i'\beta)x_{ik}$$
$$= \sum_{i=1}^{N} Y_i\big(1-\Lambda(X_i'\beta)\big)x_{ik} - (1-Y_i)\Lambda(X_i'\beta)x_{ik}$$
$$\frac{\partial^2 Q_N(\beta)}{(\partial\beta_k)^2} = \sum_{i=1}^{N} -Y_i\Lambda(X_i'\beta)\big(1-\Lambda(X_i'\beta)\big)x_{ik}^2 - (1-Y_i)\Lambda(X_i'\beta)\big(1-\Lambda(X_i'\beta)\big)x_{ik}^2$$
$$< 0$$

Demystifying the Criterion Function

- Examples of Criterion Functions
  - For maximum likelihood:

$$Q_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ln f(X_i'\theta)$$

  - For OLS:

$$Q_N(\theta) = -\frac{1}{N}\sum_{i=1}^{N} (Y_i - X_i'\theta)^2$$

  - For non-linear least square with additive error term ($Y_i = g(X_i, \theta) + \epsilon_i$):

$$Q_N(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \big(Y_i - g(X_i, \theta)\big)^2$$

  - For GMM:

$$Q_N(\theta) = -\frac{1}{N}\sum_{i=1}^{N} h(X_i, \theta)' W_N h(X_i, \theta)$$

Multinomial Choice

Cameron & Trivedi, Ch.15: 15.1 – 15.6, 15.8, 15.9 (optional 15.6 and 15.8)

❖ Multinomial choice as a generalization of binary choice

➢ Binary choice models posit a "latent variable" $Y_i^* = X_i'\beta + \epsilon_i$ with $\epsilon_i | X_i \sim F$

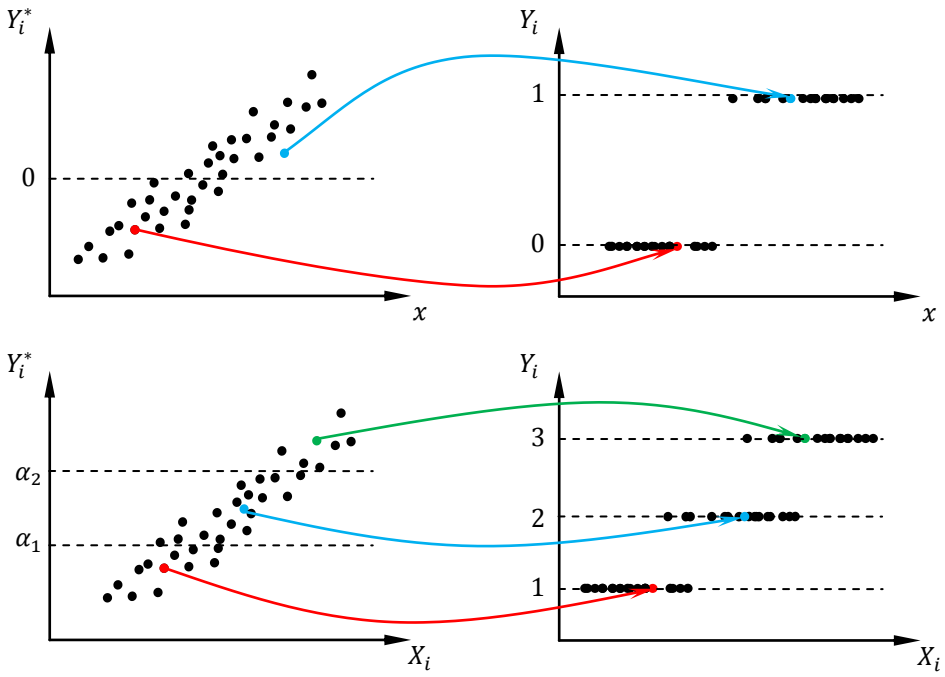
$$Y_i = \begin{cases} 1 & \text{if } Y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

▪ In multinomial choice, we generalize this to an *ordered multinomial choice*

➢ Or from another way of interpreting binary choice,

$$Y_i = 1 \;\Rightarrow\; Y_i^* \geq 0 \text{ [prefer alternative 1 over alternative 0]}$$

▪ In multinomial choice, this is generalized to *unordered multinomial choice*



❖ Model: $Y_i^* = X_i'\beta + \epsilon_i,\ \epsilon_i | X_i \sim F$, category j, $j = 1, \ldots, m$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < Y_i^* \leq \alpha_2 \\ \vdots & \vdots \\ j & \text{if } \alpha_{j-1} < Y_i^* \leq \alpha_j \\ m & \text{if } Y_i^* > \alpha_{m-1} \end{cases}$$

➢ Likelihood of i:

$$\prod_{j=1}^{m} p_{ij}^{\mathbf{1}_{\{Y_i = j\}}}$$

➢ Category 1:

$$\begin{aligned} \Pr(Y_i = 1 | X_i) &= \Pr(Y_i^* \leq \alpha_1 | X_i) \\ &= \Pr(X_i'\beta + \epsilon_i | X_i) \\ &= \Pr(\epsilon_i \leq \alpha_1 - X_i'\beta | X_i), \qquad \text{[assume normality]} \end{aligned}$$

$$= \Phi(\alpha_1 - X_i'\beta)$$

- Category j:

$$\begin{aligned}\Pr(Y_i = j|X_i) &= \Pr\left(\alpha_{j-1} < Y_i^* \le \alpha_j \middle| X_i\right)\\ &= \Pr\left(Y_i^* \le \alpha_j \middle| X_i\right) - \Pr\left(Y_i^* \le \alpha_{j-1} \middle| X_i\right)\\ &= \Phi\left(\alpha_j - X_i'\beta\right) - \Phi\left(\alpha_{j-1} - X_i'\beta\right)\end{aligned}$$

- Category m: homework.

$$\begin{aligned}\Pr(Y_i = m|X_i) &= \Pr(\alpha_{m-1} < Y_i^*|X_i)\\ &= 1 - \Pr(X_i'\beta + \epsilon_i \le \alpha_{m-1}|X_i)\\ &= 1 - \Pr(\epsilon_i \le \alpha_{m-1} - X_i'\beta|X_i)\\ &= 1 - \Phi(\alpha_{m-1} - X_i'\beta)\\ &= \Phi(X_i'\beta - \alpha_{m-1}), \qquad \text{[by symmetry]}\end{aligned}$$

❖ Suppose $m = 3$. Likelihood contribution of individual i

$$\Phi(\alpha_1 - X_i'\beta)^{\mathbf{1}_{\{Y_i=1\}}} \times \left(\Phi(\alpha_2 - X_i'\beta) - \Phi(\alpha_1 - X_i'\beta)\right)^{\mathbf{1}_{\{Y_i=2\}}} \times \left(\Phi(X_i'\beta - \alpha_{m-1})\right)^{\mathbf{1}_{\{Y_i=3\}}}$$

Parameters: β, $\alpha = (\alpha_1, \dots, \alpha_{m-1})$, and let $\theta = (\beta, \alpha)$. The log-likelihood is

$$\begin{aligned}\mathcal{L}_N(\beta, \alpha) = \frac{1}{N}\sum_{i=1}^{N} &\mathbf{1}_{\{Y_i=1\}} \ln\left(\underbrace{\Phi(\alpha_1 - X_i'\beta)}_{p_{i1}}\right) + \mathbf{1}_{\{Y_i=2\}} \ln\left(\underbrace{\Phi(\alpha_2 - X_i'\beta) - \Phi(\alpha_1 - X_i'\beta)}_{p_{i2}}\right)\\ &+ \mathbf{1}_{\{Y_i=3\}} \ln\left(\underbrace{\Phi(X_i'\beta - \alpha_2)}_{p_{i3}}\right)\\ = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3} &\mathbf{1}_{\{Y_i=j\}} \ln(p_{ij})\end{aligned}$$

❖ Interpretation of Model ($\hat{\beta}$), Linear model: $Y_i = X_i'\beta + \epsilon_i$, $E[\epsilon_i|X_i] = 0$

1) $\epsilon_i \perp X_i$

$$\beta_k = \frac{\partial E[Y_i^*|X_i]}{\partial x_{ik}}$$

As x_{ik} increases, Y_i^* changes by β_k. But this interpretation is "silly" because we don't really observe Y_i^*.

2) $\Pr(Y_i = j|X_i)$, assume $1 < j < m$ and probit,

$$\begin{aligned}\frac{\partial \Pr(Y_i = j|X_i)}{\partial x_{ik}} &= \frac{\partial\left[\Phi\left(\alpha_j - X_i'\beta\right) - \Phi\left(\alpha_{j-1} - X_i'\beta\right)\right]}{\partial x_{ik}}\\ &= \phi\left(\alpha_j - X_i'\beta\right)(-\beta_k) - \phi\left(\alpha_{j-1} - X_i'\beta\right)(-\beta_k)\\ &= \beta_k\left[\phi\left(\alpha_{j-1} - X_i'\beta\right) - \phi\left(\alpha_j - X_i'\beta\right)\right]\end{aligned}$$

In this case, the sign of β_k is not informative, because the sign of $[\cdot]$ is undetermined.

- Solution:

$$\begin{aligned}\Pr(Y_i \le j|X_i) &= \Pr(Y \in \{1, \dots, j\}|X_i)\\ &= \Pr\left(Y_i \le \alpha_j \middle| X_i\right)\\ &= \Pr\left(\epsilon \le \alpha_j - X_i'\beta \middle| X_i\right)\\ &= \Phi\left(\epsilon \le \alpha_j - X_i'\beta \middle| X_i\right)\end{aligned}$$

$$\Rightarrow \quad \frac{\partial \Pr(Y_i \le j|X_i)}{\partial x_{ik}} = -\phi\left(\alpha_j - X_i'\beta\right)\beta_k < 0$$

If β_k is positive, then the probability of Y_i being in a lower category decreases as x_{ik} goes up marginally.

❖ Unordered Multinomial. Micro-foundation to the binary choice model (aka the ***Adaptive Random Utility Model***):

$$Y_i = 1 \quad u_1 = v_1 + \epsilon_1$$
$$Y_i = 0 \quad u_0 = v_0 + \epsilon_0$$
$$Y_i = 1 \;\Leftarrow\; u_1 > u_0 \;\Leftrightarrow\; v_1 + \epsilon_1 > v_0 + \epsilon_0 \;\Leftrightarrow\; \epsilon_1 - \epsilon_0 > v_0 - v_1$$

Normalize $v_0 = 0$ and let $v_1 = X_i'\beta$, and $\epsilon = \epsilon_1 - \epsilon_0 \;\rightarrow\; \epsilon > -X_i'\beta \;\rightarrow\; \Phi(-X_i'\beta)$

Similarly, in the multinomial case, we posit the following:

$$u_j = v_j + \epsilon_j$$
$$Y_i = j \;\Leftarrow\; u_j > u_\ell \;\; \forall \ell \neq j$$
$$\Pr(Y_i = j|X_i) = \Pr\left(u_j > u_\ell \;\; \forall \ell \neq j \middle| X_i\right)$$

Assume ϵ_j independent across j (IIA), and let the distribution of ϵ_j be "Type 1 EV":

$$f\left(\epsilon_j\right) = e^{-\epsilon_j}\exp\{-e^{-\epsilon_j}\} \xrightarrow{***magic***} \Pr(Y_i = j|X_i) = \frac{e^{v_j}}{\sum_{\ell=1}^{m} e^{v_\ell}}$$

❖ Guide to 14.3 (Multinomial Logit and Conditional Logit)

$$v_j = \tilde{X}_i'\beta, \qquad [\text{multinomial logit}]$$
$$v_j = X_{ij}'\beta, \qquad [\text{conditional logit}]$$

➢ For estimation, treat everything as if it is a CL. If it happens to be ML, then construct a new set of regressors in the following way:

$$X_{i1} = \begin{bmatrix} \tilde{X}_i \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad X_{i2} = \begin{bmatrix} 0 \\ \tilde{X}_i \\ 0 \\ 0 \end{bmatrix}, \qquad X_{ij} = \begin{bmatrix} 0 \\ \vdots \\ \tilde{X}_i \\ \vdots \\ 0 \end{bmatrix}, \qquad \beta = \begin{bmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_m \end{bmatrix}$$

$$\mathcal{L}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} \underbrace{\mathbf{1}_{\{Y_i = j\}}}_{Y_i} \underbrace{\Pr(Y_i = j|X_i)}_{\beta, X}$$

➢ Marginal Effects

▪ Conditional logit

$$\Pr(Y_i = j|X_{i1}, \ldots, X_{im}) = \Pr(Y_i = j|X_i)$$

$$\frac{\partial \Pr(Y_i = j|X_i)}{\partial x_{ijk}} = \frac{\partial}{\partial x_{ijk}}\left(\frac{e^{X_{ij}'\beta}}{\sum_{\ell=1}^{m} e^{X_{i\ell}'\beta}}\right) = \frac{\left(\sum_{\ell=1}^{m} e^{X_{i\ell}'\beta}\right)\cdot \beta_k e^{X_{ij}'\beta} - \beta_k e^{X_{ij}'\beta}\cdot e^{X_{ij}'\beta}}{\left(\sum_{\ell=1}^{m} e^{X_{i\ell}'\beta}\right)^2}$$

$$\begin{bmatrix} X_{i11} \\ \vdots \\ X_{i1K} \\ X_{i21} \\ \vdots \\ X_{i2K} \\ \vdots \\ X_{imK} \end{bmatrix} \begin{matrix} \\ \\ \} \text{ Alternative 1} \\ \\ \\ \} \text{ Alternative 2} \\ \\ \\ \end{matrix}$$

$$
\begin{aligned}
&= \beta_k \underbrace{\frac{e^{x_{ij}'\beta}}{\sum_{\ell=1}^{m} e^{x_{i\ell}'\beta}}}_{p_j} - \beta_k \underbrace{\left(\frac{e^{x_{ij}'\beta}}{\sum_{\ell=1}^{m} e^{x_{i\ell}'\beta}}\right)^2}_{p_j} \\
&= \beta_k p_j - \beta_k p_j^2 \\
&= \beta_k p_j (1 - p_j)
\end{aligned}
$$

Unordered Multinomial Choice

- $j = 1, \dots, m$ alternatives, each associates with utility

$$\left.\begin{matrix} v_1 + \epsilon_1 \\ \vdots \\ v_m + \epsilon_m \end{matrix}\right\}, \qquad y = j \;\Leftrightarrow\; v_j + \epsilon_j > v_\ell + \epsilon_\ell \;\; \forall \ell \neq j$$

