1 Panel Robust Variance Estimator

The sample covariance matrix becomes

$$V\left(\hat{b}\right) = \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{X}\right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' \hat{u}_i \hat{u}_i' \tilde{X}_i\right) \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{X}\right)^{-1} \tag{1}$$

and its associated t-statistic becomes

$$t_{\hat{b}} = \frac{\hat{b}}{\sqrt{\left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{X}\right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' \hat{u}_i \hat{u}_i' \tilde{X}_i\right) \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{X}\right)^{-1}}} \tag{2}$$

Consider two regressors: First let

$$\xi_i = X_i' \hat{u}_i = [x_{1,i}\hat{u}_i \quad x_{2,i}\hat{u}_i]$$

where

$$x_{k,i} = (x_{k,i1}, ..., x_{k,iT})'$$

Then calculate $\sum_{i=1}^{N} \xi_i' \xi_i$ which is $T \times T$ matrix.

Read Lecture note in Econometric I and find out the potential issue on this panel robust variance estimator.

2 Monte Carlo Studies

2.1 Why Do We need MC?

1. Verify asymptotic results. If an econometric theory is correct, the asymptotic results should be replicatable by means of Monte Carlo studies.
   (a) Large sample theory: T or N must be very large. At least $T = 500$.
   (b) Generalize assumptions. See if a change in an assumption makes any difference in asymptotic results.
2. Examine finite sample performance. In finite sample, asymptotic results are just approximation. We don't know if or not an econometric theory works well in the finite sample.
   (a) Useful to compare with various estimators.
   (b) MSE and Bias become important to the estimation methods.
   (c) Size and Power become issues on various testing procedures & covariance estimation.

2.2 How to do MC

1. Need a data generating process (DGP), and distributional assumption.
   (a) DGP depends on an econometric theory and its assumptions.
   (b) Need to generate pseudo random variables from a certain distribution

2.2.1 Example 1: Verifying asymptotic result of OLSE

DGP:

$$\text{Model: } y_i = a + x_i\beta + u_i$$

Now we take a particular case like

$$u_i \sim iidN(0,1), \quad x_i \sim iidN(0, I_k)$$

where $a = \beta = 0$.

Step by Step procedure

1. Find out the parameters of interest. (here we are interested in consistency of OLSE)
2. Generate n pseudo random variables of u, x and y. Since $a = \beta = 0$, $y_i = u_i$.
3. Calculate OLSE for β and a. (plus the estimates of parameters of interest)
4. Repeat 2 and 3 S times. record all $\hat{\beta}$.
5. calculate mean of $\hat{\beta}$ and variance of them. (how do we know the convergence rate?)
6. Repeat 2-5 by changing n.

2.2.2 Example 2: Verifying asymptotic result of OLSE Testing

DGP:

$$\text{Model: } y_i = a + x_i\beta + u_i$$

Now we take a particular case like

$$u_i \sim iidN(0,1), \quad x_i \sim iidN(0, I_k)$$

where $a = \beta = 0$.

Step by Step procedure

1. Find out the parameters of interest. ($t-$statistic)
2. Generate n pseudo random variables of u, x and y. And calculate t ratio for β and a.
3. Repeat 2 and 3 S times. record all $t_{\hat{\beta}}$.
4. Sort $t_{\hat{\beta}}$ and find out the lower and upper 2.5% values. Compare them with the asymptotic critical value.
5. Repeat 2-4 by changing n.

2.2.3 Exercise 1: Use NW estimator and calculate t ratio. Compare the size and power of the tests (ordinary and NW t-ratios)

Asymptotic theory: Both of them are consistent. The ordinary t ratio becomes more efficient. Why?

Size of the test Change step 4 in Example 2 as follows:

Let

$$\mathbf{t}^* = \left|\hat{\mathbf{t}}_{\beta}\right|$$

sort t^*. Find when $t_j^* > 1.96$. And $1 - j^*/S$ becomes the size of the test.

Power of the test Change $\beta = 0.01$, 0.05, 0.1, 0.2.

Repeat the above procedures, and find $1 - j^*/S$. This becomes the power of the test.

2.2.4 Exercise 2: Re-do Bertrand et al.

3 Review Asymptotic Theory

3.1 Most Basic Theory

$$y_i = \beta x_i + u_i$$

where

$$u_i \sim iid\left(0, \sigma_u^2\right)$$

$$\hat{\beta} = \beta + \left(x'x\right)^{-1} x'u = \beta + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2} = \beta + \frac{\frac{1}{n}\sum_{i=1}^n x_i u_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}$$

First let

$$\frac{1}{n}\sum_{i=1}^n x_i u_i = \frac{1}{n}\sum_{i=1}^n \xi_i \to^d N\left(0, \frac{\sigma_\xi^2}{n}\right)$$

Hence we have

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n \xi_i \to^d N\left(0, n\frac{\sigma_\xi^2}{n}\right)$$

or

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \to^d N\left(0, \sigma_\xi^2\right)$$

Next,

$$\text{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^n x_i^2 = Q_x, \text{ let say.}$$

Then

$$\hat{\beta} - \beta = \frac{\frac{1}{n}\sum_{i=1}^n x_i u_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}$$

or

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i u_i}{\frac{1}{n}\sum_{i=1}^n x_i^2} \to^d N\left(0, Q_x^{-1}\sigma_\xi^2 Q_x^{-1}\right)$$

3.2 Addition Constant term

$$y_i = a + \beta x_i + u_i$$

where

$$x_i = a_x + x_i^o, \quad y_i = a_y + y_i^o.$$

$$u_i \sim iid\left(0, \sigma_u^2\right)$$

$$\hat{\beta} = \beta + \left(\tilde{x}'\tilde{x}\right)^{-1}\tilde{x}'\tilde{u} = \beta + \frac{\sum_{i=1}^n \tilde{x}_i\tilde{u}_i}{\sum_{i=1}^n \tilde{x}_i^2} = \beta + \frac{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{u}_i}{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i^2}$$

First let

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{u}_i &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(u_i-\frac{1}{n}\sum_{i=1}^{n}u_i\right)=\frac{1}{n}\sum_{i=1}^{n}x_i^o u_i^o-\frac{1}{n^2}\left(\sum x_i^o\right)\left(\sum u_i^o\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^o u_i^o-\left(\frac{1}{n}\sum x_i^o\right)\left(\frac{1}{n}\sum u_i^o\right)=\frac{1}{n}\sum_{i=1}^{n}\xi_i+O_p\left(\frac{1}{\sqrt{n}}\right)O_p\left(\frac{1}{\sqrt{n}}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\xi_i+O_p\left(\frac{1}{n}\right)
\end{aligned}$$

Hence we have

$$\begin{aligned}
\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\tilde{\xi}_i &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\xi_i+\sqrt{n}\left(\frac{1}{n}\sum x_i^o\right)\left(\frac{1}{n}\sum u_i^o\right)\\
&\rightarrow {}^{d}N\left(0,n\frac{\sigma_\xi^2}{n}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)=N\left(0,\sigma_\xi^2\right).
\end{aligned}$$

Next,

$$\text{plim}_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^2=Q_x,\ \text{let say.}$$

Then

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{x}_i\tilde{u}_i}{\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^2}\rightarrow^{d}N\left(0,Q_x^{-1}\sigma_\xi^2Q_x^{-1}\right).$$

4 Power of the Test (Local Alternative Approach)

Consider the model

$$y_i = \beta x_i + u_i$$

and under the null hypothesis, we have

$$\beta = \beta_o$$

Now we want to analyze the power of the test asymptotically. Under the alternative, we have

$$\beta = \beta_o + c$$

where $c \neq 0$.

Suppose that we are interested in comparing two estimates, let say OLSE and FGLSE ($\hat{\beta}_1$ and $\hat{\beta}_2$). Then we have

$$\frac{\sqrt{n}\left(\hat{\beta}_1 - \beta\right)}{\sqrt{V\left(\hat{\beta}_1\right)}} \rightarrow^d N(0,1) + O_p\left(N^{-1/2}\right)$$

or

$$\frac{\sqrt{n}\left(\hat{\beta}_1 - \beta_o\right)}{\sqrt{V\left(\hat{\beta}_1\right)}} \rightarrow^d N(0,1) + \sqrt{n}c + O_p\left(N^{-1/2}\right)$$

Hence as long as $c \neq 0$, the power of the test goes to one. In other words, the dominant term becomes the second term ($\sqrt{n}c$)

Similary, we have

$$\frac{\sqrt{n}\left(\hat{\beta}_2 - \beta_o\right)}{\sqrt{V\left(\hat{\beta}_2\right)}} \rightarrow^d N(0,1) + \sqrt{n}c + O_p\left(N^{-1/2}\right)$$

Hence we can't compare two tests.

Now, to avoid this, let

$$\beta = \beta_o + \frac{c}{\sqrt{n}}$$

so that $\beta \rightarrow \beta$ as $n \rightarrow \infty$. Then we have

$$\frac{\sqrt{n}\left(\hat{\beta}_\kappa - \beta\right)}{\sqrt{V\left(\hat{\beta}_\kappa\right)}} \rightarrow^d N(c,1) + O_p\left(N^{-1/2}\right).$$

Hence depending on the value of c, we can compare the power of the test (across different estimates).

5 Panel Regression

5.1 Regression Types

1. Pooled OLS estimator (POLS)

$$y_{it} = a + \beta x_{it} + \gamma z_{it} + u_{it}$$

2. Least squares dummy variables (LSDV) or Withing group (WG) or Fixed effects (FE) estimator

$$y_{it} = a_i + \beta x_{it} + \gamma z_{it} + u_{it}$$

3. Random Effect (RE) or PFGLS estimator

$$y_{it} = a + \beta x_{it} + \gamma z_{it} + e_{it}, \quad e_{it} = a_i - a + u_{it}$$

Let $X = (x_{11}, x_{12}, ..., x_{1T}, x_{21}, ..., x_{NT})'$, $\mathbf{x}_i = (x_{i1}, ..., x_{iT})'$, $\mathbf{x}_t = (x_{1t}, ..., x_{Nt})'$. Define Z, $\mathbf{z}_i$ and $\mathbf{z}_t$ in the similar way. Let $W = (X\ Z)'$. Then

5.2 Covariance estimators:

1. Ordinary estimator: $\hat{\sigma}_u^2 (W'W)^{-1}$
2. White estimator
   (a) Cross sectional heteroskedasticity: $NT (W'W)^{-1} \left(\frac{1}{N} \sum_{i=1}^{n} \hat{\mathbf{u}}_i^2 \mathbf{w}_i' \mathbf{w}_i\right) (W'W)^{-1}$
   (b) Time series heteroskedasticity: $NT (W'W)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{u}}_t^2 \mathbf{w}_t' \mathbf{w}_t\right) (W'W)^{-1}$
   (c) Cross and Time heteroskedasticity: $NT (W'W)^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2 \mathbf{w}_{it} \mathbf{w}_{it}'\right) (W'W)^{-1}$
3. Panel Robust Covariane estimator: $N (W'W)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{w}_i\right) (W'W)^{-1}$
4. LRV estimator ? Why not?

5.3 Pooled GLS Estimators

$$\hat{\delta} = \left[W' \left(\Omega^{-1} \otimes I\right) W\right]^{-1} \left[W' \left(\Omega^{-1} \otimes I\right) y\right]$$

5.3.1 How to estimate Ω :

1. Time Series Correlation:
   (a) AR1: easy to extend. $\Omega = \begin{bmatrix} 1 & & \rho^{T-1} \\ \vdots & \ddots & \vdots \\ \rho^{T-1} & & 1 \end{bmatrix}$
   (b) Unknown. $\hat{\Omega}_{sh} = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_{is}\hat{u}_{ih}$. Required small T and large N.
2. Cross sectional correlation
   (a) Spatial: Easy.
   (b) Unknown. $\hat{\Omega}_{sh} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{st}\hat{u}_{ht}$

5.4 Seemingly Unrelated Regression

$$\hat{\delta} = \left[W'\left(I \otimes \Omega^{-1}\right)W\right]^{-1}\left[W'\left(I \otimes \Omega^{-1}\right)y\right]$$

6 Bootstrap

Reference: "The BOOTSTRAP" by Joel L. Horowitz (Chapter 52 in Handbook of Econometrics Vol 5)

6.1 What is the bootstrap

It is a method for estimating the distribution of an estimator or test statistics by resampling the data.

