HW2
Labor Economics

Chia-wei, Chen[*]

October 3, 2022

1 Identification

1.1 Heuristic Identification

1. *"We don't have enough sample size to identify the causal effects of the problem."*

   The sample size doesn't effect the identification. Only the standard error of the estimation is effected.

2. *"We don't have a good identification strategy so I need to use a structural model."*

   Having a structural model does not guarantee the identification of its parameters.

3. *"Because I have a structural model, I don't need to think about identification."*

   Same as above, consider a structural model with two indistinguishable clusters. If the two clusters are the same, the identification of the parameters is impossible.

4. *"Because I can use the maximum likelihood estimator, I can identify that."*

   Let's take a counterexample. Assume the maximum likelihood estimator that we constructed is flat around its global maximum (for some reason). The estimation is unidentified in this case.

[*]R10323045

1.2 Identification of OLS

Recall that

$$\hat{\beta} = (X'X)^{-1}X'Y$$

As long as $X'X$ is nonsingular, the estimator of OLS will be certain. It turns out that this is true if there exists no perfect multicollinearity.

1.3 Identification of Factor Model

Labeling the equations,

$$y_{i,t} = \nu_{i,t} + \epsilon_{i,t} \tag{1a}$$

$$\nu_{i,t} = \rho\nu_{i,t-1} + \xi_{i,t} \tag{1b}$$

1.3.1 ρ

Substituting Eq. (1b) into Eq. (1a), we get

$$y_{i,t} = \rho\nu_{i,t-1} + \xi_{i,t} + \epsilon_{i,t} \tag{2}$$

By Eq. (1a), we know $y_{i,t-1} = \nu_{i,t-1} + \epsilon_{i,t-1}$, hence by Eq. (2) we get

$$y_{i,t} = \rho y_{i,t-1} - \rho\epsilon_{i,t-1} + \epsilon_{i,t} + \xi_{i,t} \tag{3}$$

The linear relation between $y_{i,t}$ and $y_{i,t-1}$ guarantees that there cannot exist more than one ρs. It is then obvious that ρ is identified, and is $\rho = \mathbb{E}\left(\frac{y_{i,t}}{y_{i,t-1}}\right)$

1.3.2 σ_ϵ^2

ϵ^2 is the measurement error from each observation of an AR(1) process. Given a certain time moment, we can observe the covariance of $y_{i,t}$ and $y_{i,t-1}$

$$\begin{aligned} Cov(y_{i,t}, y_{i,t-1}) &= Cov(y_{i,t-1}, \rho y_{i,t-1} - \rho\epsilon_{i,t-1} + \epsilon_{i,t} + \xi_{i,t}) \\ &= \rho Var_i(y_{i,t}) - \rho Var(\epsilon_{i,t-1}) \\ &= \rho(Var_i(y_{i,t}) - \sigma_\epsilon^2) \end{aligned}$$

Since we have the variance of all data points from the last period, and ρ can be obtained, we can calculate σ_ϵ^2

1.3.3 σ_ξ^2

We now see the variance of $y_{i,t}$

$$\begin{aligned} Var(y_{i,t}) &= Var(\rho y_{i,t-1} - \rho\epsilon_{i,t-1} + \epsilon_{i,t} + \xi_{i,t}) \\ &= \rho^2 Var(y_{i,t-1} - \epsilon_{i,t-1}) + \sigma_\epsilon^2 + \sigma_\xi^2 \\ &= \rho^2 Var(y_{i,t-1}) + (1-\rho^2)\sigma_\epsilon^2 + \sigma_\xi^2 \end{aligned}$$

Again, we have all we need from the previous subquestions, therefore σ_ϵ can be identified.

1.3.4 Estimator

The estimators are provided above. Through all of them are not efficient, it is theoretically unbiased.