- Assumption on ϵ_j
  1) $\epsilon_1, \dots, \epsilon_m$ are mutually independent
  2) $f(\epsilon_j) = e^{-\epsilon_j} \exp\{e^{-\epsilon_j}\}$, Type 1 Extreme Value distribution

$$\Pr(Y = j|X) = \frac{e^{v_j}}{\sum_\ell e^{v_\ell}}$$

- Marginal Effects for Conditional Logit

$$\Pr\left(Y = j \middle| \underbrace{X_1, X_2, \dots, X_m}_{\text{regressors per alternative}}\right) = \frac{e^{X_j'\beta}}{\sum_\ell e^{X_\ell'\beta}} \coloneqq \frac{a_j}{\sum_\ell a_\ell} \coloneqq \frac{a_j}{A}$$

Regressors $k = 1, \dots, K$

$$\begin{bmatrix} X_{i11} \\ \vdots \\ X_{i1K} \\ X_{i21} \\ \vdots \\ X_{i2K} \\ \vdots \\ X_{imK} \end{bmatrix}$$

$$\frac{\partial a_j}{\partial x_{jk}} = \beta_k e^{X_j'\beta} = \beta_k a_j, \qquad \frac{\partial A}{\partial x_{jk}} = \frac{\partial a_j}{\partial x_{jk}} = \beta_k a_j$$

Let $p_j \coloneqq \Pr(Y = j|X)$. Then,

$$\begin{aligned} \frac{\partial p_j}{\partial x_{jk}} &= \frac{A \cdot \beta_k a_j - a_j \beta_k a_j}{A^2} \\ &= \beta_k \frac{a_j}{A} - \beta_k \left(\frac{a_j}{A}\right)^2 \\ &= \beta_k p_j - \beta_k p_j^2 \\ &= \beta_k p_j (1 - p_j) \end{aligned}$$

- Marginal effects depends on p_j
- Marginal effects are larger for intermediate values of p_j

For $\ell \neq k$,

$$\frac{\partial a_\ell}{\partial x_{jk}} = 0, \qquad p_\ell = \Pr(Y = \ell|X)$$

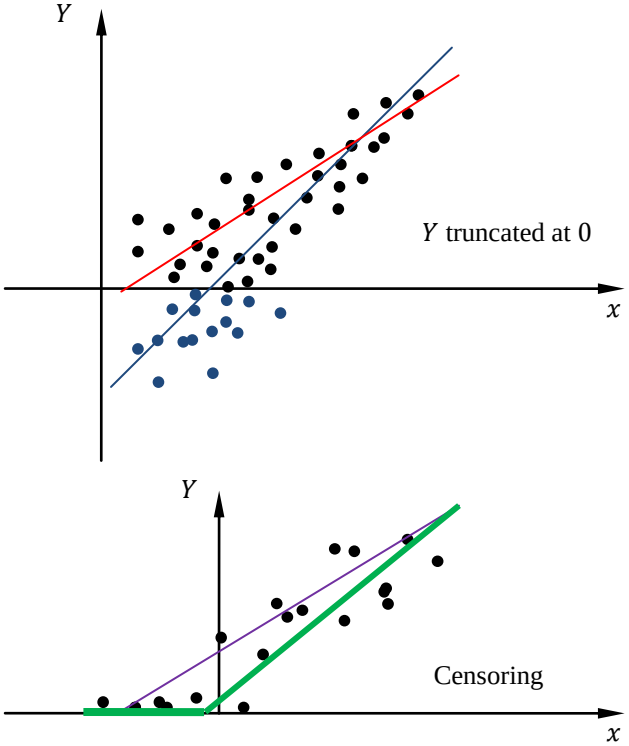
$$\frac{\partial}{\partial x_{jk}}\left(\frac{a_\ell}{A}\right) = \frac{A \cdot 0 - a_\ell \cdot \beta_k a_j}{A^2}$$
$$= -\beta_k \frac{a_\ell}{A}\frac{a_j}{A}$$
$$= -\beta_k p_\ell p_j$$
$$< 0, \qquad if \;\; \beta_k > 0$$

- What's the probability of being in another category?
  - If x_{jk} increases, p_j

Tobit + Selection

Ch. 16.1 – 16.6
Skip 16.2.4, 16.2.5, 16.3.6, 16.6.7 (and don't worry about LEF)
Plus whatever we do out of 16.10

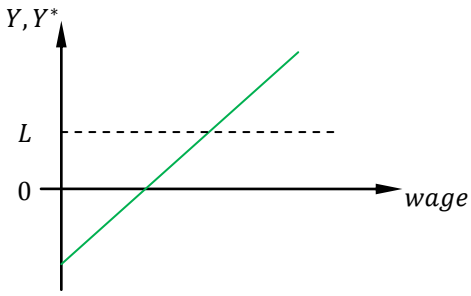
❖ Censoring and Truncation



- Censoring and truncation imply non-linearity in the model.
- Usually we start with some conditional mean, and by omitting observations below zero (truncation or censoring), the regression we get are going to have a higher condition mean for Y close to zero.

❖ A latent-variable model

- Y is the observed dependent variable (e.g. # of hours worked)
- Y^* is unobserved dependent variable (e.g. desired # of hours worked)
  - L is censoring or truncation cutoff



- Censoring: $Y = \begin{cases} Y^* & \text{if } Y^* > L \\ L & \text{if } Y^* \leq L \end{cases}$
- Truncation: $Y = \begin{cases} Y^* & \text{if } Y^* > L \\ \text{unobserved} & \text{if } Y^* \leq L \end{cases}$
- $f^*(Y^*|X)$ is the conditional pdf of Y^*
  - Censoring: $f(Y|X)$ conditional distribution of Y (observed dependent variable)

$$f(Y|X) = \begin{cases} f^*(Y^*|X) & \text{if } Y > L \\ F^*(L|X) & \text{if } Y = L \end{cases}$$

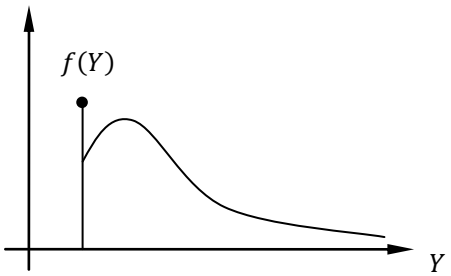
Introduce another variable:

$$d = \begin{cases} 1 & \text{if } Y > L \\ 0 & \text{if } Y = L \end{cases}$$

Then,

$$f(Y|X) = f^*(Y|X)^d F^*(L|X)^{1-d}$$

$\big((Y_1, X_1), \dots, (Y_N, X_N)\big)$ are iid copies of (Y, X)



$$L_N(\theta) = \prod_{i=1}^{N} f^*(Y_i|X_i, \theta)^{d_i} F^*(L|X_i)^{1-d_i}$$

$$\hat{\theta} = \arg\max \underbrace{\sum_{i=1}^{N} (d_i \ln f^*(Y_i|X_i, \theta) + (1 - d_i) \ln F^*(L|X_i))}_{\mathcal{L}_N(\theta)}$$

- Truncation: note the difference between $f(Y)$ and $f^*(Y)$

$$\begin{aligned} f(y|X) &= f^*(Y|X, Y > L) \\ &= \frac{P(Y, Y > L|X)}{P(Y > L|X)}, \qquad \left[\text{recall } P(A|B) = \frac{P(A \cap B)}{P(B)}\right] \\ &= \frac{f^*(y)}{1 - F^*(L|X)} \end{aligned}$$

Likelihood function:

$$L_N(\theta) = \prod_{i=1}^{N} \frac{f^*(y_i|X_i, \theta)}{1 - F^*(L|X_i, \theta)}$$

$$\hat{\theta} = \arg\max \sum_{i=1}^{N} \ln f^*(Y_i|X_i, \theta) - \ln\big(1 - F^*(L|X_i, \theta)\big)$$

❖ Tobit

- $Y^* = X'\beta + \epsilon$
- $\epsilon|X \sim \mathcal{N}(0, \sigma^2)$
- Marginal effects:
  - For all observations above the cutoff: simply change by β
  - For all below and not close to cutoff: no effect
  - For all below but close to cutoff: probability that it'll jump to above cutoff + β

- Conditional means (suppose $L = 0$)
  - $E[Y^*|X] = X'\beta$
  - Truncated mean:

$$\begin{aligned} E[Y|X, Y > 0] &= E[Y^*|X, Y^* > 0] \\ &= E[X'\beta + \epsilon|X, Y > 0] \\ &= X'\beta + E[\epsilon|X, \epsilon > -X'\beta] \\ &> E[Y^*|X] \end{aligned}$$

    - Only look at people who are already working

- Censored mean:

$$\begin{aligned} E[Y|X] &= P(Y>0|X)E[Y|X,Y>0] + P(Y=0)\underbrace{E[Y|X,Y=0]}_{=0} \\ &= P(X'\beta+\epsilon>0|X)\left(X'\beta + E\left[\epsilon\middle|X,\epsilon>-X'\beta\right]\right) \\ &= P(\epsilon>-X'\beta|X)\left(X'\beta + E\left[\epsilon\middle|X,\epsilon>-X'\beta\right]\right) \\ &= \left(1-F\left(-\frac{X'\beta}{\sigma}\right)\right)\left(X'\beta + E\left[\epsilon\middle|X,\epsilon>-X'\beta\right]\right) \end{aligned}$$

  - Look at the whole population, both working and not working

- Use the assumption $\epsilon|X \sim \mathcal{N}(0,\sigma^2)$. Define $u \sim \mathcal{N}(0,1)$ such that

$$\begin{aligned} E[u|u>c] &= \int_c^\infty u\cdot f(u|u>c)\,du \\ &= \int_c^\infty u\cdot\frac{\phi(u)}{1-\Phi(c)}du \\ &= \frac{1}{1-\Phi(c)}\int_c^\infty u\cdot\phi(u)du \\ &= \frac{1}{1-\Phi(c)}\int_c^\infty -\frac{\partial\phi(u)}{\partial u}du \\ &= \frac{1}{1-\Phi(c)}[-\phi(u)]_c^\infty \\ &= \frac{\phi(c)}{1-\Phi(c)} \end{aligned}$$

  where the 4th equality follows from

$$\phi(u) = \frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{u^2}{2}\right\} \;\Rightarrow\; \frac{\partial\phi(u)}{\partial u} = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2}\right\}(-u) = -u\cdot\phi(u)$$

  Using symmetry of Φ:

$$E[u|u>c] = \frac{\phi(c)}{1-\Phi(c)}, \qquad E[u|u>-c] = \frac{\phi(c)}{\Phi(c)} \equiv \lambda(c) = \text{inverse Mills ratio}$$

  Consider a r.v. $v \sim \mathcal{N}(0,\sigma^2)$

$$E[v|v>-c] = E\left[\sigma\cdot\frac{v}{\sigma}\middle|v>-c\right] = \sigma\cdot E\left[\frac{v}{\sigma}\middle|\frac{v}{\sigma}>-\frac{c}{\sigma}\right] = \sigma\lambda\left(\frac{c}{\sigma}\right)$$

  Let v be ϵ, and let c be $X'\beta$

$$E[\epsilon|X,\epsilon>-X'\beta] = \sigma\lambda\left(\frac{X'\beta}{\sigma}\right) = \lambda(w), \qquad w \equiv \frac{X'\beta}{\sigma}$$

  Therefore,

  - $E[Y^*|X] = X'\beta$
  - $E[Y|X,Y>0] = X'\beta + \sigma\lambda(w)$, where $w \equiv X'\beta/\sigma$
  - $E[Y|X] = \Phi(w)\cdot\left(X'\beta+\sigma\lambda(w)\right)$

Marginal effects:

$$\frac{\partial \lambda(z)}{\partial z} = \frac{\Phi(z)\cdot\left(-z\phi(z)\right) - \phi(z)^2}{\Phi(z)^2}$$
$$= -\frac{z\phi(z)}{\Phi(z)} - \left(\frac{\phi(z)}{\Phi(z)}\right)^2$$
$$= z\lambda(z) - \lambda(z)^2$$

$$\frac{\partial E[Y|X, Y>0]}{\partial x_k} = \frac{\partial}{\partial x_k}[X'\beta + \sigma\lambda(w)]$$
$$= \frac{\partial X'\beta}{\partial x_k} + \frac{\partial \sigma\lambda(w)}{\partial x_k}$$
$$= \beta_k + \sigma\cdot\frac{\partial\lambda(w)}{\partial w}\frac{\partial w}{\partial x_k}$$
$$= \beta_k + \sigma[-w\lambda(w) - \lambda(w)^2]\cdot\frac{\beta_k}{\sigma}$$
$$= \beta_k(1 - w\lambda(w) - \lambda(w)^2)$$

- Homework:

$$\frac{\partial E[Y|X]}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\Phi(w)\left(X'\beta + \sigma\lambda(w)\right)\right]$$
$$= \phi(w)\frac{\partial w}{\partial x_k}\left(X'\beta + \sigma\lambda(w)\right) + \Phi(w)[\beta_k(1 - w\lambda(w) - \lambda(w)^2)]$$
$$= \phi(w)\cdot\frac{\beta_k}{\sigma}\left(X'\beta + \sigma\lambda(w)\right) + \Phi(w)\beta_k(1 - w\lambda(w) - \lambda(w)^2)$$
$$= \beta_k\Phi(w)[\lambda(w)w + \lambda(w)^2 + 1 - w\lambda(w) - \lambda(w)^2]$$
$$= \beta_k\Phi(w)$$

Tobit Models (cont'd)

❖ $Y_1^* = X_1'\beta + \epsilon_1$ where $\epsilon_1|X_1 \sim \mathcal{N}(0, \sigma^2)$

$$Y_1 = \begin{cases} Y_1^* & \text{if } Y_1^* \geq 0 \\ 0 & \text{if } Y_1^* \leq 0 \end{cases}$$

❖ Selection (with censoring, so we see that people are below the cutoff)

$$Y_1^* = X_1'\beta_1 + \epsilon_1, \qquad Y_1 = \begin{cases} 1 & \text{if } Y_1^* \geq 0 \\ 0 & \text{if } Y_1^* < 0 \end{cases}$$

$$Y_2^* = X_2'\beta_2 + \epsilon_2, \qquad Y_2 = \begin{cases} Y_2^* & \text{if } Y_1 = 1 \\ \text{unobserved} & \text{if } Y_1 = 0 \end{cases}$$