Example 1 (Bias correction) Model

$$y_t = a + \rho y_{t-1} + e_t,$$

where e_t is a white noise process. It is well known that $\mathrm{E}(\hat{\rho} - \rho) = -\dfrac{1+3\rho}{T} + O\left(T^{-2}\right)$. Here I am explaining how to reduce Kendall bias (not eliminating) by using the following bootstrap procedure.

1. Estimate OLSE for a and ρ, denote them as $\hat{a}$ and $\hat{\rho}$. Get OLS residual $\hat{e}_t = y_t - \hat{a} - \hat{\rho} y_{t-1}$.

2. generate $T + K$ random variables from the uniform distribution of $U(1, T-1)$. Make them as integers.

   ind = rand(t+k,1)*(t-1); % generate from U(0,T-1).
   ind = 1+floor(ind); % make integers. 0.1 => 1.

3. Draw $(T+K) \times 1$ vector of e_t^* from $\hat{e}_t$.

   esta = e(ind,:);

4. Recentering e_t^* to make its mean be zero. Generate pseudo y_t^* from e_t^*, and discard the first K obs.

```
esta  = esta - mean(esta); ysta = esta;                  1
        for i=2:t+k;                                     2
        ysta(i,:) = ahat+rhohat*ysta(i-1,:) + esta(i,:); 3
        end;                                             4
ysta  =ysta(k+1:t+k,:);                                  5
```

Note that ahat should not be inside for statement. Add ahat in line 5. Precisely speaking, you have to add ahat*(1-rho) but don't need to do so if you use demeaned series to estimate rho.

5. Estimate $\hat{a}^*$ and $\hat{\rho}^*$ with y_t^*.

6. Repeat step 2 and 5 M times.

7. Calculate the sample mean of $\hat{\rho}^*$. Calculate the bootstrap bias, $B = \frac{1}{M}\sum_{m=1}^{M}\hat{\rho}_m^* - \hat{\rho}$ where $\hat{\rho}_m^*$ is the mth time bootstrapped point estimate of ρ. Subtract B from $\hat{\rho}$.

$$\hat{\rho}_{\text{mue}} = \hat{\rho} - B$$

where mue stands from mean unbiased estimator. Note that

$$E\left(\hat{\rho}_{\text{mue}} - \rho\right) = O\left(T^{-2}\right).$$

8. For $t-$statistics: Construct

$$t_{\rho,m}^* = \frac{\hat{\rho}_m^* - \hat{\rho}}{\sqrt{V\left(\hat{\rho}_m^*\right)}}$$

and then repeat M times and get the 5% critical value of $t_{\rho,m}^*$. Compare this with $t_\rho = \frac{\hat{\rho}}{\sqrt{V(\hat{\rho})}}$.

6.2 How the bootstrap works

First let the estimates be a function of T. For example, $\hat{\rho}$ be $\hat{\rho}_T$. Now define

$$\hat{\rho}_T = \frac{\sum \tilde{y}_{it}\tilde{y}_{it-1}}{\sum \tilde{y}_{it-1}^2} = g\left(z\right), \text{ let say}$$

where z is a 2×1 vector. That is, $z = (z_1, z_2)$ and $z_1 = \frac{1}{T}\sum \tilde{y}_{it}\tilde{y}_{it-1}$ and $z_2 = \frac{1}{T}\sum \tilde{y}_{it-1}^2$.

From A Tyalor expansion (or Delta method), we have

$$\hat{\rho}_T = \rho + \frac{\partial g}{\partial z}\left(z - z_o\right) + \frac{1}{2}\left(z - z_o\right)'\left(\frac{\partial^2 g}{\partial z \partial z'}\right)\left(z - z_o\right) + O_p\left(T^{-2}\right)$$

Now taking expectations yields

$$\begin{aligned} E\left(\hat{\rho}_T - \rho\right) &= E\frac{\partial g}{\partial z}\left(z - z_o\right) + \frac{1}{2}E\left(z - z_o\right)'\left(\frac{\partial^2 g}{\partial z \partial z'}\right)\left(z - z_o\right) + O\left(T^{-2}\right) \\ &= \frac{1}{2}E\left(z - z_o\right)'\left(\frac{\partial^2 g}{\partial z \partial z'}\right)\left(z - z_o\right) + O\left(T^{-2}\right) \end{aligned}$$

since $E(z - z_o) = 0$ always.

The first term in the above becomes $O\left(T^{-1}\right)$, that is $-\dfrac{1+3\rho}{T}$. We want to eliminate this part (not reduce it). The bootstrapped $\hat{\rho}_T^*$ becomes

$$\hat{\rho}_T^* = \hat{\rho}_T + \frac{\partial g}{\partial z}(z^* - z_o) + \frac{1}{2}(z^* - z_o)' \left(\frac{\partial^2 g}{\partial z \partial z'}\right)(z^* - z_o) + O_p\left(T^{-2}\right)$$

where $z^* = (z_1^*,\ z_2^*)$, and $z_1^* = \frac{1}{T}\sum \tilde{y}_{it}^* \tilde{y}_{it-1}^*$, etc. Note that we generate y_{it}^* from $\hat{\rho}_T$, $\hat{\rho}_T^*$ can be expanded around $\hat{\rho}_T$ not around the true value of ρ. Now taking expectation E^* in the sense that

$$E^* \to E \text{ as } M, T \to \infty.$$

Then we have

$$\begin{aligned} E^*(\hat{\rho}_T^* - \hat{\rho}_T) &= \frac{1}{2} E^*(z^* - z_o)' \left(\frac{\partial^2 g}{\partial z \partial z'}\right)(z^* - z_o) + O\left(T^{-2}\right) \\ &= B^* \end{aligned}$$

Note that in general

$$B^* = B + O\left(T^{-2}\right)$$

hence we have

$$\hat{\rho}_{\text{mue}} = \hat{\rho}_T - B^* = \hat{\rho}_T - E^*(\hat{\rho}_T^* - \hat{\rho}_T)$$

6.3 Bootstrapping Critical Value

Example 2. (Using the same example 1) Generate t-ratio for $\hat{\rho}_m^*$ M times. Sort them, and find 95% critical value from the bootstrapped t-ratio. Compare it with the actual t-ratio.

Asymptotic Refinement Notation:

F_0 is the true cumulative density function. For an example, cdf of normal distribution.

t_β is the t-statistic of β.

$t_{n,\beta}$ is the sample t-statistic of $\hat{\beta}$ where n is the sample size.

$G(\tau, F_0) = P(t_\beta \leq \tau)$. That is the function G is the true CDF of t_β.

$G_n(\tau, F_0) = P(t_{n,\beta} \leq \tau)$. The function G_n is the exact finite sample CDF of $t_{n,\beta}$

Asymptotically $G_n \to G$ as $n \to \infty$. Denote that $G_n(\tau, F_n)$ is the bootstrapped function for $t_{n,\beta}^*$ where F_n is the finite sample CDF.

Definition: Pivotal statistics If $G_n(\tau, F_0)$ does not depend on F_0, then $t_{n,\beta}$ is said to be pivotal.

Example 3 (exact finite sample CDF for AR(1) with a unknown constant) From Tanaka (1983, Econometrica), the exact finite sample CDF for $t_{\hat{\rho}}$ is given by

$$P(t_{T,\hat{\rho}} \leq x) = \Phi(x) + \frac{\phi(x)}{\sqrt{T}} \frac{2\rho + 1}{\sqrt{1-\rho^2}} + O\left(T^{-1}\right)$$

where Φ is the CDF of normal distribution and ϕ is PDF of normal. Here Tanaka assumes F_0 is normal. That is, y_t is distributed as normal. Of course, if y_t has a different distribution, the exact finite sample PDF is unknown. However, $t_{T,\hat{\rho}}$ is pivotal since as $T \to \infty$, its limiting distribution goes to $\Phi(x)$.

Now under some regularity conditions (see Theorem 3.1 Horowitz), we have

$$G_n(\tau, F_0) = G(\tau, F_0) + \frac{1}{\sqrt{n}} g_1(\tau, F_0) + \frac{1}{n} g_2(\tau, F_0) + \frac{1}{n^{3/2}} g_3(\tau, F_0) + O\left(n^{-2}\right)$$

uniformly over τ.

Meanwhile the bootstrapped $t_{\hat{\beta}}$ has the following properties

$$G_n(\tau, F_n) = G(\tau, F_n) + \frac{1}{\sqrt{n}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_n) + \frac{1}{n^{3/2}} g_3(\tau, F_n) + O\left(n^{-2}\right)$$

When $t_{n,\hat{\beta}}$ is not a pivotal statistic In this case, we have

$$G_n(\tau, F_0) - G_n(\tau, F_n) = [G(\tau, F_0) - G(\tau, F_n)] + \frac{1}{\sqrt{n}} [g_1(\tau, F_0) - g_1(\tau, F_n)] + O\left(n^{-1}\right)$$

Note that $G(\tau, F_0) - G(\tau, F_n) = O\left(n^{-1/2}\right)$. Hence the bootstrap makes an error of size $O\left(n^{-1/2}\right)$. Also note that $G_n(\tau, F_0)$ also makes an error of size $O\left(n^{-1/2}\right)$, so that the bootstrap does not reduce (neither increase) the size of the error.

When $t_{n,\hat{\beta}}$ is a pivotal In this case, we have

$$G(\tau, F_0) - G(\tau, F_n) = 0$$

by definition. Then we have

$$G_n(\tau, F_0) - G_n(\tau, F_n) = \frac{1}{\sqrt{n}} [g_1(\tau, F_0) - g_1(\tau, F_n)] + O\left(n^{-1}\right)$$

and $g_1(\tau, F_0) - g_1(\tau, F_n) = O\left(n^{-1/2}\right)$. Hence we have

$$G_n(\tau, F_0) - G_n(\tau, F_n) = O\left(n^{-1}\right),$$

which implies that the bootstrap reduces the size of an error.

6.4 Exercise: Sieve Bootstrap

(Read Li and Maddala, 1997)

Consider the following cross sectional regression

$$y_{it} = a + \beta x_{it} + u_{it} \tag{3}$$

We want to test the null hypothesis of $\beta = 0$. We suspect that x_{it} and u_{it} are serially correlated, but not cross correlated. Consider the following sieve bootstrap procedure

1. Run (3) and get $\hat{a}, \hat{\beta}$, and $\hat{u}_{it}$.
2. Run the following regression

$$\begin{bmatrix} x_{it} \\ u_{it} \end{bmatrix} = \begin{bmatrix} \mu_x \\ 0 \end{bmatrix} + \begin{bmatrix} \rho_x & 0 \\ 0 & \rho_u \end{bmatrix} \begin{bmatrix} x_{it} \\ u_{it} \end{bmatrix} + \begin{bmatrix} e_{it} \\ \varepsilon_{it} \end{bmatrix}$$

   and get $\hat{\mu}_x, \hat{\rho}_x, \hat{\rho}_u$ and their residuals of $\hat{e}_{it}$ and $\hat{\varepsilon}_{it}$. Recentering them.
3. Generate pseudo x^*_{it} and u^*_{it}.

```
ind = rand(t+k,1)*(t-1); % generate from U(0,T-1).
ind = 1+floor(ind); % make integers. 0.1 => 1.
F = [ehat espi]; % êit and ε̂it
Fsta = F(ind,:);  % use the same ind. Important!
repeat what you learnt before....
```

4. Generate y^*_{it} under the null,

$$y^*_{it} = \hat{a} + u^*_{it}.$$

5. Run (3) with y^*_{it} and x^*_{it}, and get the bootstrapped critical value.

Simplest Case: Consider you want to test

$$y_{jit} = a_j + u_{jit}, \quad u_{jit} = \rho_j u_{jit-1} + e_{jit} \tag{4}$$

where j stands for the jth treatment. Assume u_{jit} are cross sectionally dependent and serially correlated. However u_{jit} is exogenous. Then running the following panel AR(1) regression beceoms useless to test $H_0 : a_j = a$ for all j.

$$y_{jit} = \alpha_j + \rho_j y_{jit-1} + e_{jit}$$

since $\alpha_j = a_j \left(1 - \rho_j\right)$. In this case, one should run (4) and do a seive bootstrap with $\hat{u}_{jit}$.