1.4 Identification of MLE

$$y_i = \epsilon_i^1 + \epsilon_i^2$$

1.4.1 Likelihood Function

The sum of two normally distributed random variables is also normally distributed, hence

$$y_i \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2) \tag{4}$$

Define $\sigma_y^2 = \sigma_1^2 + \sigma_2^2$. The likelihood function is then

$$f(\sigma_1^2, \sigma_2^2; \{y_i\}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{1}{2}\left(\frac{y_i}{\sigma_y}\right)^2} \tag{5}$$

1.4.2 Simulation

```
s1 = 2
s2 = 5

# DGE
ys = numeric(2)
for(i in 1:2){
```

```
  ys[i] = rnorm(1, 0, s1) + rnorm(1, 0, s2)
}

lh = function(sigma){
  orig = 0
  sigma_1 = sigma[1]
  sigma_2 = sigma[2]

  for(y in ys){
    orig = orig - log(dnorm(y, mean = 0, sd = sqrt(sigma_1^2 + sigma_2^2)))
  }
  return(orig)
}

op = optim(c(1,2), fn = lh, method = "BFGS")

op
```

The results are 3.227986 and 6.455974, which is not so close to the parameter I initially set.

1.4.3 Distinguish σ_1 from σ_2

σ_1 and σ_2 can be interchanged and the result will remain the same. This implies that for all possible element in the parameter set $\theta_0 \in \Theta$, there will be a interchanged counterpart of the parameter θ_0 as long as $\sigma_1 \neq \sigma_2$. Therefore the parameter set is not guaranteed to be singleton.

1.4.4 $\sigma_1^2 + \sigma_2^2$

The combined random variable y is a one-dimensional normal distribution. The standard deviation can be explicitly computed using the maximum likelihood function. It is hence identified.

1.4.5 Does the procedure in question 2 make sense?

Although the parameters are not uniquely determined, we still get the two standard deviations of the structural model. It is not identified, but sure it makes sense.

P.S. In non-parametric models such as the Gaussian mixture model, we estimate the properties as well as the proportion of clusters. It is also not identified, but as long as finding the clusters is the only thing we care about, it is meaningful.

2 Potential Outcome Framework

1. $w = D_1 w_1 + (1 - D_1) w_0$, substituting the definition, we get $w = D_1(\mu_0 + \epsilon_0) + (1 - D_1)(\mu_1 + \epsilon_1)$

2. $Y_i(0)$ is the wage of labor i if he works in 0, and $Y_i(1)$ is the wage of labor i if he works in 1.

3. D_i is the choice of labor i. One chooses 1 if $w_1 > w_0$, and vice versa (neglecting the cost).

3 Control for Observables

3.1 Proof of Rosenbaum and Rubin

Define the propensity score

$$\mathscr{P}(x) \equiv \Pr(\mathcal{D} = 1 \mid X = x) \tag{6}$$

Given the conditional independence assumption (CIA)

$$\{y_0, y_1\} \perp\!\!\!\perp x \tag{7}$$

We want to show that

$$\{y_0, y_1\} \perp\!\!\!\perp D \mid \mathscr{P}(x) \tag{8}$$

Intuitively, the propensity is a mapping from the space of X to $[0, 1]$, such that under this propensity, the potential outcome is independent of its choice d.

Proof. Consider $\Pr(\mathcal{D} = 1 \mid y_0, y_1, \mathscr{P}(x))$. By the law of iterated expectation, it is equivalent to

$$\mathbb{E}\left[\Pr(\mathcal{D} = 1 \mid y_0, y_1, \mathscr{P}(x), x) \mid y_0, y_1, \mathscr{P}(x)\right]$$

Since $\mathscr{P}(x)$ is a function of x, we can neglect $\mathscr{P}$

$$= \mathbb{E}\left[\Pr(\mathcal{D} = 1 \mid y_0, y_1, x) \mid y_0, y_1, \mathscr{P}(x)\right]$$

We assume CIA, as stated in Eq. (7), hence given x, the potential outcome is independent to the choice

$$= \mathbb{E}\left[\Pr(\mathcal{D} = 1 \mid x) \mid y_0, y_1, \mathscr{P}(x)\right]$$

Notice that by definition in Eq. (6), we can write

$$= \mathbb{E}\left[\mathscr{P}(x) \mid y_0, y_1, \mathscr{P}(x)\right] = \mathscr{P}(x)$$

that gives

$$\Pr(\mathcal{D} = 1 \mid y_0, y_1, \mathscr{P}(x)) = \mathscr{P}(x)$$

We have shown that $\mathcal{D}$ given the propensity score is completely independent of the potential outcomes, therefore prooving Eq. (8).