- This is called the ***Tobit-2***
- Distribution of the error terms

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \Big| X \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right), \qquad X = union(X_1, X_2)$$

- What do we observe? (working towards $\mathcal{L}_N(\theta)$)
  - Either $\begin{cases} Y_1 = 0 \\ Y_2 \text{ nothing} \end{cases}$ with $P(Y_1 = 0|X)$
  - Or $\begin{cases} Y_1 = 1 \\ Y_2 \text{ is something} \end{cases}$ with $P(Y_1 = 1|X)$
  - So the likelihood contribution is

$$\underbrace{\left(1 - \Phi(X_1'\beta_1)\right)}_{P(Y_1=0|X)}^{1-Y_1} \underbrace{\left[P(Y_2|Y_1 = 1)P(Y_1 = 1|X)\right]}_{P(Y_2|Y_1=1)\Phi(X_1'\beta_1)}^{Y_1}$$

  - Suppose the two error terms are not correlated: $\epsilon_1 \perp \epsilon_2|X$

$$\Rightarrow \; Y_1^* \perp Y_2^*|X \;\text{ and }\; Y_1 \perp Y_2|X$$
$$\Rightarrow \; P(Y_1, Y_2) = P(Y_2|X, Y_1 = 1)P(Y_1 = 1|X) = P(Y_2|X)P(Y_1 = 1|X)$$

  - Likelihood function

$$\mathcal{L}_N(\beta_1, \beta_2) = \sum_{i=1}^{N} (1 - Y_{1i}) \ln\left(1 - \Phi(X_{1i}'\beta_1)\right) + Y_{1i} \ln \underbrace{\left(\Phi(X_{1i}'\beta_1)\right)}_{P(Y_1=1)} + \sum_{i=1}^{N} Y_{1i} \underbrace{g(\beta_2)}_{P(Y_2)}$$

    - Model Y_1 involves β_1
    - Model Y_2 involves β_2
    - Can estimate them jointly, or estimate β_1 using Probit, and β_2 using OLS

- Tobit-2

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \Big| X \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right) \;\Rightarrow\; \epsilon_2|\epsilon_1, X \sim \mathcal{N}(\sigma_{12}\epsilon_1, V)$$
$$\epsilon_2 = \sigma_{12}\epsilon_1 + \zeta, \qquad \epsilon_1 \perp \zeta|X$$

  - Conditional means: $E[Y_2^*|X] = X_2'\beta_2$

$$
\begin{aligned}
E[Y_2|X, Y_1 = 1] &= E\left[\underbrace{X_2'\beta_2 + \epsilon_2}_{Y_2^*} \middle| X, Y_1^* > 0\right] \\
&= X_2'\beta_2 + E[\epsilon_2|X, \epsilon_1 > -X_1'\beta_1] \\
&= X_2'\beta_2 + E[\sigma_{12}\epsilon_1 + \zeta|X, \epsilon_1 > -X_1'\beta_1] \\
&= X_2'\beta_2 + \sigma_{12}E[\epsilon_1|X, \epsilon_1 > -X_1'\beta_1] + \underbrace{E[\zeta|X, \epsilon_1 > -X_1'\beta_1]}_{=0} \\
&= X_2'\beta_2 + \underbrace{\sigma_{12}}_{=\sigma_2}\lambda(X_1'\beta_1)
\end{aligned}
$$

- Two stage (HECKIT)
  1) Do a Probit for $Y_1^* = X_1'\beta_1 + \epsilon_1 \rightarrow$ get $\hat{\beta}_1$
  2) Make a new variable $\lambda(X_1'\hat{\beta}_1)$
  3) Do OLS for Y_2^* on X_2 and $\lambda(X_1'\hat{\beta}_1) \rightarrow$ get $\hat{\beta}_2$ and σ_{12}
  4) $H_0 : \sigma_{12} = 0$
  5) Get the correct standard errors $\rightarrow$ usually higher than OLS ones

Stata Tricks

- Stata commands
  - `-list in 1/10-` after the `-use-` to show observations 1 to 10
  - `-list if price>10-` shows observations where price is greater than 10
  - `-list c1-wage in 1/10-` shows first 10 observations for variables "c1" to "wage"
  - `-insheet varname1 varname2 ... using dataset.csv-` loads dataset and assign names to each column of data
  - `-infix varname1 1-2 varname2 3-4 ... using dataset.fix-` loads fixed format dataset and assign names to designated columns of data.
    - In this example, columns 1 and 2 belong to variable 1, 3 and 4 to variable 2, etc.
  - There are different ways of coding missing data, and Stata treats them as `.`, `.a`, `.b`, etc. If you want to replace all these missing values use `-replace varname = 500 if varname>=.-`

- MLE in Stata
  - All we need is just the likelihood function, then we get the estimator and the standard errors for free
    - Estimator is the maximand of the likelihood function
    - Standard errors come from the Jacobian of the likelihood function
  - OLS: $y = X\beta + \epsilon$ with $\epsilon|X \sim \mathcal{N}(0, \sigma^2)$

$$f(\epsilon|X) = f(y - X\beta|X) \Rightarrow f\left(\frac{\epsilon}{\sigma}\middle|X\right) \sim \mathcal{N}(0,1)$$

$$\Rightarrow f\left(\frac{\epsilon}{\sigma}\middle|X\right) = \frac{1}{\sigma}\phi\left(\frac{y - X\beta}{\sigma}\right)$$

$$\Rightarrow (\hat{\beta}, \hat{\sigma})_{ML} = \arg\max\left(\sum_{i=1}^{N} \ln\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right)\right)$$

  - Probit: $y^* = X\beta + \epsilon$, and $y = 1 \Leftrightarrow y^* > 0$

$$\Pr(y = 1|X) = \Pr(y^* > 0|X) = \Pr(X\beta + \epsilon > 0|X) = \Pr(\epsilon > -X\beta|X) = 1 - \Phi(-X\beta)$$
$$= \Phi(X\beta)$$

Midterm Answer

❖ Question 4 (b)

$$E[z] = 0 = \Pr(z > c)\, E[z|z > c] + \Pr(z < c)\, E[z|z < c]$$
$$= [1 - \Phi(c)] \frac{\phi(c)}{1 - \Phi(c)} + \Phi(c) \cdot X$$

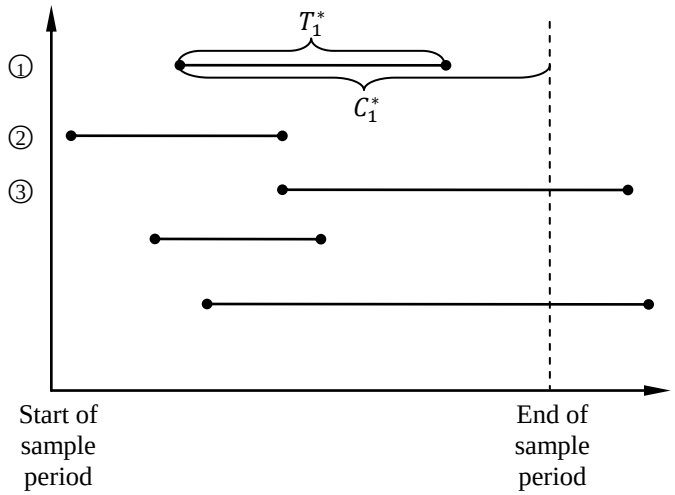
where

$$X = -\frac{1 - \Phi(c)}{\Phi(c)} \cdot \frac{\phi(c)}{1 - \Phi(c)} = -\frac{\phi(c)}{\Phi(c)}$$

Duration

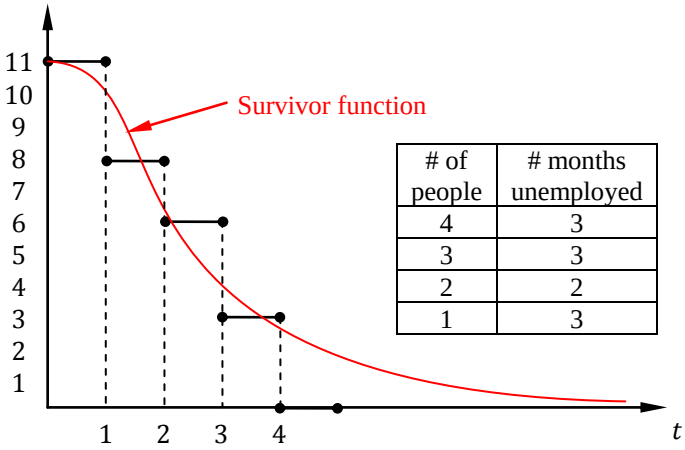
C&T Ch 17.1 – 17.4, 17.6 – 17.

❖ A picture to keep in mind



- There is truncation on the left, and censoring on the right
- T is the (true) duration of a spell
- Observed duration: $t_i = \min\{T_i^*, C_i^*\}$,
- Whether of not there is censoring: $\delta_i = \mathbf{1}_{\{T_i^* < C_i^*\}}$,
- Individual characteristics: x_i

❖ ***Survivor function***: "how many are left" $\Pr(T \geq t)$



# of people	# months unemployed
4	3
3	3
2	2
1	3

- Define a r.v. T (duration), positive, with continuous distribution on $\mathbb{R}_+$
  - cdf: $F(t) = \Pr(T \leq t)$
  - pdf: $f(t) = \partial F(t)/\partial t$
- The ***survivor function***:

$$S(t) = 1 - F(t) = \Pr(T > t)$$

- The ***hazard function***:

$$\lambda(t) \coloneqq \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

  - This is the "instantaneous probability of leaving a state conditional on survival to

time t."

- This function answers the question "What is the probability that a person's unemployment spell ends now, given he has been unemployed for T periods?"

➢ ***Cumulative hazard function***:

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

1) Survivor function:

$$S(t) = 1 - F(t)$$

2) Expectation of duration T:

$$\begin{aligned} \mathrm{E}(T) &= \int_0^\infty uf(u)du \\ &= uF(u)|_0^\infty - \int_0^\infty F(u)du \\ &= \int_0^\infty \big(1 - F(u)\big)du \\ &= \int_0^\infty S(u)du \end{aligned}$$

➢ Thus, the expectation of duration is the area under the survivor curve.

3) Hazard function:

$$\begin{aligned} \lambda(t) &= -\frac{\partial \ln S(t)}{\partial t} \\ &= -\frac{1}{S(t)}\left(\frac{\partial\big(1 - F(t)\big)}{\partial t}\right) \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

4) Alternative way of expressing survivor function:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(w)dw\right) \\ &= \exp\left(\int_0^t \frac{\partial \ln S(u)}{\partial u}du\right) \\ &= \exp(\ln S(u)|_0^t) \\ &= \exp\left(\ln S(t) - \overbrace{\ln \underbrace{S(0)}_{=1}}^{=0}\right) \\ &= \exp(\ln S(t)) \\ &= S(t) \end{aligned}$$

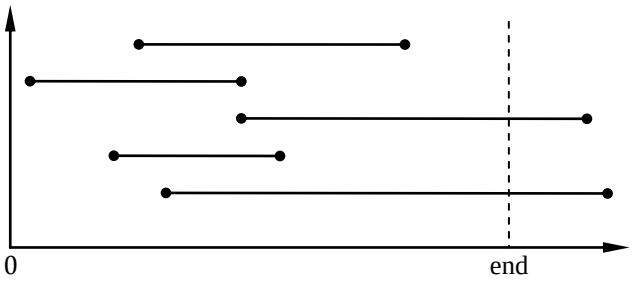
5) Cumulative hazard function:

$$\begin{aligned} \Lambda(t) &= -\ln S(t) \\ &= \int_0^t \lambda(u)du \end{aligned}$$

Thus,

$$S(t) = \exp\big(-\Lambda(t)\big) \;\Leftrightarrow\; \ln S(t) = -\Lambda(t)$$

❖ Working towards a MLE for the duration model



- $t_i = \min\{T_i^*, C_i^*\}$, $\delta_i = \mathbf{1}_{\{T_i^* < C_i^*\}}$, x_i
- $\lambda_i(t) = \lambda(t|x_i, \theta_0)$
- $\Lambda_i(t) = \Lambda(t|x_i, \theta_0)$
- Suppose $\lambda(t) = \exp(x_i'\theta) > 0$
  - $\lambda = f(t)/S(t)$

- Event ended before censoring ($t_i = T_i^*$), then $\delta_i = 1$. The probability is $[f_i(t_i)]^{\delta_i}$
- Observation censored ($t_i = C_i^*$), then $\delta_i = 0$. The probability is $[S_i(t_i)]^{1-\delta_i}$
- Likelihood contribution:

$$L(\theta) = \prod_{i=1}^{N} f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i}$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \{\delta_i \ln f_i(t_i) + (1-\delta_i) \ln S_i(t_i)\}$$

$$= \sum_{i=1}^{N} \{\delta_i (\ln \lambda_i(t_i) + \ln S_i(t_i)) + (1-\delta_i) \ln S_i(t_i)\}$$

$$= \sum_{t=1}^{N} \{\delta_i \ln \lambda_i(t_i) + \ln S_i(t_i)\}$$

$$= \sum_{t=1}^{N} \{\delta_i \ln \lambda_i(t_i) - \Lambda_i(t_i)\}$$

$$= \sum_{i=1}^{N} \{\delta_i \ln \lambda(t_i|x_i, \theta) - \Lambda(t_i|x_i\theta)\}$$

$$= \sum_{i=1}^{N} \{\delta_i x_i'\theta - t_i \exp(x_i'\theta)\}$$

- In the second equality of $\mathcal{L}$, we used the following

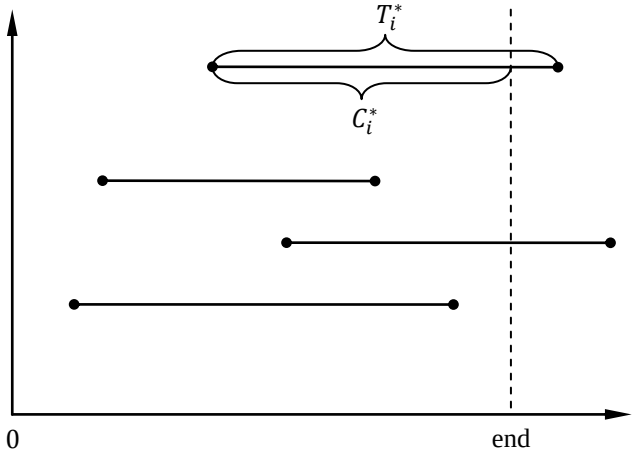
$$\lambda = \frac{f}{S} \;\Rightarrow\; f = \lambda S \;\Rightarrow \ln f = \ln(\lambda S) = \ln \lambda + \ln S$$