7 Maximum Likelihood Estimation

7.1 The likelihood function

Let $y_1, ..., y_n$, $\{y_i\}$, be a sequence of random variables which has iid$N\left(\mu, \sigma^2\right)$. Its probability density function $f\left(y|\mu, \sigma^2\right)$ can be written as

$$f\left(y = y_i|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right].$$

Note that this pdf states that with given μ and σ^2, the probability of $y_i = y$. Now let think about the joint density of y such that

$$\begin{aligned} f\left(y_1, ..., y_n|\mu, \sigma^2\right) &= f\left(y_1|\mu, \sigma^2\right) \times \cdots \times f\left(y_n|\mu, \sigma^2\right) \\ &= \prod_{i=1}^{n} f\left(y_i|\mu, \sigma^2\right) \text{ due to independence} \end{aligned}$$

That is, with given μ and σ^2, the joint pdf states that the probability of a sequence of y to be $\{y_i\}$. This concept is very useful when we do both/or MC and bootstrap.

Now consider the mirror image case. Given $\{y_i\}$, what are the most probable estimates for μ and σ^2? To answer this question, we consider the likelihood (probability) of μ and σ^2. Let $\theta = \left(\mu, \sigma^2\right)$. Then we can re-interpret the joint pdf as the likelihood function. That is,

$$f\left(y|\theta\right) = L\left(\theta|y\right).$$

And then we maximize the likelihood with given $\{y_i\}$.

$$\arg\max_{\theta} L\left(\theta|y\right).$$

However it is often difficult to maximize directly L function due to nonlinearity. Hence alternatively we maximize the log likelihood

$$\arg\max_{\theta} \ln L\left(\theta|y\right).$$

In practice (computer programming) it is much easier to minimize the negative log likelihood such that

$$\arg\min_{\theta} - \ln L\left(\theta|y\right)$$

Of course, we have to get the first order conditions with respect to θ, and find the optimal values of θ.

Example 1 Normal random variables. $\{y_i\}$, $i = 1, ..., n$. Want to estimate μ and σ^2.

$$\begin{aligned} L\left(\mu, \sigma^2 | y\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right] \\ &= \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2\right] \end{aligned}$$

since

$$\exp\left(a\right)\exp\left(b\right) = \exp\left(a + b\right).$$

Hence

$$\ln L = -\frac{n}{2}\ln\left(2\pi\right) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{\left(y_i - \mu\right)^2}{\sigma^2}\right]$$

Note that

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right) = 0, \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \mu\right)^2 = 0 \end{aligned}$$

From this, we have

$$\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_{mle}\right)^2$$

Properties of an MLE (Theorem 16.1 Green)

1. Consistency: $\hat{\theta}_{mle} \rightarrow^p \theta$
2. Asymptotic normality
$$\hat{\theta}_{mle} \rightarrow^d N\left(\theta, \left[I\left(\theta\right)^{-1}\right]\right)$$
where
$$I\left(\theta\right) = -E\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right] = -E\left(H\right)$$
where H is the Hessian matrix.
3. Asymptotic Efficiency: $\hat{\theta}_{mle}$ is asymptotically efficient and achieves the Cramer-Rao lower bound for consistent estimators.

8 Method of Moments (Chap 15)

Consider moment conditions such that

$$E\left(\xi_t - \mu\right) = 0$$

where ξ_t is a random variable and μ is the unknown mean of ξ_t. The parameter of interest, here, is μ. Consider the following minimum criteria given by

$$\arg\min_{\mu} V_T = \arg\min_{\mu} \frac{1}{T}\sum_{t=1}^{T}\left(\xi_t - \mu\right)^2$$

which becomes the minimum variance of ξ_t with respect to μ. Of course, the simple solution becomes the sample mean for μ since we have

$$\frac{\partial V_T}{\partial \mu} = -2\frac{1}{T}\sum_{t=1}^{T}\left(\xi_t - \mu\right) = 0, \quad \Longrightarrow \frac{1}{T}\sum_{t=1}^{T}\xi_t = \mu$$

The above case is the simple example of the method of moment(s).

Now consider more moments such that

$$\begin{aligned} E\left(\xi_t - \mu\right) &= 0 \\ E\left[\left(\xi_t - \mu\right)^2 - \gamma_0\right] &= 0 \\ E\left[\left(\xi_t - \mu\right)\left(\xi_{t-1} - \mu\right) - \gamma_1\right] &= 0 \\ E\left[\left(\xi_t - \mu\right)\left(\xi_{t-2} - \mu\right) - \gamma_2\right] &= 0 \end{aligned}$$

Then we have the four unknowns: $\mu, \gamma_0, \gamma_1, \gamma_2$. We have four sample moments such that

$$\frac{1}{T}\sum_{t=1}^{T}\xi_t, \frac{1}{T}\sum_{t=1}^{T}\xi_t^2, \frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_{t-1}, \frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_{t-2}$$

so that we can solve this numerically.

However, we want to impose further restriction. Suppose that we assume ξ_t follows AR(1) process. Then we have

$$\gamma_1 = \rho\gamma_0, \quad \gamma_2 = \rho\gamma_0$$

so that the total number of unknowns is reducing to three (γ_0, ρ, μ). We can increase more cross moment conditions also. Let $\psi_T = \left(\frac{1}{T}\sum_{t=1}^{T}\xi_t, \frac{1}{T}\sum_{t=1}^{T}\xi_t^2, \frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_{t-1}, \frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_{t-2}\right)'$. Then we have

$$E\frac{1}{T}\sum_{t=1}^{T}\left(\xi_t - \mu\right)^2 = E\frac{1}{T}\sum_{t=1}^{T}\xi_t^2 - \mu^2 = \gamma_0$$

so that

$$E\frac{1}{T}\sum_{t=1}^{T}\xi_t^2 = \gamma_0 - \mu^2$$

Also note that

$$E\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_{t-1} = \rho\gamma_0 - \mu^2, \text{ and so on.}$$

Hence we may consider the following estimation

$$\arg\min_{\mu,\rho,\gamma_0}\left[\psi_T - \psi\left(\theta\right)\right]'\left[\psi_T - \psi\left(\theta\right)\right]. \tag{5}$$

where θ is the parameters of interest (true parameters, μ, γ_0, ρ). The resulting estimator is called 'method of moments estimator'. Note that MM estimator is a kind of minimum distance estimators.

In general, MM estimator can be used in many cases. However, this method has one weakness. Suppose that the second moment is relatively huge than the first moment. Since V_T function assigns the same weight across moments, the minimum problem in (5) tries to minimize the second moment rather than the first and second moment both. Hence we need to design the optimal weighted method of moments, which becomes generalized method of moments (GMM).

To understand the nature of GMM, we have to study the asymptotic properties of MM estimator. (in order to find the optimal weighting matrix). Now to get the asymptotic distribution of $\hat{\theta}$, we need a Taylor expansion.

$$\psi_T = \psi\left(\theta\right) + \frac{\partial\psi_T\left(\theta\right)}{\partial\theta'}\left(\hat{\theta} - \theta\right) + O_p\left(\frac{1}{T}\right)$$

so that we have

$$\sqrt{T}\left(\hat{\theta} - \theta\right) = \sqrt{T}\left[\psi_T - \psi\left(\theta\right)\right]G\left(\theta\right)^{-1} + O_p\left(\frac{1}{\sqrt{T}}\right)$$

where $G_T\left(\theta\right) = \frac{\partial\psi_T(\theta)}{\partial\theta'}$. Note that we know that

$$\sqrt{T}\left[\psi_T - \psi\left(\theta\right)\right] \rightarrow^d N\left(0, \Phi\right)$$

Hence we have

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \rightarrow^d N\left(0, G\left(\theta\right)^{-1}\Phi G\left(\theta\right)'^{-1}\right)$$

where $G_T\left(\theta\right) \rightarrow^p G\left(\theta\right)$.

8.1 GMM

First consider infeasible generalized version of method of moments.

$$\arg\min_{\mu,\rho,\gamma_0} \left[\psi_T - \psi(\theta)\right]' \Phi^{-1} \left[\psi_T - \psi(\theta)\right].$$

where Φ is true unknown weighting matrix. Now feasible version becomes

$$\arg\min_{\mu,\rho,\gamma_0} \left[\psi_T - \psi(\theta)\right]' \mathbf{W}_T \left[\psi_T - \psi(\theta)\right] n = \arg\min_{\mu,\rho,\gamma_0} G_T(\theta)' \mathbf{W}_T G_T(\theta)$$

where $\mathbf{W}_T$ is a consistent estimator of Φ^{-1}. Let

$$V_T = \left[\psi_T - \psi(\theta)\right]' \mathbf{W}_T \left[\psi_T - \psi(\theta)\right]$$

Then GMM estimator satisfies

$$\frac{\partial V_T\left(\hat{\theta}_{GMM}\right)}{\partial \hat{\theta}_{GMM}} = 2G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T \left[\psi_T - \psi\left(\hat{\theta}_{GMM}\right)\right] = 0$$

so that we have

$$\psi\left(\hat{\theta}_{GMM}\right) = \psi_T(\theta) + G_T(\theta)\left(\hat{\theta}_{GMM} - \theta\right) + O_p\left(\frac{1}{T}\right)$$

Thus

$$\begin{aligned} & G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T \left[\psi_T - \psi\left(\hat{\theta}_{GMM}\right)\right] \\ = \; & G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T \left[\psi_T - \psi\left(\hat{\theta}_{GMM}\right)\right] + G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T G_T(\theta)\left(\hat{\theta}_{GMM} - \theta\right) = 0 \end{aligned}$$

Hence

$$\left(\hat{\theta}_{GMM} - \theta\right) = -\left\{G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T G_T(\theta)\right\}^{-1} G_T\left(\hat{\theta}_{GMM}\right)' \mathbf{W}_T \left[\psi_T - \psi\left(\hat{\theta}_{GMM}\right)\right]$$

and

$$\sqrt{T}\left(\hat{\theta}_{GMM} - \theta\right) \to^d N(0, V)$$

where

$$V = \frac{1}{T}\left\{G'\mathbf{W}G\right\}^{-1} G'\mathbf{W}\mathbf{\Phi}\mathbf{W}G\left\{G'\mathbf{W}G\right\}^{-1}.$$

When $W = \Phi^{-1}$, then we have

$$V = \frac{1}{T}\left\{G'\mathbf{\Phi}^{-1}G\right\}^{-1} G'\mathbf{\Phi}^{-1}G\left\{G'\mathbf{\Phi}^{-1}G\right\}^{-1} = \frac{1}{T}\left\{G'\mathbf{\Phi}^{-1}G\right\}^{-1}.$$

Overidentifying Restriction can be tested by calculating the following statistics

$$J = \left[\psi_T - \psi(\theta^*)\right]' \mathbf{W}_T \left[\psi_T - \psi(\theta^*)\right] \to^d \chi^2_{l-k}$$

where l is the total number of moments and k is the total number of parameters to estimate. The null hypothesis is that with given estimates, all moment conditions considered are valid. Once the overidentifying restriction is not rejected, the GMM estimates become robust.

