1 Identification

1.1 Heuristic Identification

1. "We don't have enough sample size to identify the causal effects of the problem.": As long as the identification set is singleton, we can identify the causal effects of the problem regardless of the sample size. However, smaller sample size might indeed make it harder for researchers to conclude "strong" identification, so researchers might end up with a less conclusive identification.

2. "We don't have a good identification strategy so I need to use a structural model.": It is not the case that "the more things you estimate, the better your identification is." A better strategy might be finding more useful observables for your identification process.

3. "Because I have a structural model, I don't need to think about identification.": Even if you have a structural model, you are still trying to figure out the causality. In this case, identification does matter so much.

4. "Because I can use the maximum likelihood estimator, I can identify that.": Generally, MLE is about "estimating" but not "identifying". Estimations from MLE do not guarantee identification.

1.2 Identification of OLS

Proof. We prove that β in this simple OLS model is identified by contradiction. So first, suppose that there exists $\beta^* \neq \beta$ belonging to the same identification set of β, i.e.

$$\beta^* x_i + \epsilon_i = \beta x_i + \epsilon_i$$

which implies $\beta^* = \beta (\Rightarrow\Leftarrow)$ □

1.3 Identification of a Factor Model

1. The only observable is $\{y_{it}\}$, so we start form it.

$$Cov(y_{it}, y_{it-1}) = Cov(v_{it} + \epsilon_{it}, v_{it-1} + \epsilon_{it-1})$$
$$= Cov(\rho v_{it-1} + \zeta_{it} + \epsilon_{it}, v_{it-1} + \epsilon_{it-1}) \tag{1}$$
$$= \rho Cov(v_{it-1}, v_{it-1}) \tag{2}$$
$$= \rho \sigma_v^2 \tag{3}$$

Note that form (1) to (2) we use the fact that ϵ_{it} and ζ_{it} are both i.i.d. across individuals, over time, and of each other.

Similiarly,

$$
\begin{aligned}
Cov(y_{it}, y_{it-2}) &= Cov(v_{it} + \epsilon_{it}, v_{it-2} + \epsilon_{it-2}) \\
&= Cov(\rho(\rho \cdot v_{it-2} + \zeta_{it-1}) + \zeta_{it} + \epsilon_{it}, v_{it-2} + \epsilon_{it-2}) \\
&= Cov(\rho^2 v_{it-2} + \rho\zeta_{it-1} + \zeta_{it} + \epsilon_{it}, v_{it-2} + \epsilon_{it-2}) \qquad (4) \\
&= \rho^2 Cov(v_{it-2}, v_{it-2}) \qquad (5) \\
&= \rho^2 \sigma_v^2
\end{aligned}
$$

Again from (4) to (5) we use the fact that ϵ_{it} and ζ_{it} are both i.i.d. across individuals, over time, and of each other. Therefore, ρ is identified since

$$
\rho = \frac{\rho^2 \sigma_v^2}{\rho \sigma_v^2} = \frac{Cov(y_{it}, y_{it-2})}{Cov(y_{it}, y_{it-1})}
$$

$\square$

2. Now we have already identified ρ, so we can think of using it to identify other parameters. First, notice that

$$
Var(y_{it}) = Var(v_{it} + \epsilon_{it}) = \sigma_v^2 + \sigma_\epsilon^2
$$

so we can think of identifying σ_ϵ^2 by first identifying σ_v^2. By equation (3) we can derive that

$$
\sigma_v^2 = \frac{Cov(y_{it}, y_{it-1})}{\rho}
$$

so σ_v^2 is identified. Then, σ_ϵ^2 can be identified by

$$
\sigma_\epsilon^2 = Var(y_{it}) - \sigma_\epsilon^2
$$

$\square$

3. Now we have already identified ρ, σ_v^2, and σ_ϵ^2, so we can think of using them to identify other parameters. First, notice that

$$
\begin{aligned}
Var(v_{it}) = Var(\rho v_{it-1} + \zeta_{it}) &= \rho^2 \sigma_v^2 + \sigma_\zeta^2 \\
\Rightarrow \sigma_v^2 &= \rho^2 \sigma_v^2 + \sigma_\zeta^2 \\
\Rightarrow (1 - \rho^2)\sigma_v^2 &= \sigma_\zeta^2 \\
\Rightarrow \sigma_v^2 &= \frac{\sigma_\zeta^2}{1 - \rho^2}
\end{aligned}
$$

σ_ζ^2 is thus identified. $\square$

4. We can calculate $Var(y_{it})$, $Cov(y_{it}, y_{it-1})$, and $Cov(y_{it}, y_{it-2})$. Then follow the precedures presented above, we can estimate these parameters.