- With exponential distribution, i.e.

$$\lambda(t_i|x_i, \theta) = \exp(x_i'\theta), \qquad \Lambda(t_i|x_i, \theta) = \int_0^t \lambda(s_i|x_i, \theta) ds_i = t \exp(x_i'\theta)$$

Duration Model (cont'd)

❖ Review of duration model



- Observe T_i^* if $\delta_i = 1$, and observe C_i^* if $\delta_i = 0$ (and we know that $T_i^* \geq C_i^*$). Let

$$t_i = \begin{cases} T_i^* & \text{if } \delta_i = 1 \\ C_i^* & \text{if } \delta_i = 0 \end{cases}$$

- $f_i(t_i)$ is the pdf of t_i, which potentially differs across observation i
- Survivor function $S_i(t_i) = S(t_i|x_i, \beta)$
- Log-likelihood function

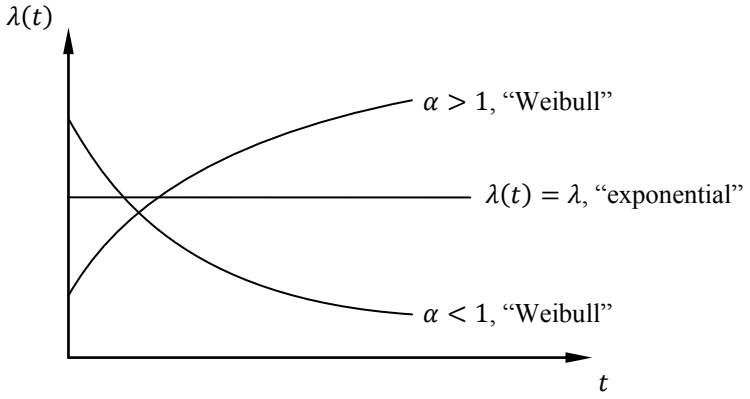
$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \{\delta_i \ln f_i(t_i) + (1 - \delta_i) \ln S_i(t_i)\}$$

- Hazard function

$$\lambda(t) \coloneqq \lim_{\Delta t \to 0} \frac{\Pr(t \leq T_i^* < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

❖ Proportional hazard model

- $\lambda(t) = \lambda$, (i.e. constant hazard rate), where $\lambda = \exp\{x'\beta\}$
- $\Lambda(t) = \int_0^t \lambda(s)ds = t\lambda$
- $\ln S(t) = -\Lambda(t) = -t\lambda \;\Rightarrow\; S(t) = \exp\{-t\lambda\} \;\Rightarrow\; F(t) = 1 - S(t) = 1 - \exp\{-t\lambda\}$



- Suppose the hazard rate is not constant. Then we'd use the ***Weibull distribution***.

$$\lambda(t) = \lambda\alpha t^{\alpha-1}$$

- Regressors: $\lambda = \exp(x'\beta)$, $\alpha = \alpha$

$$\begin{aligned}\lambda(t) &= \lambda \cdot \alpha t^{\alpha-1} \\ &= \exp\{x'\beta\} \cdot \alpha t^{\alpha-1} \\ &= \phi(x,\beta) \cdot \lambda_0(t,\alpha)\end{aligned}$$

  - The shape of the hazard function is determined by $\lambda_0(\cdot)$ (or α), the ***baseline harzard function***.
  - If you assume the PH model, then you don't need to know what λ_0 is. But proportionality is restrictive.
- ***Generalized Weibull*** (which allows for non-monotone hazard functions)

$$\lambda\alpha t^{\alpha-1}[S(t)]^{-\mu}, \qquad \alpha > 1, \ \ \mu < 0$$

  Look at the log of G-Weibull

$$\underbrace{(\alpha-1)\ln(\lambda\alpha t)}_{\text{increasing}} - \underbrace{\mu \ln\big(S(t)\big)}_{S(\cdot)\text{ decreasing in } t}$$

  Hence we can get a U-shape hazard function.

❖ Answer to Leanna's question

- Suppose data is uncensored, i.e. $\delta_i = 1$, and t_i is the length of complete spell
- Likelihood contribution

$$[f(t_i)]^{\delta_i}$$

- There are <u>two</u> random variables: (t_i, δ_i)

$$\Pr(t_i = t, \delta_i = 1) = \Pr(t_i = t|\delta_i = 1)\Pr(\delta_i = 1)$$

- Suppose t is independent of δ.

$$\begin{aligned}t \perp \delta \ &\Rightarrow \ \Pr(t_i = t|\delta_i = 1)\Pr(\delta_i = 1) = \Pr(t_i = t)\Pr(\delta_i = 1) \\ &\Rightarrow \ \ln\Pr(t_i = t|\theta_1) + \ln\Pr(\delta_i = 1|\theta_2)\end{aligned}$$

- Uninformativeness: θ_1 and θ_2 are non-overlapping

Transition (Duration) Data

- Data on durations
  - T^* is the length of completed spell
  - C^* is censoring time
  - Observe $t = \min\{T^*, C^*\}$ and $\delta = \mathbf{1}_{\{T<C\}}$, and covariates x
  - Assume model (distribution) for T^*
    - Either F, f, S, λ, or Λ will work.
  - Pick $\lambda(t) = \lambda(t|x, \beta)$
  - Collect iid data on $(t_i, \delta_i, x_i)_{i=1}^N$

- Write down likelihood function [note that we have censoring]
  - Ignore x_i (the regressors)

$$f_{t,\delta}(t_i, \delta_i) = \begin{cases} f_{t|\delta}(t_i|\delta_i = 1)\Pr(\delta_i = 1) & \text{if } \delta_i = 1 \\ f_{t|\delta}(t_i|\delta_i = 0)\Pr(\delta_i = 0) & \text{if } \delta_i = 0 \end{cases}$$

  - Assume t, δ are independent (so that T, C are independent). Then $f_{t|\delta}(t, \delta) = f_t(t)$. The log-likelihood is

$$\begin{aligned} \mathcal{L}(\cdot) &= \sum_{i=1}^N \left\{\delta_i(\ln f_t(t_i) + \ln \Pr(\delta_i = 1)) + (1-\delta_i)(\ln f_t(t_i) + \ln(1 - \Pr(\delta_i = 1)))\right\} \\ &= \underbrace{\sum_{i=1}^N \left\{\delta_i \ln f(t_i) + (1-\delta_i)\ln S(t_i)\right\}}_{\text{depends on } \theta_1 \text{ only}} \\ &\quad + \underbrace{\sum_{i=1}^N \left\{\delta_i \ln \Pr(\delta_i = 1) + (1-\delta_i)\ln(1 - \Pr(\delta_i = 1))\right\}}_{\text{depends on } \theta_2 \text{ only}} \end{aligned}$$

    - Note that $f \neq f_t$.
  - Assume

$$f(t) = f(t|\theta_1), \qquad \Pr(\delta = 1) = \Pr(\delta = 1|\theta_2)$$

    where θ_1 and θ_2 do not overlap (i.e. none of their elements coincide).
  - We're interested in θ_1. The $\hat{\theta}_1$ that maximizes $\mathcal{L}(\cdot)$ also maximizes

$$\sum_{i=1}^N \delta_i \ln f(t_i) + (1-\delta_i)\ln S(t_i)$$

    - This is the ***conditional MLE* (*CMLE*)**: we estimate θ_1 conditional on that we know the distribution of δ.

  - We've shown that

$$\begin{aligned} \sum_{i=1}^N \delta_i \ln f(t_i) + (1-\delta_i)\ln S(t_i) &= \sum_{i=1}^N \{\delta_i \ln \lambda(t_i) - \Lambda(t_i)\} \\ &= \sum_{i=1}^N \{\delta_i \ln \lambda(t_i|x_i, \beta) - \Lambda(t_i|x_i, \beta)\} \end{aligned}$$

- Popular family of models is the family of ***Proportional Hazard* (*PH*)** models.

❖ PH model (when the hazard function can be written as two components, one depends on t and the other depends on x):

$$\lambda(t|x,\beta) = \lambda_0(t|\alpha)\phi(x,\beta)$$

- With exponential: $\lambda_0(t|\alpha) = 0.01$, $\phi(x,\beta) = \exp(x'\beta) > 0$, then the hazard (rate) $\lambda(t|x,\beta)$ is constant at 0.01, given the set of regressors x. If we change the regressors, the hazard rate will shift either up or down, but still independent of t.
- The model is called "proportional" in the sense that
  1. x can only shift λ, but not its shape
  2. λ_0 is constant (in the case of exponential distribution)
- These two are also what make the model too restrictive
  - With Weibull, we can make the baseline hazard (λ_0) go up or down
  - With generalized Weibull, we can have a U-shaped λ_0
  - But these generalizations still cannot address 1., the shape is still unaffected by the regressors

- ***Cox PH model***.
  - Keep objection 1. (still a proportional model)
  - Get rid of 2 completely
    - $\phi(x,\beta)$ is parametric, but λ_0 is not restricted (we can estimate β without imposing any structure on λ_0
    - But still, even with this level of flexibility, λ_0 is still a scale factor
  - Marginal effects (in the case of exponential model $\phi(x,\beta) = \exp(x'\beta)$)

$$\frac{\partial\lambda(t|x,\beta)}{\partial x_k} = \lambda_0(t)\beta_k \exp(x'\beta) = \beta_k\lambda(t|x,\beta)$$

In general,

$$\frac{\partial\lambda(t|x,\beta)}{\partial x_k} = \lambda_0(t)\cdot\frac{\partial\phi(x,\beta)}{\partial x_k} = \lambda(t|x,\beta)\cdot\frac{\partial\phi(x,\beta)/\partial x_k}{\phi(x,\beta)}$$

❖ Estimation of the PH model.

- Think of it as discrete time: $t_1 < t_2 < \cdots < t_k$ ("failure time") and $N \geq k$
  - Order observations according the length of the durations, from smallest to largest
- Risk set: $R(t_j) = \{\ell : t_\ell \geq t_j\}$ is the set of observation that didn't "die" yet.
- "Death" set: $D(t_j) = \{\ell : t_\ell = t_j\}$, with $d_j = \#\{D(t_j)\}$
- hazard $= \frac{\#D(t_j)}{\#R(t_j)}$
- Ties: $d_j > 1$, i.e. more than one spell ends at t_j

- Assume no ties. The probability of the jth observation dying at time j is

$$\Pr\left(T_j = t_j \middle| R(t_j)\right) = \frac{\Pr(T_j = t_j | T_j \geq t_j)}{\sum_{\ell\in R(t_j)}\Pr(T_\ell = t_j | T_\ell \geq t_j)} = \frac{\lambda(t_j|x_j,\beta)}{\sum_{\ell\in R(t_j)}\lambda(t_j|x_\ell,\beta)}$$

$$= \frac{\lambda_0(t_j)\phi(x_j,\beta)}{\sum_{\ell\in R(t_j)}\lambda_0(t_j)\phi(x_\ell,\beta)} = \frac{\phi(x_j,\beta)}{\sum_{\ell\in R(t_j)}\phi(x_\ell,\beta)}$$

- If we have ties, then

$$\Pr\left(T_j = j \middle| j \in R(t_j)\right) \simeq \frac{\prod_{m\in D(t_j)} \phi(x_m, \beta)}{\left[\sum_{\ell\in R(t_j)} \phi(x_\ell, \beta)\right]^{d_j}}$$

- The "likelihood function" is

$$L_p(\beta) = \prod_{j=1}^{k} \Pr\left(T_j = j \middle| j \in R(t_j)\right) = \prod_{j=1}^{k} \frac{\prod_{m\in D(t_j)} \phi(x_m, \beta)}{\left[\sum_{\ell\in R(t_j)} \phi(x_\ell, \beta)\right]^{d_j}}$$

  - Choose β to maximize L_p
  - Note that λ_0 is not in the likelihood function
  - But we have results showing that $\hat{\beta}$ is a consistent estimator of this model

❖ We are often not interested in $\lambda_0(t)$. But how do we estimate $\lambda_0(t)$ if we ever get interested?

$$\lambda_0(t_j) = 1 - \alpha_j \;\Rightarrow\; \lambda(t|x,\beta) = \underbrace{(1-\alpha_j)}_{\text{Cox: hazard free}} \underbrace{\phi(x,\beta)}_{\text{PH: fully specified}}$$

$$\Rightarrow\; S(t|x,\beta) = \underbrace{[S_0(t)]}_{\text{depends on } \alpha_j}^{\phi(x,\beta)}$$

Likelihood:

- Think of observation j, dies at t_j, and its likelihood contribution is

$$S(t_j|x_i,\beta) - S(t_{j+1}|x_i,\beta)$$

This leads to a likelihood function

$$\mathcal{L}(\alpha_1, \ldots, \alpha_k, \beta) = \sum_{i=1}^{N} \ln(\text{likelihood contribution})$$

Then, use the knowledge of $\hat{\beta}$ to back out the α's by estimating

$$\frac{\partial \mathcal{L}(\alpha, \hat{\beta})}{\partial \alpha} = 0$$

GMM Review

C&T Ch.6.1 – 6.6

❖ Newey and McFadden (1994), Figure 1

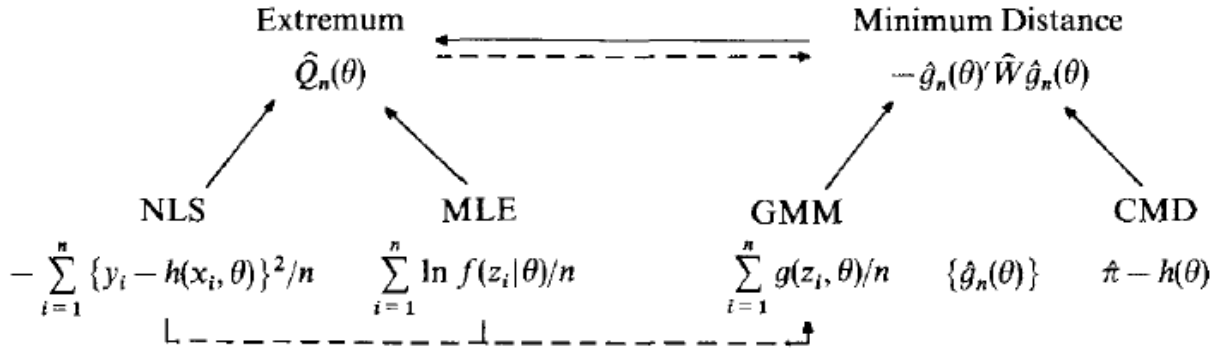


Figure 1.