9 Sample Midterm Exam

Model

$$y_{it} = a_i + \beta x_{it} + u_{it}, \quad t = 1, ..., T; i = 1, ..., N \tag{6}$$

$$u_{it} = \rho u_{it-1} + v_{it}, \quad x_{it} = \rho x_{it-1} + e_{it} \text{ for time series and panel cases} \tag{7}$$

where v_{it},e_{it} are independent each other.

9.1 Matlab Exercise:

1. Estimators
   (a) Cross section: Let $t = 1$, $N = n$. Provide matlab codes for OLS, WLS (weighted least squares)
   (b) Time series: Let $N = 1, T = T$. Provide matlab codes for OLS.
   (c) Panel data: Provide matlab codes for POLS, LSDV, PGLS (infeasible GLS)
2. t-statistics
   (a) Cross section: provide t ratios for ordinary and white.
   (b) Time series: provide t ratios for ordinary and NW.
   (c) Panel Data: provide t ratios for ordinary and panel robust.
3. Monte Carlo Study. Assume all innovations are iidN(0,1). (Don't need to write up matlab codes)
   (a) want to show that $\hat{\beta}_{LSDV}$ is inconsistent. Write down how you can do by means of MC.
   (b) want to show that $t_{\hat{\beta}} = \hat{\beta}_{LSDV} / \sqrt{\hat{\sigma}_u^2 / \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \frac{1}{T} \sum_{t=1}^{T} x_{it} \right)^2 \right\}}$ suffers from size distortion. Write down step by step procedure how you can show it by means of MC.
4. Bootstrap.
   (a) write up the bootstrap procedure (step by step) how to construct the bootstrapped critical value for $t_{\hat{\beta}}$ in 3.b. under the null hypothesis $\beta = 0$.

9.2 Theory

1. Basic: Derive the limiting distribution of $\hat{\beta}_{LSDV}$ in (1) and (2)

2. Suppose that

$$u_{it} = \theta_t + \varepsilon_{it} \tag{8}$$

where θ_t is independent from x_{it}.

   (a) You run eq. (1). (y on a_i and x_{it}). Show $\hat{\beta}_{LSDV}$ is consistent.

   (b) Further assume that ε_{it} is a white noise but θ_t follows an AR(1) process. How can you obtain more efficient estimator by using a simple transformation. (Don't think about MLE)

3. Now we have

$$u_{it} = \lambda_i \theta_t + \varepsilon_{it}$$

   (a) Show that $\hat{\beta}_{LSDV}$ is still consistent as long as u_{it} is independent from x_{it}.

   (b) Can you eliminate $\lambda_i \theta_t$? If so, how?

4. DGP is given by

$$y_{it} = a + \beta x_{it} + \omega_{it}, \quad \omega_{it} = (a_i - a) + u_{it}, \quad u_{it} = \rho u_{it-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim iidN\left(0, \sigma_\varepsilon^2\right). \tag{9}$$

   (a) you want to estimate the set of parameters by maximizing log likelihood function. Write down the set of parameters.

   (b) Write down the log likelihood function and its F.O.C.

   (c) Derive MLEs when $\rho = 0$ and this information is given to you.

10 Binary Choice Model: (Chapter 23, Green)

10.1 Cross Sectional Regression

$$y_i = 1\{a + bx_i + u_i > 0\} \tag{10}$$

where $1\{\cdot\}$ is a binary function. That is

$$y_i = \begin{cases} 1 \text{ if } a + bx_i + u_i > 0 \\ 0 \text{ otherwise.} \end{cases}$$

10.1.1 Regression Type: Linear Probability Model (LPM)

$$y_i = a_1 + b_1 x_i + e_i = \mathbf{x}'\boldsymbol{\beta} + \mathbf{e} \tag{11}$$

Let

$$\begin{aligned} \Pr(y = 1|\mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \Pr(y = 0|\mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}) \end{aligned}$$

where F is a CDF, typically assumed to be symmetric about zero. That is, $F(u) = 1 - F(-u)$. Then we have

$$E(y|\mathbf{x}) = 1 \times F(\mathbf{x}, \boldsymbol{\beta}) + 0 \times \{1 - F(\mathbf{x}, \boldsymbol{\beta})\} = F(\mathbf{x}, \boldsymbol{\beta}).$$

Now

$$y = E(y|\mathbf{x}) + (y - E(y|\mathbf{x})) = E(y|\mathbf{x}) + \mathbf{e} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{e}$$

Properties

1. $a_1 + b_1 x_i + e_i$ should be either 1 or 0. That is,

$$\begin{aligned} a_1 + b_1 x_i + e_i &= 1 \iff e_i = 1 - a_1 - b_1 x_i \text{ with } F(\mathbf{x}, \boldsymbol{\beta}) \\ a_1 + b_1 x_i + e_i &= 0 \iff e_i = -a_1 - b_1 x_i \text{ with } 1 - F \end{aligned}$$

2. Hence

$$Var(\mathbf{e}|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}\left(1 - \mathbf{x}'\boldsymbol{\beta}\right)$$

3. Easy to interpret.

$$\frac{1}{n}\sum y_i = \text{ estimated probability that } y = 1$$

$$\frac{1}{n}\sum y_i = a_1 + b_1 \frac{1}{n}\sum x_i$$

Criticism

1. If (10) is true, then (11) is false.

2. $\mathbf{x}'\boldsymbol{\beta}$ is constrained to be between 0 and 1.

10.1.2 Logit and Probit Model

Assume that (here I delete constant term for notational convenience). Both logit and probit model work with the latent model given by

$$y_i^* = bx_i + u_i$$

and

$$y_i = 1\left\{y_i^* > 0\right\}$$

Then

$$\Pr\left[y_i = 1|x_i\right] = \Pr\left[u_i > bx_i\right] = F\left(bx_i\right) = 1 - F\left(-bx_i\right)$$

Two common choices are

$$F\left(bx\right) = \frac{\exp\left(bx_i\right)}{1+\exp\left(bx_i\right)} : \text{ logit},$$

and

$$F\left(bx_i\right) = \int_{-\infty}^{bx_i} \phi\left(t\right) dt = \Phi\left(bx_i\right) : \text{ probit}.$$

How to interpret the regression coefficient: Logit and probit models are not linear. Hence the interpretation of the regression coefficients must be done in the following way.

$$\begin{aligned} \frac{\partial E\left[y|x\right]}{\partial x} &= \left\{\frac{dF\left(bx_i\right)}{d\left(bx_i\right)}\right\} b = f\left(bx_i\right) b \\ &= \begin{cases} \phi\left(bx_i\right) b : \text{ probit} \\ \frac{\exp(bx_i)}{(1+\exp(bx_i))^2} = \Lambda\left(bx_i\right)\left[1-\Lambda\left(bx_i\right)\right] : \text{ logit} \end{cases} \end{aligned}$$

Two way to calculate slope.

1. use sample mean of x_i

$$\frac{\partial E\left[y|x\right]}{\partial x} = \begin{cases} \phi\left(b\bar{x}\right) b : \text{ probit} \\ \Lambda\left(b\bar{x}\right)\left[1-\Lambda\left(b\bar{x}\right)\right] : \text{ logit} \end{cases}$$

2. use sample mean of slopes across x_i

$$\frac{\partial E[y|x]}{\partial x} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial E[y|x_i]}{\partial x_i} = \begin{cases} \frac{1}{n}\sum \phi(bx_i)\, b: \text{ probit} \\ \frac{1}{n}\sum \Lambda(bx_i)[1-\Lambda(bx_i)]: \text{ logit} \end{cases}$$

1 and 2 are similar. So usually 1 is used.

10.1.3 Estimation of Logit and Probit: Using MLE.

Assumption: independent and identical distributed.

Then the joint pdf is given by

$$\Pr(y_1, ..., y_n|x) = \prod_{y_i=0}[1-F(bx_i)]\prod_{y_i=1}F(bx_i)$$

and the likelihood function can be defined as

$$L(b) = \prod_{i=1}^{n}[1-F(bx_i)]^{1-y_i}F(bx_i)^{y_i}$$

and its log likelihood is given by

$$\ln L = \sum_{i=1}^{n}[y_i \ln F(bx_i) + (1-y_i)\ln\{1-F(bx_i)\}].$$

Now F.O.C. is

$$\frac{\partial \ln L}{\partial b} = \sum\left[\frac{y_i f_i}{F_i} + (1-y_i)\frac{-f_i}{(1-F_i)}\right]x_i = 0$$

where $f_i = dF_i/d(bx_i)$. More specifically, we have

$$\frac{\partial \ln L}{\partial b} = \begin{cases} \sum(y_i - \Lambda_i)x_i = 0 \\ \sum \lambda_i x_i = 0 \end{cases}$$

where

$$\lambda_i = \frac{(2y_i-1)\phi((2y_i-1)bx_i)}{\Phi((2y_i-1)bx_i)}.$$

Estimation of Covariance Matrix 1. Inverse Hessian matrix.

$$H = \frac{\partial^2 \ln L}{\partial b \partial b'} = \begin{cases} -\sum \Lambda_i x_i x_i': \text{ logit} \\ \sum -\lambda_i(\lambda_i + bx_i)x_i x_i': \text{ probit} \end{cases}$$

2. Berndt, Hall, Hall and Hausman Estimator

$$B = \begin{cases} \sum(y_i - \Lambda_i)^2 x_i x_i': \text{ logit} \\ \sum \lambda_i^2 x_i x_i': \text{ probit} \end{cases}$$

3. Robust Covariance Estimator

$$V\left(\hat{b}\right) = \left[\hat{H}\right]^{-1} \hat{B} \left[\hat{H}\right]^{-1}$$

Estimation of Covariance of Marginal Effects Marginal Effects $= f\left(\hat{b}\bar{x}\right)\hat{b} = \hat{F}$, let say. How to estimate its variance?

$$V\left(\hat{F}\right) = \left(\frac{\partial \hat{F}}{\partial b}\right)' V \left(\frac{\partial \hat{F}}{\partial b}\right)$$

where

$$V = V(\hat{b})$$

Issue on Binary Choice Models First consider linear model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

but you run

$$y = X_1\beta_1 + u$$

Then A) if X_2 is correlated with X_1, $\hat{\beta}_1$ is inconsistent. B) If $\frac{1}{n}X_2'X_1 = 0$ (orthogonal), then $\hat{\beta}_1$ is unbiased even when $\beta_2 \neq 0$.

Now consider the binary choice model

$$y = 1\left\{X_1\beta_1 + X_2\beta_2 + \varepsilon > 0\right\}$$

Assume $EX_1X_2 = 0$ but $\beta_2 \neq 0$. And you run

$$y = 1\left\{X_1\beta_1 + u > 0\right\}$$

Then

$$\text{plim}\hat{\beta}_{1|\text{wo } X_2} = c_1\beta_1 + c_2\beta_2 \neq \beta_1$$

Likelihood Ratio Test (LR) Let

$$H_0 : \beta_2 = 0$$

Then we can test this null hypothesis by using likelihood ratio test given by

$$-2\left(\ln L_R - \ln L_U\right) \sim \chi_k^2$$

where k is the number of restriction.

Measuring Goodness of Fit R^2 does not work since the regression erros are inside the nonlinear function. Several measures are used in practice.

1. McFadden's likelihood ratio index
$$LR1 = 1 - \frac{\ln L}{\ln L_0}$$
where L_0 is the likelihood only with constant term.