1.4   Simulation of MLE

1. By the additive property of normal distribution,

$$y_i \sim N(0, \sigma_1^2 + \sigma_2^2)$$

so the probability density function (PDF) of y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{2}\left(\frac{y}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2}$$

Therefore, the likelihood function is

$$\mathcal{L}(\sigma_1^2 + \sigma_2^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{2}\left(\frac{y_i}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2}$$

Moreover, the log-likelihood function is

$$
\begin{aligned}
\ell(\sigma_1^2 + \sigma_2^2) &= \sum_{i=1}^{N} ln\left[\frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{2}\left(\frac{y_i}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2}\right] \\
&= \sum_{i=1}^{N} ln\left[\frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}}\right] + \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{y_i}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2\right] \\
&= -\frac{N}{2} ln(\sigma_1^2 + \sigma_2^2) - \frac{N}{2} ln 2\pi - \frac{1}{2(\sigma_1^2 + \sigma_2^2)} \sum_{i=1}^{N} y_i^2
\end{aligned}
$$

2. R script 1.4_MLE_Simulation.R is in my ECON-7069 repository [1]. Please find in the folder "PS2".

3. No, they can not be seperatly identified since the log-likelihood function can only identify parameters with observables. In this case, our only observable is $\{y_i\}_{i=1}^{N}$, which only enables us to identify $\sigma_1^2 + \sigma_2^2$.

---

[1] https://github.com/JayChang426/ECON-7069.git

4. Taking first order derivative of our log-likelihood function, first order condition is

$$
\frac{d\ell(\sigma_1^2 + \sigma_2^2)}{d(\sigma_1^2 + \sigma_2^2)} = -\frac{N}{2}\frac{1}{\sigma_1^2 + \sigma_2^2} - \frac{1}{2}\frac{-1}{(\sigma_1^2 + \sigma_2^2)^2}\sum_{i=1}^{N} y_i^2 = 0
$$

$$
\Rightarrow -\frac{N}{2} + \frac{1}{2}\frac{1}{\sigma_1^2 + \sigma_2^2}\sum_{i=1}^{N} y_i^2 = 0
$$

$$
\Rightarrow \frac{1}{\sigma_1^2 + \sigma_2^2}\sum_{i=1}^{N} y_i^2 = N
$$

$$
\Rightarrow \sigma_1^2 + \sigma_2^2 = \frac{\sum_{i=1}^{N} y_i^2}{N} \qquad \square
$$

Thus we can identify $\sigma_1^2 + \sigma_2^2$.

5. Even if we add up the results of $\sigma_1^2 + \sigma_2^2$ in 2. instead of taking them seperately, the sample size is still too small to make sense.

2   Potential Outcome Framework

1. With exact potential outcome notations in class, we can rewrite the Roy model into

$$
w_i = D_i \cdot w_{1i} + (1 - D_i) \cdot w_{0i}
$$

where w_i is often denoted by Y_i, w_{1i} is often denoted by $Y_i(1)$, and w_{0i} is often denoted by $Y_i(0)$.

2. As explained above,

   - $Y_i(0)$ is w_{0i}, which means the potential outcome (wage) without treatment (not to immigrate).
   - $Y_i(1)$ is w_{1i}, which means the potential outcome (wage) with treatment (immigrate).

3. D_i means the state of being treated (immigrate) or not.
   More specifically, D_i is a dummy that takes value of 1 when $w_{1i} > w_{0i}$, i.e. $Y_i(1) > Y_i(0)$ and takes value of 0 when $w_{1i} < w_{0i}$, i.e. $Y_i(1) < Y_i(0)$.

3   Control for Observables

3.1   Rosenbaum and Rubin

First, we need to define some notations following Rosenbaum and Rubin (1983):

1. Similar to what was defined in potential outcome framework, the ith observation has both a response r_{1i} under treatment 1, and a response r_{0i} under treatment 0, i.e. not treated.

2. For all observations, $Z_i = 1$ if treated and $Z_i = 0$ if not treated.

3. A balancing score, $b(x)$, is a function of the observed covariates x such that the conditional distribution of x given $b(x)$ is the same for treated ($Z_i = 1$) and control ($Z_i = 0$) groups, i.e.

$$x \perp Z \mid b(x)$$

4. The propensity score, $e(x) = Pr(Z_i = 1|x)$, which is the propensity towards exposure to treatment 1 given the observed covariates x.

Theorem 1. Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is,

$$x \perp Z \mid e(x)$$

Proof. Since **Theorem 1.** is just a special case of **Theorem 2.**, please see the proof of **Theorem 2.** as follows.

Theorem 2. Let $b(x)$ be a function of x, then $b(x)$ is a balancing score, i.e. $x \perp Z \mid b(x)$ if and only if ($\Longleftrightarrow$) $b(x)$ is finer than $e(x)$ in the sense that $e(x) = f\{b(x)\}$ for some function $f(\cdot)$.

Proof. We first prove the $\Leftarrow$) direction, i.e. if $b(x)$ is finer than $e(x)$ in the sense that $e(x) = f\{b(x)\}$ for some function $f(\cdot)$, then $b(x)$ is a balancing score.

To show $b(x)$ is a balancing score, it suffices to show that

$$Pr[Z = 1 \mid b(x)] = e(x) \tag{6}$$

because $e(x) = Pr(Z_i = 1|x)$, which is by construction a balancing score.