- MLE is not consistent if model specification is wrong – if $\epsilon \nsim \mathcal{N}$, and you used probit, then $\hat{\beta}$ is not consistent
- GMM is a semi-parametric estimator, and OLS is an example of this estimator
  - We say that $\mathrm{E}[\epsilon] = 0$, we do not need to assume $\epsilon \sim \mathcal{N}$ for OLS (hence semi-parametric)
- Solid line in the figure → estimator is included in the class.
  - Every estimator is an extremum estimator. Proposition 5.1 applies to all these estimators (note that CMD stands for "conditional minimum distance")
- Broken line in the figure → if maximum occurs as a solution to the FOC (interior + differentiability), then we have these connections
  - GMM covers NLS and MLE; so GMM is a big class of estimators.

❖ Generalized Method of Moments (GMM)

1) We have a r.v. X and a parameter space Θ, elements θ with true value θ_0
2) We know that the expectation of some function $g(X,\theta)$ is zero at θ_0
3) Get data, iid copies of $X_1, \dots, X_N$
4) "Analog principle": look at the sample equivalent of the expectation in 2)
5) Set 4) to zero

- Example 1. Suppose we have an r.v X with mean μ_0
  - Steps (1) and (2):
$$\mathrm{E}(X) = \mu_0 \;\Leftrightarrow\; \mathrm{E}(X - \mu_0) = 0 \;\Rightarrow\; g(X,\mu) = X - \mu$$
  - Step (3): get data $X_1, \dots, X_N$
  - Steps (4) and (5):
$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu}) = 0 \;\Rightarrow\; \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- Example 2. Suppose we have an OLS model, $y = X\beta_0 + \epsilon$, with $\mathrm{E}(\epsilon|X) = 0$.
$$\mathrm{E}(\epsilon|X) = 0 \;\Rightarrow\; \mathrm{E}(h(X)\epsilon|X) = 0 \;\Rightarrow\; \mathrm{E}(h(X)\epsilon) = 0 \;\Rightarrow\; \mathrm{E}(X'\epsilon) = 0$$

- Note that $y \in \mathbb{R}$, $\beta \in \mathbb{R}^k$, $X \in \mathbb{R}^{1\times k}$, $\epsilon \in \mathbb{R}$, and hence $\mathrm{E}(X'\epsilon) \in \mathbb{R}^k$
- Steps (1) and (2):

$$y = X\beta_0 + \epsilon \;\Rightarrow\; \epsilon = y - X\beta_0$$
$$\mathrm{E}(X'\epsilon) = 0$$

$\big((y_1, X_1), \dots, (y_N, X_N)\big)$ is a random sample of (y, X). Hence,

$$\mathrm{E}(X'\epsilon) = \mathrm{E}\big(X'(y - X\beta_0)\big) = 0$$

- Sample analog

$$\frac{1}{N}\sum_{i=1}^{N} X_i'\big(y_i - X_i'\hat{\beta}\big) = 0 \;\Rightarrow\; \hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i'X_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i'y_i\right)$$

- Example 3. $y = X\beta_0 + \epsilon$, $\mathrm{E}(\epsilon|Z) = 0$ with $Z \in \mathbb{R}^k$, and hence

$$\mathrm{E}(Z'\epsilon) = 0 \;\Rightarrow\; \mathrm{E}\big(Z'(y - X\beta_0)\big) = 0$$

Random sample: $(y_i, X_i, Z_i)_i$. Thus the sample analog:

$$\frac{1}{N}\sum_{i=1}^{N} Z_i'\big(y_i - X_i\hat{\beta}\big) = 0 \;\Rightarrow\; \hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'X_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'y_i\right)$$

- Example 4. $\mathrm{E}[g(X, \theta_0)] = 0$, where $g(\cdot)$ and θ_0 have the same dimension.
  - However, assume $\theta_0 \in \Theta \subset \mathbb{R}^p$. Then

$$g : \mathbb{R}^d \times \Theta \to \mathbb{R}^q, \qquad \text{where } q \geq p$$

    - Note that we've assumed $q = p$ in the previous examples.
  - Function g has q components, and θ has p components. There are more components of g than parameters.

GMM (cont'd)

❖ We have $\mathrm{E}\big(h(X,\theta_0)\big)=0$, iid data $(X_1,\dots,X_N)$, and the analog principle:

$$\frac{1}{N}\sum_{i=1}^{N} h\big(X_i,\hat{\theta}\big)=0$$

where $h(\cdot)$ is $q\times 1$ and θ is $p\times 1$. But the equality can hold only when $q=p$.

➢ When $q>p$, one solution is to introduce a matrix $A_N\in\mathbb{R}^{p\times q}$

$$A_N\frac{1}{N}\sum_{i=1}^{n} h\big(X_i,\hat{\theta}\big)=0$$

This leads to a criterion function

$$Q_N(\theta)=\left(\frac{1}{N}\sum_{i=1}^{N} h(X_i,\theta)\right)' A_N' A_N\left(\frac{1}{N}\sum_{i=1}^{N} h(X_i,\theta)\right) \;\Rightarrow\; \hat{\theta}=\arg\max_{\theta\in\Theta} Q_N(\theta)$$

➢ Example. $y=X\beta_0+\epsilon$ with $\mathrm{E}(Z'\epsilon)=0$, where $X',\beta_0\in\mathbb{R}^p$, $Z'\in\mathbb{R}^q$, $q>p$.

$$\mathrm{E}\big(Z'(y-X\beta)\big)=0 \;\Rightarrow\; \frac{1}{N}\sum_{i=1}^{N} Z_i'\big(y_i-X_i\hat{\beta}\big)\simeq 0$$

The criterion function is

$$\left(\frac{1}{N}\sum_i Z_i'(y_i-X_i\beta)\right)' W_N\left(\frac{1}{N}\sum_i Z_i'(y_i-X_i\beta)\right)$$

notations:

$$h(X,\theta),\qquad h(\theta)=\mathrm{E}\big(h(X,\theta)\big),\qquad h_N(\theta)=\frac{1}{N}\sum_i h(X_i,\theta)$$

Hence the criterion function simplifies to

$$Q_N(\theta)=h_N(\theta)'W_N h_N(\theta)$$

The derivative

$$\frac{\partial Q_N(\theta)}{\partial\theta}=2\underbrace{\left(\frac{\partial h_N}{\partial\theta}\right)}_{G_N(\theta)}W_N h_N(\theta)$$

Let

$$G_N(\theta)=\sum_i\frac{\partial h(X_i,\theta)}{\partial\theta}\bigg|_\theta=\frac{\partial\sum_i h(X_i,\theta)}{\partial\theta}\bigg|_\theta=\frac{\partial h_N(\theta)}{\partial\theta}\bigg|_\theta$$

Set

$$G_N\big(\hat{\theta}\big)'W_N\underbrace{h_N\big(\hat{\theta}\big)}_{Z'y-Z'X\hat{\beta}}=0 \;\Rightarrow\; -(Z'X)'W_N\big(Z'y-Z'X\hat{\beta}\big)=0$$

$$\Rightarrow\; X'ZW_NZ'y=(X'ZW_NZ'X)\hat{\beta}$$

$$\Rightarrow\; \hat{\beta}=(X'ZW_NZ'X)^{-1}(X'ZW_NZ'y)$$

❖ Panel data

$$y_{it}=X_{it}\beta+\epsilon_{it},\qquad t=1,\dots,T,\;\; i=1,\dots,N$$

where T is fixed and $N\to\infty$. The data looks like

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \qquad X_i = \begin{pmatrix} x_{i1,1} & \cdots & x_{i1,K} \\ \vdots & \ddots & \vdots \\ x_{iT,1} & \cdots & x_{iT,K} \end{pmatrix}$$

- There are Tk moment conditions
$$\mathrm{E}(X_{it}'\epsilon_{it}) = 0, \qquad \forall t = 1, \dots, T$$
where $X_{it} = (x_{it,1}, \dots, x_{it,K})$ and each regressor $x_{it,k}$ are exogenous, i.e. $\mathrm{E}(x_{it,k}\epsilon_{it}) = 0 \ \forall k$.
- Another way to impose exogeneity is to have
$$\mathrm{E}(x_{is}'\epsilon_{it}) = 0, \qquad \forall s, t = 1, \dots, T, \ \ s \leq t$$
Here we have KT^2 moment conditions.
- These show the advantage of GMM: you have more moment conditions, depending how you assume the exogeneity between variables across time, etc., you'll get even more.

GMM Theory: Consistency, Asymptotic Normality, and Efficiency

❖ *Definition.*

$$\hat{\theta} = \arg\min_{\theta\in\Theta} h_N(\theta)'W_N h_N(\theta) = \arg\min_{\theta\in\Theta} Q_N(\theta)$$

where $h_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} h(X_i, \theta)$.

❖ **Theorem (Consistency).** If

1) $X_1, \dots, X_N$ are iid
2) $\mathrm{E}\big(h(X,\theta)\big) = 0 \;\Leftrightarrow\; \theta = \theta_0$
3) Uniform convergence of h_N to h. This is satisfied when
   a. $h(X,\theta)$ is continuous in θ
   b. $\mathrm{E}(\sup_{\theta\in\Theta}|h(X,\theta)| < \infty)$
   c. Θ is compact
4) $W_N \xrightarrow{p} W$ a positive definite and symmetric matrix

then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. General steps

- (3) gives uniform convergence of h_N to h, i.e.

$$\sup_{\theta\in\Theta}|h_N(\theta) - h(\theta)| \xrightarrow{p} 0$$

- Imply uniform convergence of Q_N to Q_0?
- Imply pointwise convergence of $Q_0\big(\hat{\theta}\big)$ to $Q_0(\theta_0)$?
- Imply $\hat{\theta} \xrightarrow{p} \theta_0$?

Now the proof:

$$\begin{aligned}|Q_N(\theta) - Q_0(\theta)| &= |h_N(\theta)'W_N h_N(\theta) - h(\theta)'Wh(\theta)| \\ &= |(h_N - h)'W_N(h_N - h) + h'(W_N + W_N')(h_N - h) + h'(W_N - W)h|\end{aligned}$$

Thus we've shown that

$$\sup_{\theta\in\Theta}|Q_N(\theta) - Q_0(\theta)| \xrightarrow{p} 0$$

We know that

$$\begin{aligned}0 < Q_0\big(\hat{\theta}\big) - Q_0(\theta_0) &= Q_0\big(\hat{\theta}\big) - Q_N\big(\hat{\theta}\big) + Q_N\big(\hat{\theta}\big) - Q_0(\theta_0) \\ &< Q_0\big(\hat{\theta}\big) - Q_N\big(\hat{\theta}\big) + Q_N(\theta_0) - Q_0(\theta_0) \\ &\le 2\sup_{\theta\in\Theta}|Q_N(\theta) - Q_0(\theta)| \\ &\xrightarrow{p} 0\end{aligned}$$

Since $\mathrm{E}\big(h(X,\theta)\big) = 0 \;\Leftrightarrow\; \theta = \theta_0$,

$$\forall\epsilon > 0, \exists\delta > 0 : \big|\hat{\theta} - \theta_0\big| > \delta \;\Rightarrow\; \big|Q_0\big(\hat{\theta}\big) - Q_0(\theta_0)\big| > \epsilon$$

$$\Rightarrow\; \Pr\big(\big|\hat{\theta} - \theta_0\big| > \delta\big) \le \underbrace{\Pr\big(\big|Q_0\big(\hat{\theta}\big) - Q_0(\theta_0)\big| > \epsilon\big)}_{\xrightarrow{p} 0}$$

$$\Rightarrow\; \hat{\theta} \xrightarrow{p} \theta_0$$

This completes the proof.

❖ Recall the derivative of h is

$$G_N(\theta) = \left.\frac{\partial h_N(\theta)}{\partial \theta}\right|_\theta, \qquad G_0(\theta) = \mathrm{E}\left(\left.\frac{\partial h(X,\theta)}{\partial \theta}\right|_\theta\right)$$

Assume $\theta_0 \in \mathrm{int}(\Theta)$ and consistency of $\hat{\theta}$. Then,

$$-2G_N(\hat{\theta})' W_N h_N(\hat{\theta}) = 0 \;\Rightarrow\; G_N(\hat{\theta})' W_N h_N(\hat{\theta}) = 0$$

Apply the Mean value expansion

$$h_N(\hat{\theta}) = h_N(\theta_0) + G_N(\tilde{\theta})(\hat{\theta} - \theta_0)$$

where

$$\hat{\theta} \xrightarrow{p} \theta_0, \qquad \tilde{\theta} \in (\hat{\theta}, \theta_0), \qquad \tilde{\theta} \xrightarrow{p} \theta_0$$

Therefore,

$$\begin{aligned}
G_N(\hat{\theta})' W_N h_N(\hat{\theta}) &= 0 \\
G_N(\hat{\theta})' W_N \left(h_N(\theta_0) + G_N(\tilde{\theta})(\hat{\theta} - \theta_0)\right) &= 0 \\
G_N(\hat{\theta})' W_N \sqrt{N} h_N(\theta_0) &= -G_N(\hat{\theta})' W_N G_N(\tilde{\theta}) \sqrt{N}(\hat{\theta} - \theta_0) \\
\sqrt{N}(\hat{\theta} - \theta_0) &= -\left(G_N(\hat{\theta})' W_N G_N(\tilde{\theta})\right)^{-1} G_N(\hat{\theta})' W_N \sqrt{N} h_N(\theta_0) \\
&\xrightarrow{p} \left(G_0(\theta_0)' W G_0(\theta_0)\right)^{-1} G_0(\theta_0) W \underbrace{\sqrt{N} h_N(\theta_0)}_{\xrightarrow{d} \mathcal{N}(0, S_0)} \\
&\xrightarrow{d} \mathcal{N}(0, \Omega)
\end{aligned}$$

where

$$\Omega = (G_0' W G_0)^{-1} G_0' W S_0 W G_0' (G_0' W G_0)^{-1}$$

❖ **Theorem (Asymptotic Normality).** If

1) All conditions of the consistency theorem hold
2) $\mathrm{E}[h(X,\theta_0)h(X,\theta_0)'] = S_0 < \infty$, i.e. variance of the moment condition at the true value is finite
3) $\frac{\partial h(X,\theta)}{\partial \theta_0}$ is continuous and uniformly bounded
   - Note also that $G_N \xrightarrow{p} G_0$ pointwise. So pointwise convergence together with continuity and uniform boundedness gives uniform convergence of G_N to G_0.
4) $G_0(\theta_0)$ has full rank

then $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$.