2. Ben-Akiva, Lerman, Kay and Little
$$R^2_{BL} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i\hat{F}_i + (1-y_i)\left(1-\hat{F}_i\right)\right]$$

3. See Green page 791 for other criteria

10.2 Time Series Regression

Models are given by
$$y_t = 1\{a + bx_t + u_t > 0\}.$$
Note that there is no difference in terms of estimation and statistical inference from cross sectional binary choice model. However, in time series case, the persistent response becomes an important issue. In fact, most of time series binary choices are very persistent, and especially the source of such persistency becomes an important issue. There are three explanations

1. Fixed effects
$$y_t^* = a + bx_t + u_t > 0 \text{ because } a >> 0: \text{ M1}$$
In this case, y_t^* has all positive values for most of all times.

2. y_t^* is highly persistent.
$$y_t^* = a + \rho y_{t-1}^* + u_t\text{: M2}$$

3. y_t is depending on y_{t-1} (past choice).
$$y_t = 1\{a + \rho y_{t-1} + u_t\}: \text{ M3}$$

Model1 and Model 2 can be identified from Model 3. However if two models are mixed, then it is impossible to identify the order of serial correlation. In other words, it the true model is given by
$$y_t = 1\left\{y_t^* = a + \rho y_{t-1} + \phi y_{t-1}^* + u_t\right\}: \text{ M4}$$
Then ϕ and ρ are not in general identifiable.

Heckman Run Test

Assumption 1 (Cross Section Independence) The binary choice is not cross sectionally dependent. That is, $a_i \sim i.i.d\left(0, \sigma_a^2\right)$, and $e_{it} \sim i.i.d\left(0, \sigma_i^2\right)$.

Assumption 2 (Initial Condition) For M2, $y_{i-1} = 0$. That is, $y_{i1} = 1\{a + e_{i1} \geq 0\}$. For M3, $y_{i0} = 1\{a + u_{i0} \geq 0\}$ and $u_{i0} \sim i.i.d\left(0, \sigma_i^2/\left(1-\rho^2\right)\right)$.

Under these two assumptions, Heckman suggests the so-called 'run' test by looking running patterns of y_{it}. To fix the idea, let $T = 1, 2, 3$ and consider the true probablity of each run

Model	Run Patterns	
M1	P(110) =P(011) =P(101)	P(100) =P(001) =P(010)
M2	P(110) =P(011) ≠P(101)	P(100) =P(001) ≠P(010)
M3	P(011) >P(101) >P(110)	P(001) >P(010) >P(100)
M4	All probabilities distinct but unordered	

By using these runs patterns, Heckman constructs the following two sequential null hypotheses to distinguish the first three models.

$$\begin{aligned} H_{01} &: \mathrm{P}(011) = \mathrm{P}(101) \ \& \ \mathrm{P}(001) = \mathrm{P}(010) \ \text{under M1} \\ H_{A1} &: \mathrm{P}(011) \neq \mathrm{P}(101) \ \text{or} \ \mathrm{P}(001) \neq \mathrm{P}(010) \ \text{under M2, M3, and M4} \end{aligned}$$

When the first null hypothesis is rejected, then the second null hypothesis can be tested.

$$\begin{aligned} H_{02} &: \mathrm{P}(110) = \mathrm{P}(011) \ \text{and} \ \mathrm{P}(100) = \mathrm{P}(001) \ \text{under M2} \\ H_{A2} &: \mathrm{P}(110) \neq \mathrm{P}(011) \ \text{or} \ \mathrm{P}(100) \neq \mathrm{P}(001) \ \text{under M3 and M4} \end{aligned}$$

Heckman suggests Peason's score χ^2 statistics for both null hypotheses. For H_{01} and H_{02}, the test statistics are given by

$$\begin{aligned} \mathcal{P}_{01} &= \frac{\left(F_{011} - \mathrm{E}F_{011}\right)^2}{\mathrm{E}F_{011}} + \frac{\left(F_{101} - \mathrm{E}F_{101}\right)^2}{\mathrm{E}F_{101}} \Longrightarrow \chi_1^2 \\ \mathcal{P}_{02} &= \frac{\left(F_{110} - \mathrm{E}F_{110}\right)^2}{\mathrm{E}F_{110}} + \frac{\left(F_{011} - \mathrm{E}F_{011}\right)^2}{\mathrm{E}F_{011}} \Longrightarrow \chi_1^2 \end{aligned}$$

where F_{011} is the observed frequency for the outcome of the '$0, 1, 1$' response. Similary F_{ijk} is defined in the same way. $\mathrm{E}F_{011} = \mathrm{E}F_{101}$ under H_{01} while $\mathrm{E}F_{110} = \mathrm{E}F_{011}$ under H_{02}.

Several issues are arised in Heckman's run tests. First, Heckman's run test requires to estimate the expected frequency under the null hypothesis. When $\alpha_i \neq \alpha$ in M3 or M2, it is

hard to estimate the expected frequency from the models. Second, M3 is hard to distinguished from M4 if the current choice depends on many past lagged depedent variables. In fact, Heckman does not provide any formal test to distinguish M3 from M4. Third, when the binary panel data shows severe persistency, the numbers of observations in each case for the two null hypotheses are decreased significantly. In fact, Heckman (1978) couldn't reject the first null hypothesis by using 198 individuals over three years of the data: Out of 198 individuals, 165 individuals show either '111' or '000' flat response, and only less than 17% of individuals show heterogeneous responses. Finally, such runs patterns become useless if the first observation does not start from $t = 1$. For example, if econometricians don't observe the first k observations, and if they treated as if the $k+1$th observation as the first observation, they can't obtain the heterogeneous running patterns for M2 and M3. With a moderate large k, P(110) =P(011) and P(100) =P(001) both under M2, M3 and M4. Hence the second null hypothesis can't be tested by using runnig patterns.

11 Panel Binary Choice

11.1 Multivariate Models (See 23.8 Green)

Model

$$\begin{aligned} y_1^* &= x_1\beta_1 + \varepsilon_1, \\ y_2^* &= x_2\beta_2 + \varepsilon_2, \end{aligned}$$

where

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

Consider the bivaraite normal cdf given by

$$\Pr\left[X_1 < x_1, X_2 < x_2\right] = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \phi_2\left(z_1, z_2, \rho\right) dz_1 dz_2$$

where

$$\phi_2\left(z_1, z_2, \rho\right) = \frac{\exp\left[-\frac{1}{2}\left(x_1^2 + x_2^2 - 2\rho x_1 x_2\right) / \left(1-\rho^2\right)\right]}{2\pi\sqrt{\left(1-\rho^2\right)}} \tag{12}$$

Likelihood Function Let

$$\begin{aligned} q_{i1} &= 2y_{i1} - 1, \quad q_{i2} = 2y_{i2} - 1 \\ z_{ij} &= x_{ij}'\beta_j, \quad w_{ij} = q_{ij} z_{ij}, \quad \rho_{i^*} = q_{i1} q_{i2} \rho \end{aligned}$$

Then the log likelihood function can be written as

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2 \left(w_{i1}, w_{i2}, \rho_{i^*} \right)$$

Marginal Effects

$$\frac{\partial \Phi_2}{\partial x_1} = \phi \left(x_1' \beta_1 \right) \Phi \left(\frac{x_2' \beta - \rho x_1' \beta}{\sqrt{1 - \rho^2}} \right) \beta_1$$

Testing for zero correlation

$$\lambda_{LR} = 2 \left[\ln L_{\text{multi}} - (\ln L_1 + \ln L_2) \right] \rightarrow^d \chi_1^2$$

11.2 Recursive Simulatneous Equations

$$\Pr \left[y_1 = 1, y_2 = 1 | x_1, x_2 \right] = \Phi \left(x_1 \beta_1 + \gamma y_2, x_2 \beta_2, \rho \right)$$

At least one endogeneous variable is expressed only with exogenous variables.

Results (Maddala 1983)

1. We can ignore the simultaneity
2. Use log-likelihood estimation. Not LPM.

11.3 Panel Probit Model

Model

$$y_{it} = 1 \left\{ y_{it}^* > 0 \right\}$$

Results

1. If $T < N$, then don't use fixed effects: Let $y_{it}^* = a_i + \beta x_{it} + u_{it}$. Note that y_{it} is either 1 or 0. When T is small, a_i is impossible to identify.
2. If $T > N$, then use multivariate probit or logit.
3. If $T < N$, then you can use random effects but have to know that it is very complicated.
4. Overall, Don't use panel probit.

11.4 Conditional Panel Logit Model with Fixed Effects

Set $T = 2$, consider the cases

$$\begin{aligned} \text{C1} &: \quad y_{i1} = 0, y_{i2} = 0 \\ \text{C2} &: \quad y_{i1} = 1, y_{i2} = 1 \\ \text{C3} &: \quad y_{i1} = 0, y_{i2} = 1 \\ \text{C4} &: \quad y_{i1} = 1, y_{i2} = 0 \end{aligned}$$

The unconditional likelihood becomes

$$L = \prod \Pr\left(Y_{i1} = y_{i1}\right) \Pr\left(Y_{i2} = y_{i2}\right),$$

which is not helpful to eliminate fixed effects.

For C1, consider this

$$\Pr\left[y_{it} = 1 | x_{it}\right] = \frac{\exp\left(a_i + x_{it}\beta\right)}{1 + \exp\left(a_i + x_{it}\beta\right)}$$

Now, we want to eliminate the fixed effects from logistic distribution. How? Consider the following probability

$$\begin{aligned} \Pr\left[y_{i1} = 0, y_{i2} = 1 | sum = 1\right] &= \frac{\Pr\left[0, 1 \text{ and } sum = 1\right]}{\Pr\left[sum = 1\right]} \\ &= \frac{\Pr\left[0, 1\right]}{\Pr\left[0, 1\right] + \Pr\left[1, 0\right]} \end{aligned}$$

Hence for this pair of obs, the conditional probability is given by

$$\frac{\frac{1}{1+\exp(a_i+x_{i1}\beta)} \frac{\exp(a_i+x_{i2}\beta)}{1+\exp(a_i+x_{i2}\beta)}}{\frac{1}{1+\exp(a_i+x_{i1}\beta)} \frac{\exp(a_i+x_{i2}\beta)}{1+\exp(a_i+x_{i2}\beta)} + \frac{1}{1+\exp(a_i+x_{i2}\beta)} \frac{\exp(a_i+x_{i1}\beta)}{1+\exp(a_i+x_{i1}\beta)}} = \frac{\exp\left(x_{i1}\beta\right)}{\exp(x_{i1}\beta) + \exp\left(x_{i2}\beta\right)}$$

In other words, conditioning on the sum of the two observations, we can remove the fixed effects. Now the log likelihood function is given by

$$\ln L = \sum_{i=1}^{n} d_i \left[y_{i1} \ln\left(\frac{\exp\left(x_{i1}\beta\right)}{\exp(x_{i1}\beta) + \exp\left(x_{i2}\beta\right)}\right) + y_{i2} \ln\left(\frac{\exp\left(x_{i2}\beta\right)}{\exp(x_{i1}\beta) + \exp\left(x_{i2}\beta\right)}\right)\right]$$

where

$$d_i = 1 \text{ if } y_{i1} + y_{i2} = 1, \text{ and 0 otherwise.}$$

11.5 Conditional Panel Logit Model with Fixed and Common Time Effects

Model

$$y_{it} = 1\left\{y_{it}^* > 0\right\}$$

where

$$y_{it}^* = a_i + \theta_t + x_{it}\beta + u_{it}$$

Consider

$$\Pr\left[y_{i1} = 0, y_{i2} = 1 | sum = 1\right] = \frac{\exp\left(\theta_2 + x_{i2}\beta + a_i\right)}{\exp\left(\theta_2 + x_{i2}\beta + a_i\right) + \exp\left(\theta_1 + x_{i1}\beta + a_i\right)}$$

Let divide both sides by $\exp\left[\theta_1 + x_{i1}\beta + a_i\right]$, then we have

$$\begin{aligned} & \frac{\exp\left(\theta_2 + x_{i2}\beta + a_i\right) / \exp\left[\theta_1 + x_{i1}\beta + a_i\right]}{\exp\left(\theta_2 + x_{i2}\beta + a_i\right) / \left[\theta_1 + x_{i1}\beta + a_i\right] + \exp\left(\theta_1 + x_{i1}\beta + a_i\right) / \left[\theta_1 + x_{i1}\beta + a_i\right]} \\ = & \frac{\exp\left(\theta_2 - \theta_1 + \left(x_{i2} - x_{i1}\right)\beta\right)}{\exp\left(\theta_2 - \theta_1 + \left(x_{i2} - x_{i1}\right)\beta\right) + \exp\left(0\right)} = \frac{\exp\left(\Delta\theta + \Delta x_i\beta\right)}{\exp\left(\Delta\theta + \Delta x_i\beta\right) + 1} \end{aligned}$$

Hence the condition log-likelihood function can be written as

$$\ln L = \sum_{i=1}^{n} d_i \left[y_{i1} \ln\left(\frac{\exp\left(\Delta\theta + \Delta x_i\beta\right)}{1 + \exp\left(\Delta\theta + \Delta x_i\beta\right)}\right) + y_{i2} \ln\left(\frac{1}{1 + \exp\left(\Delta\theta + \Delta x_i\beta\right)}\right)\right]$$

Remark Fixed and common time effects can't be estimated. Use panel profit with random effects to estimate their variances if you are interested in them.