Now by the definition of $e(x)$ in point 4. of the notation definitions

$$Pr[Z = 1 \mid b(x)] = \mathbb{E}[e(x) \mid b(x)]$$

But since $b(x)$ is finer than $e(x)$, i.e. $e(x) = f\{b(x)\}$ for some function $f(\cdot)$,

$$\mathbb{E}[e(x) \mid b(x)] = \mathbb{E}\{e(x) \mid f^{-1}[e(x)]\} = e(x)$$

by $\mathbb{E}(X|X) = X$, and therefore equation (6) is satisfied.

For the opposite $\Rightarrow$) direction, i.e. if $b(x)$ is a balancing score, then $b(x)$ is finer than $e(x)$ in the sense that $e(x) = f\{b(x)\}$ for some function $f(\cdot)$.

Suppose that $b(x)$ is a balancing score but not finer than $e(x)$, so there exist x_1 and x_2 such that $e(x_1) \neq e(x_2)$ but $b(x_1) = b(x_2)$. Then $Pr(Z = 1 \mid x_1) \neq Pr(Z = 1 \mid x_2)$ and thus Z and x are not conditionally independent given $b(x)$, so $b(x)$ is not a balancing score. ($\Rightarrow\Leftarrow$) □

Theorem 3. If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any $b(x)$, that is,

$$(r_{1i}, r_{0i}) \perp Z \mid x \Rightarrow (r_{1i}, r_{0i}) \perp Z \mid b(x)$$

Proof. for the above statement to hold, we need to prove that

$$Pr[Z_i = 1 \mid r_{1i}, r_{0i}, b(x)] = Pr[Z_i = 1 \mid b(x)]$$

which by equation (6), is equivalent to show that

$$Pr[Z_i = 1 \mid r_{1i}, r_{0i}, b(x)] = e(x)$$

Now we start from

$$Pr[Z_i = 1 \mid r_{1i}, r_{0i}, b(x)] = \mathbb{E}[Pr(Z_i = 1 \mid r_{1i}, r_{0i}, x) \mid r_{1i}, r_{0i}, b(x)]$$

which by assumption equals $\mathbb{E}[Pr(Z_i = 1 \mid x) \mid r_{1i}, r_{0i}, b(x)]$ and by definition equals $\mathbb{E}[e(x) \mid r_{1i}, r_{0i}, b(x)]$. Finally, since $b(x)$ is finer than $e(x)$, it equals to $e(x)$. $\qquad\square$

3.2   Propensity Score

1. Let's pick $\beta_1 = 0.1$ and $\beta_2 = 0.05$.

2. X_1 can be total year of education.
   Define D_i as

   $$\begin{cases} D_i = 1 \ \text{ when } \ w_{1i} - w_{0i} \geq 0 \\ D_i = 0 \ \text{ when } \ w_{1i} - w_{0i} < 0 \end{cases}$$

   that is, $D_i = 1$ if

   $$\mu_1 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1 - (\mu_0 + \beta_1 X_1 + \epsilon_0) \geq 0$$
   $$\Rightarrow (\mu_1 - \mu_0) + \beta_2 X_2 + (\epsilon_1 - \epsilon_0) \geq 0$$

   Therefore, we cannot identify β_1.

3. We define propensity score $= Pr(D_i = 1 \mid X = x)$

4. Recall that we defined $v = \epsilon_1 - \epsilon_0$ in problem set 1.

Now according to the model here, we'll define $z = \frac{\mu_0 - \mu_1 - \beta_2 X_2}{\sigma_v}$.

$$
\begin{aligned}
Pr(D_i = 1 \mid X = x) &= Pr(\mu_1 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1 - (\mu_0 + \beta_1 X_1 + \epsilon_0) \geq 0 \mid X = x) \\
&= Pr(\epsilon_1 - \epsilon_0 \geq \mu_0 - \mu_1 - \beta_2 X_2 \mid X = x) \\
&= Pr(\frac{v}{\sigma_v} \geq \frac{\mu_0 - \mu_1 - \beta_2 X_2}{\sigma_v} \mid X = x) \\
&= 1 - \Phi(z \mid X = x)
\end{aligned}
$$

$\square$

5. R script 3.propensity_score_simulation.R is in my ECON-7069 repository [2]. Please find in the folder "PS2".

6. See the code in 3.propensity_score_simulation.R is in my ECON-7069 repository.

7. Correlation coefficient is 0.9996493, nearly 1.

8. For the code, see 3.propensity_score_simulation.R is in my ECON-7069 repository.
   When using the derived formula, the intercept is 2.5775018 and coefficient on I is 0.2587068.
   When using logit, the intercept is 2.5781721 and coefficient on I is 0.2546473.

9. For the code, see 3.propensity_score_simulation.R is in my ECON-7069 repository.
   The intercept is 2.5279400 and coefficient on I is 0.3592014, so we cannot recover the parameters.

10. For the code, see 3.propensity_score_simulation.R is in my ECON-7069 repository.
    The intercept is 2.22723110, the coefficient on I is 0.20549727, and the coefficient on X_2 is 0.03377549, so we still cannot recover the parameters.

---

[2] https://github.com/JayChang426/ECON-7069.git