GMM Theory (cont'd)

❖ Recall from last time:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, Q_N(\theta), \qquad Q_0(\theta) = 0 \;\Leftrightarrow\; \theta = \theta_0$$

$$Q_N(\theta) = h_N(\theta)' W_N h_N(\theta), \qquad \text{where } h_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} h(X,\theta)$$

- Consistency: $\hat{\theta}_{W_N} \xrightarrow{p} \theta_0$
- Asymptotic normality: $\sqrt{N}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, \Omega)$, where

$$\Omega = (G'WG)^{-1}G'WS_0WG(G'WG)^{-1}$$
$$S_0 = \mathrm{E}(h(X,\theta_0)h(X,\theta_0)')$$
$$G = \mathrm{E}\left(\left.\frac{\partial h(X,\theta)}{\partial \theta}\right|_{\theta=\theta_0}\right)$$
$$W_N \xrightarrow{p} W$$

❖ Optimal W is S_0^{-1}

$$\begin{aligned}(G'WG)^{-1}G'WS_0WG(G'WG)^{-1} &\overset{W=S_0^{-1}}{=} (G'S_0^{-1}G)^{-1}G'S_0^{-1}S_0S_0^{-1}G(G'S_0^{-1}G)^{-1}\\ &= (G'S_0^{-1}G)^{-1}(G'S_0^{-1}G)(G'S_0^{-1}G)^{-1}\\ &= (G'S_0^{-1}G)^{-1}\end{aligned}$$

❖ To show that $W = S_0^{-1}$ is indeed optimal, we show that the expression

$$(G'WG)^{-1}G'WS_0WG(G'WG)^{-1} - (G'S_0^{-1}G)^{-1}$$

is positive definite. Pre- and post-multiply $(G'WG)$,

$$\begin{aligned}G'WS_0WG - G'WG(G'S_0^{-1}G)^{-1}G'WG &= G'WS_0^{1/2}\Psi S_0^{1/2}WG\\ &= G'WS_0^{1/2}\left[I - S_0^{-1/2}G(G'S_0^{-1}G)^{-1}G'S_0^{-1/2}\right]S_0^{1/2}WG\end{aligned}$$

Thus we want to show that Ψ is positive definite. $A'\Psi A$ is PD if Ψ is PD.

- Ψ is PD if Ψ is idempotent, i.e. $\Psi = \Psi\Psi$ and $\Psi = \Psi'$.
- If Ψ is idempotent, then $I - \Psi$ is idempotent.
- Notice that Ψ is like a projection matrix (let $X' = G'S_0^{-1/2}$), $X(X'X)^{-1}X'$, which is idempotent. Thus we can conclude that $W^* = S_0^{-1}$

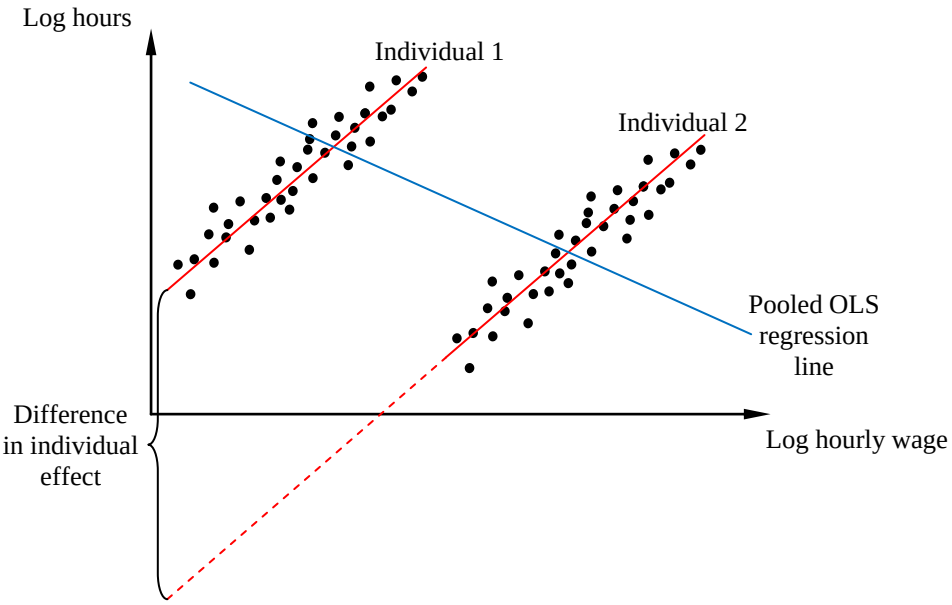
❖ To implement GMM, the traditional way:

- Choose W_1
- Get $\hat{\theta}_{W_1}$
- Estimate S_0 by $\hat{S}$
- Get $\hat{\theta}_{\hat{S}^{-1}}$

Panel Data

C&T Ch. 21, skip 21.5

- Example



- Panel data model

$$y_{it} = \alpha_{it} + x_{it}\beta_{it} + \epsilon_{it}$$

usually let $\alpha_{it} = \alpha$ and $\beta_{it} = \beta_i$.

- Benefits of using Panel data
  - Simply have more data, more information, allows
  - Allows us to deal with omitted variable bias, e.g. laziness of people
  - Get free instrument

Panel Data (cont'd)

C&T Ch.21.2.3 will not be lectured, but should be read—there will be a final exam question on this section

❖ Use the following model by default:

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}, \qquad \text{where } i = 1, \dots, N;\ t = 1, \dots, T$$

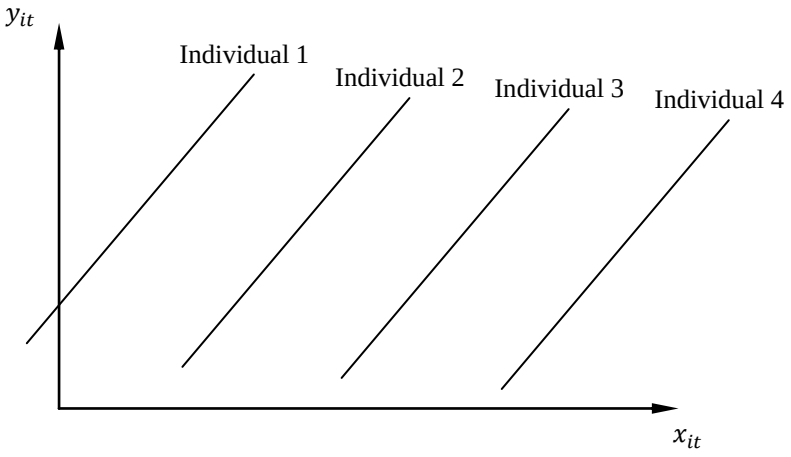
with T fixed and $N \to \infty$

- Models
  - Fixed effects
  - Random effects
- Estimators
  - FE
  - RE
  - Pooled OLS
  - Between
  - First-difference

❖ Model

- $\mathrm{E}(\alpha_i | x_{i1}, \dots, x_{iT})$
- Relation between $(\epsilon_{i1}, \dots, \epsilon_{iT})$ and $(x_{i1}, \dots, x_{iT})$
  - Assume *strict/strong exogeneity*: $\mathrm{E}(\epsilon_{it} | x_{i1}, \dots, x_{iT}) = 0$ in Ch.21.
  - Will relax this in Ch.22.

- ***Random Effects* (*RE*) *model*** assumes $\mathrm{E}(\alpha_i | x_{i1}, \dots, x_{iT}) = 0$
  - No individual effects
- ***Fixed Effects* (*FE*) *model*** assumes nothing about $\mathrm{E}(\alpha_i | x_{i1}, \dots, x_{iT})$



  - This picture violates that α_i doesn't change with x_{it}, i.e. $\mathrm{E}(\alpha_i | x_{i1}, \dots, x_{iT}) = 0$
  - Practical advice: always use FE models (except perhaps in experiments)

- Conditional means and marginal effects
  - FE model:

$$\mathrm{E}(y_{it} | \alpha_i, x_{it}) = \alpha_i + x_{it}\beta$$

The marginal effect is

$$\frac{\partial \mathrm{E}(y_{it}|\alpha_i, x_{it})}{\partial x_{it,k}} = \beta_k$$

If we don't know the individual effects α_i,

$$\mathrm{E}(y_{it}|x_{it}) = \mathrm{E}(\alpha_i|x_{it}) + x_{it}\beta = \alpha(x_{it}) + x_{it}\beta \;\Rightarrow\; \frac{\partial \mathrm{E}(\cdot)}{\partial x} = \alpha'(\cdot) + \beta$$

- RE model:

$$\mathrm{E}(y_{it}|x_{it}) = \mathrm{E}(\alpha_i|x_{it}) + x_{it}\beta = x_{it}\beta$$

❖ Take a RE model. Let $\epsilon_{it} \sim iid$ with mean zero and variance σ_ϵ^2, $\alpha_i \sim iid$ with mean zero and variance σ_α^2

$$y_{it} = x_{it}\beta + (\alpha_i + \epsilon_{it}) = x_{it} + u_{it}, \qquad \mathrm{E}(u_{it}|x_{it}) = 0$$

Impose restriction that α_i and ϵ_i are uncorrelated. Then,

$$\mathrm{Cov}(y_{i1}, y_{i2}|X) = \mathrm{Cov}(\alpha_i + \epsilon_{i1}, \alpha_i + \epsilon_{12}) = \mathrm{Cov}(\alpha_i, \alpha_i) = \mathrm{Var}(\alpha_i) = \sigma_\alpha^2$$

Then,

$$\mathrm{Var}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} X = \begin{pmatrix} \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & 0 & 0 & 0 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 & 0 & 0 & 0 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ 0 & 0 & 0 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 \\ 0 & 0 & 0 & \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{pmatrix}$$

Hence, we can estimate β consistently using (F)GLS.

➢ How is this related to GLS? Recall that in OLS, we assume

$$y_{it} = x_{it}\beta + \epsilon_{it}, \qquad \begin{cases} \mathrm{E}(\epsilon_{it}|x_{it}) = 0 \\ \mathrm{Var}(\epsilon_{it}|x_{it}) = \sigma^2 \end{cases}$$

Thus,

$$\mathrm{Var}\begin{pmatrix} y_{11} \\ \vdots \\ y_{NT} \end{pmatrix} X = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

Under this condition, OLS is consistent. Allowing for heteroskedasticity and autocorrelation, we use GLS to estimate β, if we can estimate the variance-covariance matrix.

❖ Estimators for panel data

➢ First, notice that there are two dimensions in panel data: N and T

- This affects the kind of regressors
  - Time variant: $x_{it} = x_{it} \neq x_{is}$ for all $s \neq t$
  - Time invariant: $x_{it} = x_i$ for all t

➢ It is important to be clear about the assumed relationships between x_{it}, α_i, and ϵ_{it}. These assumptions are crucial for determining when to use which estimator, and whether the chosen estimator is consistent.

➢ ***Pooled OLS***: run OLS on all data

- Assume RE model: $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it} = x_{it}\beta + (\alpha_i + \epsilon_{it})$, with $\mathrm{E}(\alpha_i|x_{it}) = 0$,

$\mathrm{E}(\epsilon_{it}|x_{it}) = 0$. Then P-OLS is consistent
- If $\mathrm{Var}(\alpha_i|x_{it}) = 0$, then P-OLS is efficient
- Assume FE model, $y_{it} = x_{it}\beta + (\alpha_i + \epsilon_{it})$, where $\mathrm{E}(\alpha_i|x_{it}) \neq 0$. Then P-OLS is not consistent.

➢ ***Between estimator***
- Take time averages,

$$(y_{i1}, \dots, y_{iT}) \mapsto \frac{1}{T}\sum_{t=1}^{T} y_{it} = \bar{y}_i, \qquad x_{it} \mapsto \bar{x}_i, \qquad NT \mapsto N$$

Do OLS on the transformed data:

$$\bar{y}_i = \alpha_i + \bar{x}_i\beta + \bar{\epsilon}_i$$

Under the RE assumption, i.e. $\mathrm{E}(\alpha_i|x_{i1}, \dots, x_{iT}) = 0 \;\Rightarrow\; \mathrm{E}(\alpha_i|\bar{x}_i) = 0$, this estimator is consistent. However, under the FE model, this is not consistent.
- This is really collapsing the T dimension and using only the N dimension to compare individuals only. "Between" means we're looking at the differences *between* the individuals.

➢ ***Within estimator*** (FE)
- Subtract the "between" equation from the FE model:

$$\begin{aligned} y_{it} - \bar{y}_i &= (\alpha_{it} + x_{it}\beta + \epsilon_{it}) - (\alpha_i + \bar{x}_i\beta + \bar{\epsilon}_i) \\ &= 0 + (x_{it} - \bar{x}_i)\beta + (\epsilon_{it} - \bar{\epsilon}_i) \\ \Rightarrow\; \tilde{y}_{it} &= \tilde{x}_{it}\beta + \tilde{\epsilon}_{it} \end{aligned}$$

This estimator is consistent under both FE and RE, but hinges (crucially) on the assumption that $(\epsilon_{i1}, \dots, \epsilon_{iT}) \perp (x_{i1}, \dots, x_{iT})$
- This estimator requires weaker assumption than in RE and Pooled OLS to yield consistency.
- By subtracting the individual average $\bar{y}_i$ from y_{it}, this estimator allows us to focus on the T dimension of the data.