STATA CODE: xtlogit y x, fe. Marginal effects: mxf, predict(pu0)

12 Multinomial Choice Models

Two types of multinomial choices: Unordered choice v.s. ordered choice model. First we consider unordered choice.

12.1 Unordered Choice or Multinomial Logit Model (23.11 Green)

Example: How to commute to school.

(1) automobile (2) bike (3) walk (4) Bus (5) Train

$$y_i = J\left\{y_{ij}^* > y_{ij^c}^*\right\}$$

where $J\{\cdot\}$ is an interger J function if $\{\cdot\}$ is true. That is, $J = 0, 1, 2, ..., K$. Note that an individual j will choose j if y_{ij}^* (utility) is greater than any other choice, $y_{ij^c}^*$.

Now, let

$$y_{ij}^* = a_j x_i + e_{ij}$$

Note that the coefficient on x_i is varying across choices, j. Why? Suppose that $a_j = a$ across j. Then

$$y_{ij}^* = y_i^* \text{ for all } j$$

so that an individual i does not make any choice (since there is no dominant choice). Similary, let

$$y_{ij}^* = a_j x_i + \beta z_i + u_{ij}$$

then the coefficient β is not identifiable due to the same reason.

Hence when you model for multinomial choice, you may want to include some variable which will differ across j. For example, the cost of transportation must be different across j. In this case, we can setup the model given by

$$y_{ij}^* = a_j x_i + \beta z_{ij} + u_{ij}.$$

In this case, the probability is given by

$$\Pr[Y_i = j] = \frac{\exp(a_j x_i + \beta z_{ij})}{\sum_{j=0}^{K} \exp(a_j x_i + \beta z_{ij})}$$

Now, let's consider only x_i case by setting $\beta = 0$. Then we have

$$\Pr[Y_i = j] = \frac{\exp(a_j x_i)}{\sum_{j=0}^{K} \exp(a_j x_i)},$$

so that the first choice will not be identified since the sum of probability should be one. Hence we let (usually)

$$\Pr[Y_i = j] = \frac{\exp(a_j x_i)}{1 + \sum_{j=1}^{K} \exp(a_j x_i)} \text{ by setting } a_0 = 0.$$

The log likelihood function is given by

$$\ln L = \sum_{i=1}^{n} \sum_{j=0}^{K} d_{ij} \ln \left(\frac{\exp(a_j x_i)}{1 + \sum_{j=1}^{K} \exp(a_j x_i)} \right)$$

where $d_{ij} = 1$ if $y_i = j$, otherwise 0.

Issue: 1. Independence of Irelevant Alternatives (IIA).Individual preference should not be dependent on what other choices are available.

2. Panel multinomial logit: pooled one is okay. random effects mlogit is okay. fixed effects clogit is not available yet.

12.2 Ordered Choices (Chapter 23.10 in Green)

Example: Recommendation letter.

(1) Outstanding (2) excellent (3) good (4) average (5) poor

Usually rating is corresponding to the distribution of the grades.

Then we can say

$$y = \begin{cases} 0 \text{ if } y^* \leq 0 \\ 1 \text{ if } 0 < y^* \leq c_1 \\ \vdots \\ k \text{ if } c_{k-1} < y^* \end{cases}$$

Then we have

$$\begin{aligned} \Pr(y = 0|x) &= \Phi(-x'\beta) \\ \Pr(y = 1|x) &= \Phi(c_1 - x'\beta) - \Phi(x'\beta) \\ &\vdots \\ \Pr(y = k|x) &= 1 - \Phi(c_{k-1} - x'\beta) \end{aligned}$$

so that the likelihood function becomes

$$\ln L = \sum_{i=1}^{n} \sum_{j=0}^{k} d_{ij} \ln \left\{ \Phi(c_j - x'\beta) - \Phi(c_{j-1} - x'\beta) \right\}$$

where $d_{ij} = 1$ if $y_i = j$, otherwise 0.

13 Truncated and Censored Regressions

13.1 Truncated Distributions

x_i has a nontruncated distribution. For an example, $x \sim iidN(0,1)$. If x_i is truncated around a, then its pdf is changed to

$$f(x|x>a) = \frac{f(x)}{\Pr(x>a)} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)}{1-\int_{-\infty}^{a}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)dx}$$

For an example, if $a=0$, then we have

$$\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)dx = 0.5$$

so that

$$f(x|x>0) = \frac{f(x)}{\Pr(x>0)} = 2\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right) \text{ from } x>0$$

Now consider its mean

$$E[x|x>a] = \int_a^\infty xf(x|x>a)\,dx = \int_a^\infty x\frac{2}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)dx = \frac{\sqrt{2}}{\sqrt{\pi}}e^{-\frac{1}{2}a^2}$$

For the case of $a=0$, we have

$$E[x|x>0] = \frac{\sqrt{2}}{\sqrt{\pi}} = 0.798$$

More generally, we have the following fact.

Let $x \sim iidN\left(\mu,\sigma^2\right)$ and a is a constant, then

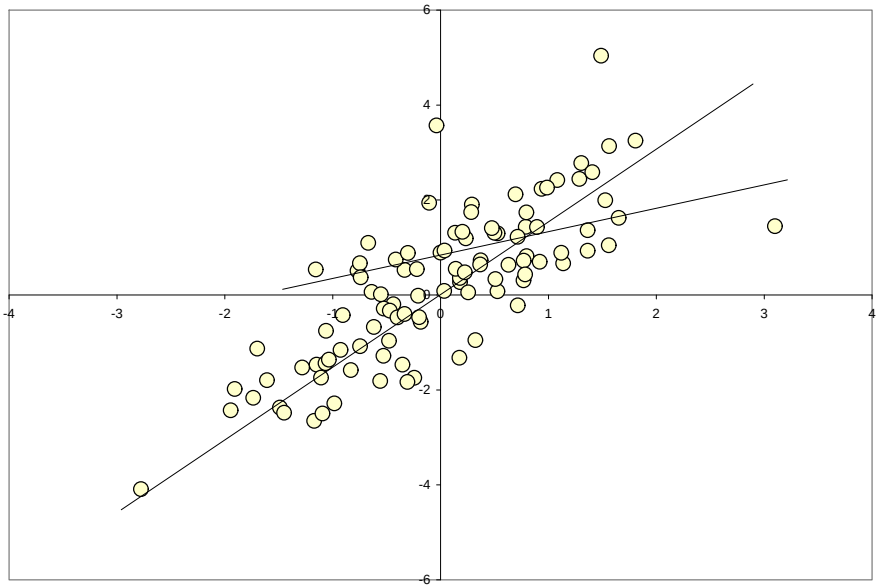
$$\begin{aligned} E(x|x>a) &= \mu+\sigma\lambda(\alpha) \\ V(x|x>a) &= \sigma^2[1-\delta(\alpha)] \end{aligned}$$

where

$$\alpha = \frac{a-\mu}{\sigma}$$

and

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1-\Phi(\alpha)}, \quad \delta(\alpha) = \lambda(\alpha)(\lambda(\alpha)-\alpha).$$



13.2 Truncated Regression

Example: y_i = Test score, x_i = parents income. Data selection: High Test score only. Suppose that we have the following regression

$$y_i = x_i\beta + \varepsilon_i$$

The conditional expectation of y given x is equal to

$$\begin{aligned} E_x\left(y|y>0\right) &= E_x\left(y^*|y^*>0\right) = E_x\left(x\beta+\varepsilon|\varepsilon>-x\beta\right) \\ &= x\beta + E_x\left(\varepsilon|\varepsilon>-x\beta\right) \neq x\beta \end{aligned}$$

since

$$E_x\left(\varepsilon|\varepsilon>-x\beta\right) \neq 0.$$

Hence typical LS estimator becomes biased and inconsistent. We call this bias sample selection bias.

The solution for this problem is using MLE based on the truncated log likelihood function given by

$$\ln L = \sum_{i=1}\left[\ln\left\{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right\} - \ln\Phi\left(\frac{x_i\beta}{\sigma}\right)\right].$$

We will consider more solution later.

Now, what happen if x_i is truncated? Note that

$$\begin{aligned} E_x\left[y_i|x_i > a\right] &= E_x\left[x\beta + \varepsilon|x > a\right] \\ &= x\beta + E_x\left[\varepsilon|x > a\right] = x\beta \end{aligned}$$

Hence the typical OLS estimator becomes consistent.

13.3 Censored Distribution

We say that y_i is censored if

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* \leq 0 \\ y_i &= y_i^* \text{ if } y_i^* > 0 \end{aligned}$$

Hence y_i^* has a continuous (or non-censored) distribution.