➢ ***First-difference***
- Take first-difference of the model

$$\begin{aligned} y_{it} - y_{i,t-1} &= (\alpha_i + x_{it}\beta + \epsilon_{it}) - \left(\alpha_i + x_{i,t-1}\beta + \epsilon_{i,t-1}\right) \\ \Rightarrow\; \Delta y_{it} &= 0 + \Delta x_{it}\beta + \Delta\epsilon_{it} \end{aligned}$$

For consistency, we need to impose

$$\mathrm{E}\left(\epsilon_{it} - \epsilon_{i,t-1}\middle|x_{it} - x_{i,t-1}\right) = 0$$

which is a weaker condition than the "strong exogeneity" assumption at the beginning
- This estimator works under both RE and FE, since α_i is not there

➢ ***Random effects***: it's like "Pooled Feasible GLS"
- This estimator can be thought of as the "optimal" combination of the between and within estimators, which focus on the N and the T dimensions, respectively.

Random Effect v.s. Fixed Effects

- Strict exogeneity, i.e. $\mathrm{E}(\epsilon_{it}|x_{i1}, \dots, x_{iT}) = 0$.
- In addition, if we…
  - assume $\mathrm{E}(\alpha_i|x_{i1}, \dots, x_{iT}) = 0$, we'd have the RE model
    - Here we can think of α_i as the time invariant omitted variable
  - or assume nothing about α_i, then we'd have the FE model

- Consistent estimators for panel data models:
  - FE model → the *Within estimator* and the *First-Difference estimator*
  - RE model → the *Pooled OLS*, *Between estimator*, and *RE* (or *Pooled GLS*) *estimator*

- FE estimator

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}\beta + \epsilon_{it} \\ - \quad \bar{y}_i &= \alpha_i + \bar{x}_i\beta + \bar{\epsilon}_i \quad \text{average over time} \\ \hline \tilde{y}_{it} &= 0 + \tilde{x}_{it}\beta + \tilde{\epsilon}_{it} \quad \text{deviation from mean} \end{aligned}$$

Then, consistency of $\hat{\beta}_{FE}$:

$$\begin{aligned} \hat{\beta}_{FE} &= \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{y}_{it}\right) \\ &= \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'(\tilde{x}_{it}\beta + \tilde{\epsilon}_{it})\right) \\ &= \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)\beta \\ &\qquad + \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{\epsilon}_{it}\right) \\ &= \beta + \Sigma_{xx}^{-1}\cdot\frac{1}{NT}\sum_{i,t}\tilde{x}_{it}'\tilde{\epsilon}_{it} \\ &\xrightarrow{p} \beta, \qquad \text{if } \frac{1}{NT}\sum_{i,t}\tilde{x}_{it}'\tilde{\epsilon}_{it} \xrightarrow{p} 0 = \mathrm{E}(\tilde{x}_{it}'\tilde{\epsilon}_{it}) \end{aligned}$$

  - In Ch.21, $\mathrm{E}(\tilde{x}_{it}'\tilde{\epsilon}_{it}) = 0$ is guaranteed by strictly exogeneity! But strict exogeneity is a much stronger condition. It would be sufficient to just require $\tilde{\epsilon}_{it}$ to be exogenous.
  - If we have time-invariant regressors, i.e. $x_{it} = x_i$, then we cannot estimate the coefficients for them:

$$\begin{aligned} y_{it} &= x_i\beta + \epsilon_{it} + \alpha_i \\ - \quad \bar{y}_i &= x_i\beta + \bar{\epsilon}_i + \alpha_i \\ \hline \tilde{y}_{it} &= 0\cdot\beta + \tilde{\epsilon}_{it} + 0 \end{aligned}$$

    So in FE models, <u>NO time-invariant coefficients can estimated</u>.

- Hausman Test. H_0 : RE model is true.
  - If H_0 is true, then, RE estimator is consistent and efficient; and the FE estimator is

consistent (but not efficient).

- Suppose H_0 is not true. Then, RE estimator is inconsistent; but FE estimator is consistent.
- If $\left|\hat{\theta}_{RE} - \hat{\theta}_{FE}\right|$ is large, this is evidence against H_0.
- Let $\hat{\theta}_{FE,1}$ and $\hat{\theta}_{RE,1}$ be the time-variant estimates, where 1 denotes variant regressors
- So the test is

$$H = \left(\hat{\theta}_{RE,1} - \hat{\theta}_{FE,1}\right)' \left(\text{Var}\left(\hat{\theta}_{RE,1} - \hat{\theta}_{FE,1}\right)\right)^{-1} \left(\hat{\theta}_{RE,1} - \hat{\theta}_{FE,1}\right)$$

- Homework:
  - Show that

$$H \xrightarrow{d} \chi^2_{d_1}, \qquad \text{where } d_1 \text{ is \# of time variant regressors}$$

  - Show that

$$\text{Var}\left(\hat{\theta}_{RE,1} - \hat{\theta}_{FE,1}\right) = \text{Var}\left(\hat{\theta}_{FE,1}\right) - \text{Var}\left(\hat{\theta}_{RE,1}\right)$$

Dynamics in Panel Data

C&T Ch.22, skip 22.2.7, 22.4.4, 22.4.5, 22.5.5

❖ In dealing with the "dynamics" in panel data, we...
- ➢ relax the assumption of *strict exogeneity*, i.e. $\mathrm{E}(x_{it}'\epsilon_{is}) = 0$ for all s, t
- ➢ study dynamics in y, i.e. using $y_{i,t-1}$ as regressor

❖ Notations

$$y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it}, \qquad i = 1, \dots, N, \;\; t = 1, \dots, T$$

- ➢ So far, we've been assuming $\mathrm{E}(x_{it}'\epsilon_{is}) = 0 \;\; \forall s, t$.
- ➢ Let the individual i be the basic unit of observation. Define notations

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix} \in \mathbb{R}^T, \qquad X_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1K} \\ \vdots & \ddots & \vdots \\ x_{iT1} & \cdots & x_{iTK} \end{pmatrix}, \qquad \iota_T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{pmatrix}$$

Then the model can be written as

$$y_i = \alpha_i \iota_T + X_i\beta + \epsilon_i$$

❖ ***Instrument***: a variable z for which we can write down an exogeneity condition. In linear models, this means $\mathrm{E}(z_{it}'u_{it}) = 0$.
- ➢ In RE models, we have $\alpha_i \perp X_i$. With strict exogeneity, this means

$$\mathrm{E}(x_{is}'\epsilon_{it}) = 0 \;\; \Rightarrow \;\; \mathrm{E}\big(x_{is}'(\alpha_i + \epsilon_{it})\big) = 0$$

- ➢ Relaxing strict exogeneity
  - "***Summation***"

$$\mathrm{E}\left(\sum_{t=1}^{T} x_{it}'\epsilon_{it}\right) = 0$$

Let $u_{it} = \alpha_i + \epsilon_{it}$, and $u_i = \alpha_i\iota_T + \epsilon_i$. Then,

$$\mathrm{E}\left(\sum_{t=1}^{T} x_{it}'u_{it}\right) = 0 \;\; \Leftarrow \;\; \mathrm{E}(x_{it}'\epsilon_{it}) = 0$$

So this assumption is weaker than strict exogeneity. It is saying that if ϵ_{it} is positively correlated with x_{it} for some time period, then the two will be negatively correlated in other ones; and overall, the two correlations cancel each other out. This is like doing Pooled-OLS.

- Consider the following example:

$$h(X_i, \beta) = \sum_t x_{it}'(y_{it} - x_{it}'\beta), \qquad \text{where } \mathrm{E}\big(h(X_i, \beta)\big) = 0$$

Let z_i contain the instruments, and $u_i = y_i - X_i'\beta$. Then

$$z_i'(y_i - X_i'\beta) = z_i'\begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix}.$$

Under the "summation" assumption, we have

$$\mathrm{E}(z_i'u_i) = [z_{i1}' \quad \cdots \quad z_{iT}']\begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix} = \sum_{t=1}^{T} z_{it}'u_{it} = 0$$

- ***Contemporaneous exogeneity***

$$\mathrm{E}(x_{it}'u_{it}) = 0, \qquad \forall t = 1, \dots, T$$

  - This implies "summation" (i.e. stronger than the "summation" assumption)
  - Over-identification: there are $K \cdot T$ moment conditions but only K parameters

$$\mathrm{E}(x_{itk}'\epsilon_{it}) = 0, \qquad \forall t, k$$

    Have to use GMM.
  - consider the example

$$Z_i'(y_i - X_i'\beta)$$

    Here Z_i' is a $(TK \times T)$ matrix: since u_i is $T \times 1$ so that Z_i' has to have T columns for the multiplication to be conformable; since there are TK moment conditions, there has to be TK rows in Z_i'. So that

$$Z_i'u_i = \underbrace{\begin{pmatrix} x_{i1}' & 0 & \cdots & 0 \\ 0 & x_{i2}' & 0 & \vdots \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & x_{iT}' \end{pmatrix}}_{KT \times T} \underbrace{\begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{pmatrix}}_{T \times 1} = \begin{pmatrix} x_{i1}'u_{i1} \\ x_{i2}'u_{i2} \\ \vdots \\ x_{iT}'u_{iT} \end{pmatrix}$$

    where each x_{it} is a $K \times 1$ vector.
    The sample analog of the moment conditions

$$\mathrm{E}\big(Z_i'(y_i - X_i'\beta)\big) = 0 \;\Rightarrow\; \frac{1}{N}\sum_{i=1}^{N} Z_i'(y_i - X_i'\beta) = 0$$

$$\Rightarrow\; \frac{1}{N}Z'u = \frac{1}{N}Z'(y - X\beta)$$

    where

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{pmatrix}, \qquad y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{NT} \end{pmatrix}$$

    In GMM, the criterion function is

$$Q_N(\beta) = \big(y - X\hat{\beta}\big)'ZWZ'\big(y - X\hat{\beta}\big) = 0$$

    The FOC is

$$\frac{\partial Q_N(\beta)}{\partial \beta} = -2X'ZWZ'\big(y - X\hat{\beta}\big) = 0 \;\Rightarrow\; X'ZWZ'\big(y - X\hat{\beta}\big) = 0$$

$$\Rightarrow\; X'ZWZ'y = X'ZWZ'X\hat{\beta}$$

$$\Rightarrow\; \hat{\beta} = (X'ZWZ'X)^{-1}X'ZWZ'y$$

- ***Weak exogeneity***

$$\mathrm{E}(x_{is}'\epsilon_{it}) = 0, \qquad \forall s \le t$$

  - This implies contemporaneous exogeneity, and hence also "summation"
  - This assumption is saying that current regressors are uncorrelated with future error terms.

- Construct matrix for moment conditions

$$Z_i'u_i = \underbrace{\begin{pmatrix} x_{i1}' & 0 & \cdots & 0 \\ 0 & x_{i1}' & \ddots & \vdots \\ \vdots & x_{i2}' & 0 & \vdots \\ \vdots & 0 & x_{i1}' & \vdots \\ \vdots & \vdots & x_{i2}' & \vdots \\ \vdots & \vdots & x_{13}' & 0 \\ \vdots & \vdots & 0 & x_{i1}' \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \vdots \end{pmatrix}}_{D\times T} \underbrace{\begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix}}_{T\times 1} = \underbrace{\begin{pmatrix} x_{i1}'u_{i1} \\ x_{i1}'u_{i2} \\ x_{i2}u_{i2} \\ x_{i1}u_{i3} \\ x_{i2}u_{i3} \\ x_{i3}u_{i3} \\ \vdots \end{pmatrix}}_{D\times 1}$$

where $D = 1 + 2 + \cdots + T$

❖ Dynamic panels

$$y_{i,t} = \alpha_i + x_{i,t}\beta + \gamma y_{i,t-1} + \epsilon_{i,t}$$

- $N \to \infty$, iid observations and T fixed
- $\mathrm{E}(\epsilon_{is}\epsilon_{it}) = 0$ if $s \neq t$, i.e. no error terms are correlated across periods
- $|\gamma| < 1$, stationarity of y
- Weak exogeneity

- Pooled-OLS

$$y_{it} = \alpha_i + \gamma y_{i,t-1} + \epsilon_{it} \;\Rightarrow\; y_{it} = \gamma y_{i,t-1} + (\alpha_i + \epsilon_{it})$$
$$y_{i,t-1} = \alpha_i + \gamma y_{i,t-2} + \epsilon_{i,t-1}$$

For the estimate γ to be consistent, we need $\mathrm{E}(y_{i,t-1}\alpha_i) = 0$. But clearly, $\mathrm{E}(y_{i,t-1}\alpha_i) \neq 0$.Thus OLS estimate is inconsistent.

- FE model

$$y_{it} - \bar{y}_i = 0 + \gamma(y_{i,t-1} - \bar{y}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

But notice that

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}\,, \qquad \bar{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$$

and hence $\mathrm{E}[\epsilon_{it} - \bar{\epsilon}_i | y_{it} - \bar{y}_i] \neq 0$. Thus, the estimate is not consistent either.

- First difference.

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta\epsilon_{it}, \qquad \text{where } \Delta y_{it} = y_{it} - y_{i,t-1},\ \Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{i,t-1}$$

But this is still inconsistent, because $\Delta\epsilon_{it}$ contains $\epsilon_{i,t-1}$ which is correlated with $y_{i,t-1}$, a component of Δy_{it}.

- Homework: how do we solve this inconsistency problem in dynamic panel data models?

Dynamic Panel Data (cont'd)

❖ Wrapping up dynamic panel data:

$$y_{i,t} = \alpha_i + \gamma y_{i,t-1} + \epsilon_{it} \;\Rightarrow\; \Delta y_{it} = 0 + \gamma \Delta y_{i,t-1} + \Delta \epsilon_{i,t}$$

Note that $\epsilon_{i,t-1}$ and $y_{i,t-1}$ are strongly correlated.

- Assume there's no serial correlation. Then $y_{i,t-2}$ is a good instrument, and so are $y_{i,t-3}$, $y_{i,t-4}$, etc.
- The moment condition is $\mathrm{E}(Z_i' u_i) = 0$, where

$$u_i = \begin{pmatrix} \epsilon_{i3} - \epsilon_{i2} \\ \epsilon_{i4} - \epsilon_{i3} \\ \vdots \\ \epsilon_{iT} - \epsilon_{i,T-1} \end{pmatrix}, \qquad Z_i' = \begin{pmatrix} y_{i1} & 0 & \cdots & 0 & \\ 0 & y_{i1} & & \ddots & \\ 0 & y_{i2} & & & \\ \vdots & 0 & y_{i1} & & \\ \vdots & 0 & y_{i2} & & \\ & & y_{i3} & & \\ & & & \ddots & \\ 0 & \cdots & \cdots & 0 & y_{i,T-2} \end{pmatrix}$$

❖ Binary choice with panel data

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad \text{where } y_i^* = x_i'\beta + \epsilon_i,\ \ \epsilon_i | x \sim \mathcal{N}(0,1),\ \ \Pr(y = 1) = \Phi(x_i'\beta)$$

In panel data,

$$\Pr(y_{it} = 1 | x_{it}, \alpha_i, \beta) = \begin{cases} (1) & \Pr(y_{it} | x_{it}, \beta) \\ (2) & \alpha_i + \Pr(y_{it} | x_{it}, \beta) \\ & \alpha_i \cdot \Pr(y_{it} | x_{it}, \beta) \end{cases}$$

- Problem with (1):
  - Neglects unobserved heterogeneity
- Problem with (2):
  - It is possible that $\Pr(y_{it} | x_{it}, \alpha_i, \beta) \notin [0,1]$
  - Not consistent with economic theory:

$$y_{it}^* = \alpha_i + x_{it}'\beta + \epsilon_{it}, \qquad \epsilon_{it} | x_{it} \sim \Lambda(0,1), \qquad y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then the condition mean is

$$\Pr(y_{it} | x_{it}, \alpha_i, \beta) = \Lambda(\alpha_i + x_{it}'\beta)$$

in which case we cannot isolate α_i in the estimation.

❖ ***Incidental parameter problem***.

$$x_{it} \sim \mathcal{N}(\alpha_i, \sigma^2), \qquad i = 1, \dots, N,\ \ t = 1, \dots, T$$

The likelihood function

$$L\big((x_{it})_{i,t} \big| (\alpha_i)_i, \sigma\big) = \prod_t \prod_i \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x_{it} - \alpha_i}{\sigma} \right)^2 \right\}$$

and the log-likelihood

$$\mathcal{L}\big((x_{it})_{i,t} \big| (\alpha_i)_i, \sigma\big) \propto -NT \ln \sigma - \frac{1}{2} \sum \sum \left(\frac{x_{it} - \alpha_i}{\sigma} \right)^2$$

Take derivative w.r.t. α_i and σ, and set them equal zero we get

$$\hat{\alpha}_i = \frac{1}{T}\sum_t x_{it} = \bar{x}_i$$

$$\hat{\sigma}^2 = \frac{1}{NT}\sum_i\sum_t (x_{it} - \bar{x}_i)^2 \quad \rightsquigarrow \quad \underbrace{\frac{1}{T}\sum_t \left(\mathrm{E}_i(x_{it}) - \mathrm{E}_{i,t}(x_{it})\right)^2}_{\chi^2_{T-1}}$$

$$\Rightarrow \quad \mathrm{E}(\hat{\sigma}^2) = \frac{1}{T}(T-1)\sigma^2$$

- This means when $T = 2$, $\hat{\sigma}^2$ will be 50% away from the true value, even if we allow $N \to \infty$!!!

- A possible solution for the case of logit. $\Lambda(\alpha_i + x_{it}'\beta)$. The likelihood contribution is

$$\Pr(y_{it}|x_{it}, \alpha_i, \beta) = \left[\frac{\exp\{\alpha_i + x_{it}'\beta\}}{1 + \exp\{\alpha_i + x_{it}'\beta\}}\right]^{y_{it}} \left[\frac{1}{1 + \exp\{\alpha_i + x_{it}'\beta\}}\right]^{1-y_{it}}$$

Assume mutually independent ϵ_{it}. $y_i = (y_{i1}, \dots, y_{iT})$.

$$\Pr(y_i|x_i, \alpha_i, \beta) = \prod_{t=1}^{T}\left[\frac{\exp\{\alpha_i + x_{it}'\beta\}}{1 + \exp\{\alpha_i + x_{it}'\beta\}}\right]^{y_{it}} \left[\frac{1}{1 + \exp\{\alpha_i + x_{it}'\beta\}}\right]^{1-y_{it}}$$

$$= \frac{\exp\{\sum_{t=1}^{T} y_t(\alpha_i + x_{it}'\beta)\}}{\prod_{t=1}^{T}(1 + \exp\{\alpha_i + x_{it}'\beta\})}$$

$$= \frac{\exp\{\alpha_i \sum_{t=1}^{T} y_{it}\} \exp\{(\sum_{t=1}^{T} y_{it}x_{it}')\beta\}}{\prod_{t=1}^{T}(1 + \exp\{\alpha_i + x_{it}'\beta\})}$$

Condition on $\sum_t y_{it} = c$

$$\Pr\left(y_i \middle| \alpha_i, x_i, \beta, \sum_t y_{it}\right)$$

$$\Pr\left(y_i \middle| \sum_t y_{it} = c\right) = \frac{\Pr(y_i \wedge \sum_t y_{it} = c)}{\Pr(\sum_t y_{it} = c)}$$

Program Evaluation

Drop 25.4.5
Read 3.3, 3.4

- There is a program (i.e. a treatment), and we're trying to evaluate what kind of effect this treatment has.
  - Main problem with program evaluation: self-selection issue
  - In economics, it's hard to conduct a "perfect experiment" with random assignment into treatment and control groups

- Main application: labor and development
  - In the labor market programs, the self-selection issue is who chooses to enrol in these programs: maybe those who expect to find a job after these programs would actually enrol. On the other hand, it could also be that the program does have an effect on improving the chance of the enrolled finding a job later; this is the treatment effect. Essentially, program evaluation wants to disentangle these two effects.

- The model

$$y_{i,1} = \text{outcome if treated}, \qquad D_i = 1 = \text{treatment indicator}$$
$$y_{i,0} = \text{outcome if control}, \qquad D_i = 0$$

  - ***Fundamental problem of causal inference***: we can see either $y_{i,1}$ or $y_{i,0}$, i.e.

$$y_i = D_i y_{i,1} + (1 - D_i) y_{i,0}$$

N is the number of individuals. Among them N_1 have $D_i = 1$, i.e. the size of treatment group; and $N - N_1$ is the size of control group. Data is (y_i, D_i, x_i).

  - What we are interested in
    - Average treatment effect.

$$\Delta \equiv y_{i,1} - y_{i,0}$$
$$- \quad \mathrm{E}(\Delta) \quad \text{average treatment effect}$$
$$\mathrm{E}(\Delta | D_i = 1) \quad \text{average treatment effect for the treated}$$

      This is the ***average treatment effect for the treated*** (or ***ATT*** or ***ATET***)

  - Randomized Control Trial: $(y_0, y_1) \perp D$, so that outcomes are independent of treatment.

$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:D_i=1} y_i - \frac{1}{N - N_1} \sum_{i:D_i=0} y_i$$

Since $(y_0, y_1) \perp D \Rightarrow \mathrm{E}(y_0|D) = \mathrm{E}(y_0)$, we infer that

$$\begin{aligned} \mathrm{E}(y_1 - y_0) &= \mathrm{E}(y_1) - \mathrm{E}(y_0) \\ &= \mathrm{E}(y_1|D=1) - \mathrm{E}(y_0|D=0) \\ &= \frac{1}{N_1} \sum_{i:D_i=1} y_i - \frac{1}{N - N_1} \sum_{i:D_i=0} y_i, \qquad \text{[sample analog]} \end{aligned}$$

By the LLN, $\widehat{ATE}$ is a consistent estimate of $\mathrm{E}(y_1 - y_0)$.

  - Two alternative assumptions
    - Unconfoundedness (or conditional independence): $(y_0, y_1) \perp D|X$. This means that, conditional on x, outcomes are independent of treatment. But there is no test of

whether there is enough controls in the X. This implies $\mathrm{E}(y_1|D,x) = \mathrm{E}(y_1|x)$.

- Overlap (or matching assumption): $\Pr(x) = \Pr(D=1|x)$ where $\Pr(x) \in (0,1)$ for all x. This is saying that we are not sure whether or not somebody is in the treatment group for a given value of x. In other words, there are both treated and untreated individuals for each x. This way, we can compare treated and untreated individuals for any given x

To estimate the ATE,

$$\begin{aligned}\mathrm{E}(y_1 - y_0) &= \mathrm{E}_x(\mathrm{E}(y_1 - y_0|x)) \\ &= \mathrm{E}_x(\mathrm{E}(y_1|x) - \mathrm{E}(y_0|x)) \\ &= \mathrm{E}_x(\mathrm{E}(y_1|D=1,x) - \mathrm{E}(y_0|D=0,x))\end{aligned}$$

Suppose $x \in \{x_1, \dots, x_m\}$, and for $j = 1, \dots, m$, let N_{x_j} be the number of observations with $x = x_j$, and N_{1,x_j} the number of observations with $x = x_j$ and $D = 1$.

$$\widehat{ATE}(x_j) = \underbrace{\frac{1}{N_{1,x_j}} \sum_{i:D_i=1 \wedge x_i=x_j} y_i}_{\xrightarrow{p} \mathrm{E}(y_1|D=1,x=x_j)} - \underbrace{\frac{1}{N_{x_j} - N_{1,x_j}} \sum_{i:D_i=0 \wedge x_i=x_j} y_i}_{\xrightarrow{p} \mathrm{E}(y_0|D=0,x=x_j)}$$

Thus, $\widehat{ATE}(x_j)$ consistently estimates $ATE(x_j)$.

$$ATE = \mathrm{E}_x(ATE(x)) = \sum_j \Pr(x = x_j)\, ATE(x_j)$$

This motivates

$$\widehat{ATE} = \sum_j \frac{N_{x_j}}{N} \widehat{ATE}(x_j)$$

But this approach is hard to implement in practice because for each x_j we need to have a large number of observations (to have the convergence in probability). This means that our dataset needs to be extremely large.

- Propensity Score Matching (with unconfoundedness and overlap assumptions)
  - ***Propensity score***: $\Pr(x) \equiv \Pr(D=1|x)$, where $\Pr(x)$ is estimable (e.g. from probit or logit of D on x)
  - Instead of $\mathrm{E}(y_1 - y_0)$, look at

$$\begin{aligned}\mathrm{E}\left(\frac{D}{\Pr(x)} y_1 - \frac{1-D}{1-\Pr(x)} y_0\right) &= \mathrm{E}_x\left[\mathrm{E}\left(\frac{D}{\Pr(x)} y_1 - \frac{1-D}{1-\Pr(x)} y_0\right)\middle| x\right] \\ &= \mathrm{E}_x\left[\mathrm{E}_y\left(\frac{\Pr(x)}{\Pr(x)} y_1 - \frac{1-\Pr(x)}{1-\Pr(x)} y_0\right)\middle| x\right] \\ &= \mathrm{E}(y_1 - y_0)\end{aligned}$$

    Note that

$$\mathrm{E}(D|x) = 1 \cdot \Pr(D=1|x) + 0 \cdot \Pr(D=0|x) = \Pr(x)$$

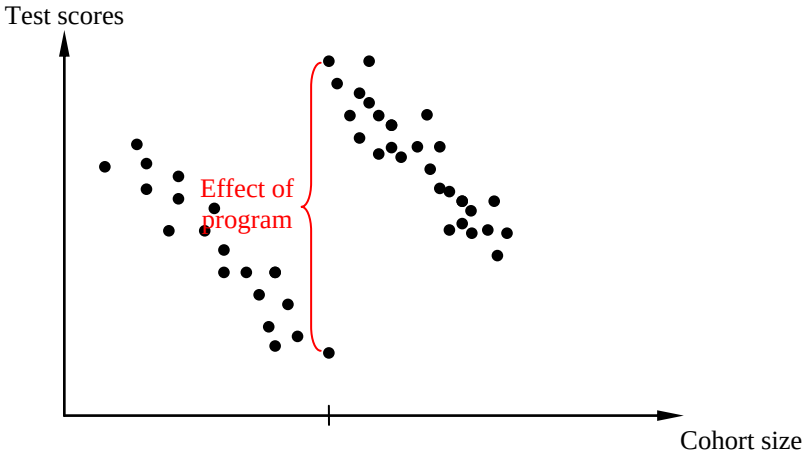
  - Suppose we have $\Pr(x)$ [we know that ($\widehat{\Pr}(x) \xrightarrow{p} \Pr(x)$)], but after logit/probit,

$$\frac{1}{N} \sum_i \frac{D_i y_i}{\widehat{\Pr}(x)} - \frac{(1-D_i) y_i}{1-\widehat{\Pr}(x)} \xrightarrow{p} \mathrm{E}\left(\frac{Dy}{\Pr(x)} - \frac{(1-D)y}{1-\Pr(x)}\right) = \cdots = ATE$$

Program Evaluation (cont'd)

C&T 25.6

❖ Regression Discontinuity (RD)



➢ Sample slightly to the left of the threshold is assumed to be almost identical to the sample slightly to the right of the threshold. So by comparing the two samples, we get the effect of the program.

➢ Mathematically,
- s is a continuous variable with observation $s_i = (s_1, \dots, s_N)$
- $\bar{s}$ is the threshold
- Construct

$$D = \begin{cases} 1 & \text{if } s \geq \bar{s}, \text{ treatment} \\ 0 & \text{if } s < \bar{s}, \text{ control} \end{cases}, \qquad \Pr(D = 1|X) \in (0,1)$$

- Let the model be

$$y_i = \beta + \alpha D_i + u_i$$

where s determines D, and s may be correlated with u (thus so may D).
- Since s fully determines D,

$$\mathrm{E}(u|D, s) = \mathrm{E}(u|s) = k(s)$$

Then the model can be rewritten as

$$y_i = \beta + \alpha D_i + k(s_i) + \underbrace{(u_i - \mathrm{E}(u_i|D_i, s_i))}_{\epsilon_i}$$

$$= \beta + \alpha D_i + k(s_i) + \epsilon_i$$

Now $\mathrm{E}(\epsilon_i|D_i, s_i) = \mathrm{E}(u_i|D_i, s_i) - \mathrm{E}(u_i|D_i, s_i) = 0$. We can use

$$k(s) = \sum_{j=0}^{J} \lambda_j s^j$$

to approximate $k(s)$. So we can do OLS on this new equation. This is called the ***control function approach***.