Example: If $y^* \sim iidN\left(\mu, \sigma^2\right)$, and $y = a$ if $y^* \leq a$ or else $y = y^*$, then

$$\begin{aligned} E\left(y\right) &= \Pr\left(y = a\right) a + \Pr\left(y > a\right) E\left(y^*|y^* > a\right) \\ &= \Pr\left(y^* \leq a\right) a + \Pr\left(y^* > a\right) E\left(y^*|y^* > a\right) \\ &= \Phi\left(a\right) a + \left\{1 - \Phi\left(a\right)\right\} \left\{\mu + \sigma\lambda\left(\alpha\right)\right\} \end{aligned}$$

and

$$V\left(y\right) = \sigma^2\left(1 - \Phi\right)\left[\left(1 - \delta\right) + \left(\alpha - \lambda\right)^2 \Phi\right],$$

where

$$\alpha = \left(a - \mu\right)/\sigma,\ \lambda = \phi\left(\alpha\right)/\left\{1 - \Phi\left(\alpha\right)\right\},\ \delta = \lambda^2 - \lambda\alpha$$

Remark Let $y^* \sim iidN\left(0, 1\right)$, and $y = 0$ if $y_i^* \leq 0$ or else $y = y^*$. Then

$$\lambda = \frac{\phi\left(0\right)}{1 - \Phi\left(0\right)} = \frac{\frac{1}{\sqrt{2\pi}}}{0.5} = 0.798,\ \delta = 0.798^2 = 0.637$$

$$\begin{aligned} E\left(y|a = 0\right) &= 0.5\left(0 + 0.798\right) = 0.399 \\ V\left(y\right) &= 0.5\left[\left(1 - 0.637\right) + 0.637 \times 0.5\right] = 0.637 \end{aligned}$$

13.4 Censored Regression (Tobit) Model

Tobin proposed this model. So we call it Tobit model.

$$y_i^* = x_i\beta + \varepsilon_i$$

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* \leq 0 \\ y_i &= y_i^* \text{ if } y_i^* > 0 \end{aligned}$$

The conditional expectation of y given x is equal to

$$\begin{aligned} E[y|x] &= \Pr[y=0]\cdot 0 + \Pr[y>0]\, E_x(y|y>0) \\ &= \Pr[y>0]\, E_x(y|y>0) \\ &= \Pr[\varepsilon > -x\beta]\, E_x(y^*|\varepsilon > -x\beta) \\ &= \Pr[\varepsilon > -x\beta]\, E_x(x\beta+\varepsilon|\varepsilon > -x\beta) \\ &= \Pr[\varepsilon > -x\beta]\, \{x\beta + E_x(\varepsilon|\varepsilon > -x\beta)\} \\ &= \Phi\left(\frac{x_i\beta}{\sigma}\right)(x_i\beta + \sigma\lambda_i) \end{aligned}$$

where

$$\lambda_i = \frac{\phi[-x_i\beta/\sigma]}{1-\Phi[-x_i\beta/\sigma]} = \frac{\phi[x_i\beta/\sigma]}{\Phi[x_i\beta/\sigma]}$$

since ϕ is a symmetric distribution. Also note that the OLS estimator becomes biased and inconsistent.

Similar to the truncated regression, the ML estimator based on the following likelihood function becomes consistent.

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \ln\sigma^2 + \frac{(y_i - x_i\beta)^2}{\sigma^2}\right] + \sum_{y_i=0}\ln\left[1-\Phi\left(\frac{x_i\beta}{\sigma}\right)\right]$$

Now, the marginal effect is given by

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \beta\Phi\left(\frac{x_i\beta}{\sigma}\right).$$

13.5 Balanced Trimmed Estimator (Powell, 1986)

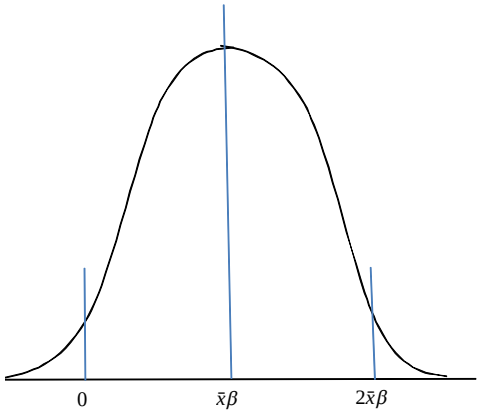
Truncation and censored problems arise due to asymmetric truncation. Now consider the following truncation rule

$$y_i^* = x_i\beta + u_i$$

$$y_i = 0 \text{ if } y_i^* \leq 0 \iff y_i = 0 \text{ if } x_i\beta \leq u_i$$

$$y_i = y_i^* \text{ if } y_i^* > 0 \iff y_i = y_i^* \text{ if } x_i\beta > -u_i$$

We can't do anything about censored or truncated parts, but can modify the non-censored or non-truncated part to balance up the symmetry. Consider the below figure. The vertical axis represents the density of y_i^*. When y_i^* is either censored or truncated about 0, the mean of y_i^* shifts due to asymmetry of its pdf. To avoid this, we can censor or truncate the right side distribution about $2\bar{x}\beta$.



Powell (1986) suggests the following criteria of nonlinear LS estimator. For truncated regressions,

$$T(\beta) = \sum_{i=1}^{n} \left[y_i - \max\left(\frac{1}{2} y_i, x_i\beta \right) \right]^2$$

and for censored regressions

$$C(\beta) = \sum_{i=1}^{n} \left[y_i - \max\left(\frac{1}{2} y_i, x_i\beta \right) \right]^2 + \sum_{i=1}^{n} 1\{y_i > 2x_i\beta\} \left[\left(\frac{1}{2} y_i \right)^2 - \max(0, x_i\beta)^2 \right]$$

The F.O.C for $T(\beta)$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} 1\{y_i < 2x_i\beta\} (y_i - x_i\beta) x_i = 0$$

and the F.O.C. for $C(\beta)$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} 1\{x_i\beta > 0\} \{\min[y_i, 2x_i\beta] - x_i\beta\} x_i = 0$$

13.6 Sample Selection Bias: Heckman's two stage estimator

13.6.1 Incidental Truncated Bivariate Normal Distribution

Let y and z have a bivariate normal distribution

$$\begin{bmatrix} y \\ z \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}, \begin{bmatrix} \sigma_y & \sigma_{yz} \\ \sigma_{yz} & \sigma_z \end{bmatrix} \right)$$

and let

$$\rho = \frac{\sigma_{yz}}{\sqrt{\sigma_y \sigma_z}}$$

Then we have

$$\begin{aligned} E\left[y|z > a\right] &= \mu_y + \rho\sigma_y \lambda\left(\alpha_z\right) \\ V\left[y|z > a\right] &= \sigma_y^2\left[1 - \rho^2 \delta\left(\alpha_z\right)\right] \end{aligned}$$

where

$$\alpha_z = \frac{a - \mu_z}{\sigma_z}, \quad \lambda\left(\alpha_z\right) = \frac{\phi\left(\alpha_z\right)}{1 - \Phi\left(\alpha_z\right)}, \quad \delta\left(\alpha_2\right) = \lambda\left(\alpha_z\right)\left[\lambda\left(\alpha_z\right) - \alpha_z\right]$$

13.6.2 Sample Selection Bias

Consider

$$\begin{aligned} z_i^* &= w_i\gamma + u_i \\ y_i &= x_i\beta + \varepsilon_i \end{aligned}$$

where z_i^* is unobservable. Suppose that

$$y_i = \begin{cases} n.a. \text{ if } z_i^* \leq 0 \\ y_i \text{ otherwise} \end{cases}, \text{ Truncation based on } z_i^*$$

then

$$\begin{aligned} E\left(y_i|z_i^* > 0\right) &= E\left(y_i|u_i > -w_i\gamma\right) \\ &= x_i\beta + E\left(\varepsilon_i|u_i > -w_i\gamma\right) \\ &= x_i\beta + \rho\sigma_\varepsilon \lambda_i\left(\alpha_u\right) \end{aligned}$$

Hence the OLS estimator becomes biased and inconsistent.

Solution (Heckman's two stage estimation) (1) Assume normality

We rewrite the model as

$$z_i = 1\left\{z_i^* = w_i\gamma + u_i > 0\right\}$$

and

$$y_i = x_i\beta + \varepsilon_i \text{ if } z_i = 1$$

Step 1 Probit regression with z_i. Estimate $\hat{\gamma}_{\text{probit}}$, and compute $\hat{\lambda}_i$ and $\hat{\delta}_i$ given by

$$\hat{\lambda}_i = \frac{\phi\left(w_i\hat{\gamma}_{\text{probit}}\right)}{\Phi\left(w_i\hat{\gamma}_{\text{probit}}\right)}, \quad \hat{\delta}_i = \hat{\lambda}_i\left(\hat{\lambda}_i + w_i\hat{\gamma}_{\text{probit}}\right).$$

Step 2 Estimate β and $\beta_\lambda = \rho\sigma_\varepsilon$ by OLS

$$y_i = x_i\beta + \beta_\lambda\hat{\lambda}_i + \text{ error}$$

What if $x_i = w_i$? Then we have

$$\begin{aligned} z_i &= 1\left\{y_i^* = x_i\beta + u_i > 0\right\} \\ y_i &= = x_i\beta + \varepsilon_i \text{ if } z_i = 1 \end{aligned}$$

Step 1 Probit regression with z_i. Estimate $\hat{\beta}_{\text{probit}}$, and compute $\hat{\lambda}_i$ given by

$$\hat{\lambda}_i = \frac{\phi\left(x_i\hat{\beta}_{\text{probit}}\right)}{\Phi\left(x_i\hat{\beta}_{\text{probit}}\right)},$$

Step 2 Estimate β and $\beta_\lambda = \rho\sigma_\varepsilon$ by OLS

$$y_i = x_i\beta + \beta_\lambda\hat{\lambda}_i + \text{ error}$$

13.7 Panel Tobit Model

$$y_{it}^* = a_i + \beta x_{it} + u_{it}$$

$$y_{it} = \begin{cases} n.a. \text{ if } y_{it}^* \leq 0 \\ y_{it}^* \text{ otherwise} \end{cases}, \quad y_{it} = \begin{cases} 0 \text{ if } y_{it}^* \leq 0 \\ y_{it}^* \text{ otherwise} \end{cases}$$

Assume u_{it} is iid and indepenent of x_{it}.

Likelihood Function (Treat a_i random)

$$L\left(\text{truncated}\right)=\prod_{i=1}^{n}\int\left[\prod_{t=1}^{T}\frac{f\left(y_{it}-\beta x_{it}-a_i\right)}{1-F\left(-\beta x_{it}-a_i\right)}\right]g\left(a_i\right)da_i$$

$$L\left(\text{Censored}\right)=\prod_{i=1}^{n}\int\left[\prod_{y_{it}=0}^{T}F\left(-\beta x_{it}-a_i\right)\prod_{y_{it}>0}^{T}f\left(y_{it}-\beta x_{it}-a_i\right)\right]g\left(a_i\right)da_i$$

Symmetric Trimmed LS Estimator (Horone, 1992) Let

$$y_{it}=a_i+x_{it}\beta+u_{it}$$

$$\begin{aligned}y_{it} &= E\left(y_{it}|x_{it},a_i,y_{it}>0\right)+\epsilon_{it}\\ &= a_i+x_{it}\beta+E\left(u_{it}|u_{it}>-a_i-x_{it}\beta\right)+\epsilon_{it}\end{aligned}$$

so that we have

$$y_{it}=a_i+x_{it}\beta+E\left(u_{it}|u_{it}>-a_i-x_{it}\beta\right)+\epsilon_{it}$$

Now take the first s difference

$$\begin{aligned}y_{it}-y_{is} &= \left(x_{it}-x_{is}\right)\beta+E\left(u_{it}|u_{it}>-a_i-x_{it}\beta\right)\\ &\quad -E\left(u_{is}|u_{is}>-a_i-x_{is}\beta\right)+\epsilon_{it}-\epsilon_{is}\end{aligned}$$

In general, we have

$$E\left(u_{it}|u_{it}>-a_i-x_{it}\beta\right)-E\left(u_{is}|u_{is}>-a_i-x_{is}\beta\right)\neq 0$$

Hence a typical sth differencing does not work.

Now consider the following sample truncation

$$y_{it}>\left(x_{it}-x_{is}\right)\beta,\quad y_{is}>-\left(x_{it}-x_{is}\right)\beta$$

Otherwise, drop the sample.

Then we have when $-\left(x_{it}-x_{is}\right)\beta>0$,

$$\begin{aligned}E\left(y_{is}|a_i,x_{it},x_{is},y_{is}>-\left(x_{it}-x_{is}\right)\beta\right) &= a_i+x_{is}\beta+E\left(u_{is}|u_{is}>-a_i-x_{is}\beta-\left(x_{it}-x_{is}\right)\beta\right)\\ &= a_i+x_{is}\beta+E\left(u_{is}|u_{is}>-a_i-x_{it}\beta\right)\end{aligned}$$

Note that due to iid condition, we have

$$E\left(u_{is}|u_{is}>-a_i-x_{it}\beta\right)=E\left(u_{it}|u_{it}>-a_i-x_{it}\beta\right)$$

Similarly when $(x_{it} - x_{is})\beta > 0$,

$$\begin{aligned} E(y_{it}|a_i, x_{it}, x_{is}, y_{it} > (x_{it} - x_{is})\beta) &= a_i + x_{it}\beta + E(u_{it}|u_{it} > -a_i - x_{it}\beta - (x_{it} - x_{is})\beta) \\ &= a_i + x_{it}\beta + E(u_{it}|u_{it} > -a_i - x_{is}\beta) \end{aligned}$$

Note that due to iid condition, we have

$$E(u_{is}|u_{is} > -a_i - x_{is}\beta) = E(u_{it}|u_{it} > -a_i - x_{is}\beta)$$

Hence if we use the observations where (1) $y_{it} > (x_{it} - x_{is})\beta$, (2) $y_{is} > -(x_{it} - x_{is})\beta$, (3) $y_{it} > 0$, (4) $y_{is} > 0$, then we have

$$(y_{it} - y_{is}) = (x_{it} - x_{is})\beta + (\epsilon_{it} - \epsilon_{is}).$$

The OLS estimator becomes consistent.

Since β is unknown, the LS estimator can be obtained by maximinzing the following sum of square errors.

$$\begin{aligned} &\sum_{i=1}^{n} \Big\{ (\Delta y_i - \Delta x_i\beta)^2 1\{y_{i1} \geq -\Delta x_i\beta, y_{i2} \geq \Delta x_i\beta\} \\ &+ y_{i1}^2 1\{y_{i1} > -\Delta x_i\beta, y_{i2} < \Delta x_i\beta\} \\ &+ y_{i2}^2 1\{y_{i1} < -\Delta x_i\beta, y_{i2} > \Delta x_i\beta\} \Big\} \end{aligned}$$

14 Treatment Effects

14.1 Definition and Model

$$d_i = \begin{cases} 1 \\ 0 \end{cases}, \text{ treatment}$$

$$y_i = \begin{cases} y_{i1} \text{ if } d_i = 1 \\ y_{i0} \text{ if } d_i = 0 \end{cases}, \text{ outcome}$$

We can't observe both y_{i1} and y_{i0} *at the same time*

14.1.1 Regression Model

$$y_i = a + \beta d_i + \varepsilon_i \tag{13}$$

If $\hat{\beta} \neq 0$ significantly, we say treatment is effective. This becomes true if d_i is not correlated with ε_i.

Suppose that

$$\begin{aligned} d_i^* &= w_i\gamma + u_i \\ d_i &= 1\{d_i^* > 0\} \end{aligned}$$

and u_i is correlated with ε_i. Then we have

$$\begin{aligned} E(y_i|d_i = 1) &= a + \beta + E(\varepsilon_i|d_i = 1) \\ &= a + \beta + \rho\sigma_\varepsilon\lambda(-w_i\gamma) \neq a + \beta \end{aligned}$$

Hence the treatment effect will be over-estimated.

14.1.2 Bias in Average Treatment Effects

In general, the true treatment effect is given by

$$E(y_{i1} - y_{i0}|d_i = 1) = TE,$$

but it is impossible to observe $E(y_{i0}|d_i = 1)$. Instead of this, we are using

$$ATE = E(y_{i1}|d_i = 1) - E(y_{i0}|d_i = 0),$$

which is called 'average treatment effect'. In the above, we show that this will be upward biased (and inconsistent). To see this, we expand

$$\begin{aligned} E(y_{i1}|d_i=1) - E(y_{i0}|d_i=0) &= E(y_{i1}-y_{i0}|d_i=1) + E(y_{i0}|d_i=1) - E(y_{i0}|d_i=0) \\ &= ATE + [E(y_{i0}|d_i=1) - E(y_{i0}|d_i=0)] \\ &\neq ATE \end{aligned}$$

Of course, if treatment effects are not endogenous (alternatively exogenous or forced to get the treatment), then

$$E(y_i|d_i=1) = a + \beta + E(\varepsilon_i|d_i=1) = a + \beta$$

so that there will be no bias. We call this type of treatment 'randomized treatment'.

14.2 Estimation of Average Treatment Effects (ATE)

14.2.1 Linear Regression

The inconsistency of $\hat{\beta}$ in (13) can be interpreted as endogeneous inconsistency due to missing observations. The typical solution in this case is including control variables, w_i, in the regression. That is,

$$y_i = a + \beta d_i + \gamma w_i + \epsilon_i. \tag{14}$$

This is the most crude estimation method for estimating average treatment effects. The consistency of $\hat{\beta}$ requires the following restriction.

Definition: Unconfoundedness: Conditional on a set of covariate w, the pair of counterfactual outcomes, (y_{i0}, y_{i1}), is independent of d. That is

$$(y_{i0}, y_{i1}) \perp d \mid w$$

Under unconfoundedness, the OLS estimator in (14) becomes consistent, that is

$$\hat{\beta} \to^p \beta$$

and the estimator of ATE becomes $\hat{\beta}$.

Typical linear treatment regression is given by

$$y_i = a + \beta d_i + \gamma_1 w_i + \gamma_2 w_i^2 + \varepsilon_i,$$

but there is no theoretical justification of why d_i has a linear relationship with w_i

14.2.2 Propensity Score Weighting Method

We learn first what propensity score is.

Definition: Propensity Score The conditional probability of receiving the treatment is call 'propensity score'

$$e(w) = \Pr[d_i = 1 | w_i = w] = E[d_i | w_i = w]$$

How to estimate the propensity score: Use LPM, logit or probit, and estimate the propensity scores.

$$d_i = 1\{d_i^* = w_i\gamma + u_i > 0\}$$

$$e(\hat{w}_i) = F(w_i\hat{\gamma}) = \frac{\exp(w_i\hat{\gamma})}{1+\exp(w_i\hat{\gamma})} \text{ for a logit case}$$

Now, the average outcomes for treated and controls are given by

$$\hat{\tau} = \frac{\sum d_i y_i}{\sum d_i}_{|treated} - \frac{\sum (1-d_i) y_i}{\sum (1-d_i)}_{|controlled}$$

is biased.

Consider the following expectation of the simple weighting

$$E\left[\frac{d\cdot y}{e(w)}\right] = E\left[\frac{d\cdot y(1)}{e(w)}\right] = E\left\{E\left[\frac{d\cdot y(1)}{e(w)}|w\right]\right\} = E\left\{E\left[\frac{e(w)\cdot y(1)}{e(w)}\right]\right\} = E\{y(1)\}$$

Similarly we have

$$E\left[\frac{(1-d)y}{1-e(w)}\right] = E\{y(0)\},$$

which implies that

$$\hat{\tau}_p = \frac{1}{N}\sum_{i=1}^{n}\left\{\frac{d_i y_i}{e(w_i)} - \frac{(1-d_i)y_i}{1-e(w_i)}\right\} = \frac{1}{N}\sum_{i=1}^{n}\{\omega_i^\tau y_i - \omega_i^c y_i\}$$

where ω_i^τ is the weight for treated units.

However, this estimator is not an attractive estimator since the weight is not always one in the finite sample. To balance out the weight, we consider the following weighting over weighting estimator

$$\hat{\tau}_{pw} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{\omega_i^\tau}{\sum_{i=1}^n \omega_i^\tau} y_i\right) - \frac{1}{N}\sum_{i=1}^{n}\left(\frac{\omega_i^c}{\sum_{i=1}^n \omega_i^c} y_i\right).$$

Hirano, Imbens and Ridder (2003) show that this estimator is efficient.

14.2.3 Matching

There are several matching methods. Among them, propensity score matching is usually used. The idea of matching is simple. Suppose that a subject j in the controlled group has the same covariate value (w_j) with a subject i in the treated group. That is

$$w_i = w_j.$$

In this case, we can calculate the average treatment effect without any bias. Several matching methods are available. Alternatively we can use propensity score to match. That is

$$p_i = p_j.$$

Of course, it is inefficient if many observations should be dropped. Here I show only two examples of matching methods.

Exact Matching Drop subjects in treated and controlled if $w_i \neq w_j$ or $p_i \neq p_j$.

Propensity Matching 1. Cluster samples such that

$$\begin{aligned} |p_i - p_j| &< \varepsilon_1 \text{ for the first group} \\ \varepsilon_1 &\leq |p_i - p_j| < \varepsilon_2 \text{ for the second group} \\ &\vdots \end{aligned}$$

2. Calculate the average treatment effects by taking

$$\hat{\tau}_m = \frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{n_s}\sum_{i=1}^{n_s}\left(y_{is}\left(1\right) - y_{js}\left(0\right)\right)\right\}$$

14.3 Panel Treatment Effects

14.3.1 When $T = 2$

Example: (Card and Krueger, 1994) New Jersey raised the minimum wage in Jan. 1990. (I don't know the exact year). Meanwhile Pennsylvania didn't do so both in 1990 and 1991. In the below, the number of net employed persons are shown during these periods.

	Before	After	Difference
NJ	20.44	21.03	0.59
PENN	23.33	21.17	-2.16
Difference	-2.89	-0.14	2.76

Now estimate the effect of the higher minimum wage.

Let y_{it} be the outcome and x_{it} be treatment. Then for NJ, we have

$$\begin{aligned} E\left(y_{10}|x_{10}=0\right) &= \alpha_1 \\ E\left(y_{11}|x_{11}=1\right) &= \alpha_1+\gamma+\delta \end{aligned}$$

Meanwhile for Penn,

$$\begin{aligned} E\left(y_{20}|x_{20}=0\right) &= \alpha_2 \\ E\left(y_{21}|x_{21}=0\right) &= \alpha_2+\gamma \end{aligned}$$

Note that the withine difference is

$$\begin{aligned} E\left(y_{11}|x_{11}=1\right)-E\left(y_{10}|x_{10}=0\right) &= \gamma+\delta \\ E\left(y_{21}|x_{21}=0\right)-E\left(y_{20}|x_{20}=0\right) &= \gamma \end{aligned}$$

so that we can get the treatment effects by taking difference in difference.

$$\left[E\left(y_{11}|x_{11}=1\right)-E\left(y_{10}|x_{10}=0\right)\right]-\left[E\left(y_{21}|x_{21}=0\right)-E\left(y_{20}|x_{20}=0\right)\right]=\delta$$

In the regression context, we run

$$y_{it}=a+\beta S_i+\gamma t+\delta t x_{it}+\varepsilon_{it}, \quad \text{for } t=1,2$$

where S_i is a state dummy, t is trend.

Now for a large t, we consider a case of $(0,1,1)$ and $(0,0,0)$.

$$\begin{aligned} E\left(y_{10}|x_{10}=0\right) &= \alpha_1 \\ E\left(y_{11}|x_{11}=1\right) &= \alpha_1+\gamma_1+\delta \\ E\left(y_{12}|x_{12}=1\right) &= \alpha_1+\gamma_2+\delta \\ E\left(y_{20}|x_{20}=0\right) &= \alpha_2 \\ E\left(y_{21}|x_{21}=0\right) &= \alpha_2+\gamma_1 \\ E\left(y_{22}|x_{22}=0\right) &= \alpha_2+\gamma_2 \end{aligned}$$

Then we can estimate the ATE by

$$\begin{aligned} E\left(y_{11}|x_{11}=1\right)-E\left(y_{21}|x_{21}=0\right) &= \alpha_1-\alpha_2+\delta \\ E\left(y_{10}|x_{10}=0\right)-E\left(y_{20}|x_{20}=0\right) &= \alpha_1-\alpha_2 \end{aligned}$$

and also we can do

$$E\left(y_{12}|x_{12}=1\right)-E\left(y_{22}|x_{22}=0\right)=\alpha_1-\alpha_2+\delta.$$

Hence overall we can estimate the ATE by running

$$y_{it}=a_i+\theta_t+\delta x_{it}+\varepsilon_{it}$$