$\square$

3.2 Propensity Score Simulation

3.2.1 Code for Simulation

Note that this section covers all the subquestions in this problem.

```
library(MASS)
library(data.table)
library(dplyr)
library(xtable)
library(Matrix)
library(stargazer)
setwd("C:/Users/tedb0/Documents/111-1/Labor Economic/HW2/src")

    ## Parameters settings
N = 1e7
mu0 = 10
mu1 = 15
sigma0 = 3
sigma1 = 4.5
sigma_01 = 2
sigma_matrix = matrix(c(sigma0^2, sigma_01,
                        sigma_01, sigma1^2),
                      ncol=2)
C = 3

## ===== New Parameter

beta1 = 0.5
beta2 = -1
mux1 = 3
mux2 = 6

sigma_matrix = bdiag(sigma_matrix, diag(1,2,2))
```

```r

    ## Creating error terms, saving to data.table
wage = data.table(
  mvrnorm(n=N,mu=c(0, 0, mux1, mux2 ), Sigma=sigma_matrix)
)

  ## Rename
setnames(wage, "V1", "e0")
setnames(wage, "V2", "e1")
setnames(wage, "V3", "x1")
setnames(wage, "V4", "x2")

    ## Creating variables
wage[, W0 := e0 + x1* beta1 + mu0]
wage[, W1 := e1 + x1* beta1 + x2* beta2 + mu1]
wage[, Nu := e1 - e0]
wage[, D  := W1 > W0 + C]
wage[, W  := W1 * D + W0 * (1-D)]
    ## Calculate the theoretical value

## Calculate rho_nu
sigma_nu = sqrt( sigma0^2 + sigma1^2 - 2 * sigma_01 )

wage[, p_score_theor := dnorm((mu1 - mu0 + beta2 * x2 - C)/sigma_nu)
    ]

## Estimate logit to get propensity score

wage[, p_score_est := predict(
  glm(D~x2, family = binomial(link = "logit"))
  , type = "response"),]

cor_p_score = wage[,cor(p_score_theor, p_score_est),]

wage[, IPW_theor := W/ifelse(D, p_score_theor, -1 + p_score_theor)]
wage[, IPW_est :=  W/ifelse(D, p_score_est, -1 + p_score_est)]

ATE_theor = mu1 - mu0 + beta2 * mux2
ATE_prop_theor = wage[, mean(IPW_theor)]
ATE_prop_est = wage[, mean(IPW_est)]


model_ols_1 = wage[,lm(W~D)]
stargazer(model_ols_1,model_ols_2, out = "OLS_D.tex", table.
   placement = "h",
          keep.stat = c("rsq"), label = "tab:reg")
```

3.2.2 Example of X_1

X_1 is the common factor for both outcomes which has the same marginal effect. In the example of migration, think of X_1 as the experience, which in both countries increases its wage in the same scale.

β_1 does not affect the choice of outcomes, but it equally affects both w_0 and w_1, therefore it can be identified by looking at the value of w.

3.2.3 Define the Propensity Score

$$w_0 = \mu_0 + \beta_1 X_1 + \epsilon_0 \tag{9a}$$
$$w_1 = \mu_1 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1 \tag{9b}$$

The propensity score is defined as

$$\mathscr{P}(x_1, x_2) = \Pr(\mathcal{D} = 1 \mid X_1 = x_1, X_2 = x_2) \tag{10}$$

3.2.4 Derive the Propensity Score

Note that in Roy model, people choose $\mathcal{D} = 1$ if $w_1 > w_0 + C$, therefore

$$\begin{aligned}
\mathscr{P}(x_1, x_2) &= \Pr(w_1 > w_0 + C) \\
&= \Pr(\mu_1 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_1 > \mu_0 + \beta_1 x_1 + \epsilon_0 + C) \\
&= \Pr(\epsilon_0 - \epsilon_1 < \mu_1 - \mu_0 + \beta_2 x_2 - C) \\
&= \Pr\left(\frac{\nu}{\sigma_\nu} < \frac{\mu_1 - \mu_0 + \beta_2 x_2 - C}{\sigma_\nu}\right)
\end{aligned}$$

Where $\nu \equiv \epsilon_0 - \epsilon_1$

3.2.5 Theoretical Propensity Score

Demonstrated in code.

3.2.6 Propensity Score from Logit

Demonstrated in code.

3.2.7 Correlation between Theoretical and Empirical Propensity Score

The correlation coefficient in this simulation is 0.979.

3.2.8 Conduct IPW estimate

We construct IPW estimate by dividing w_0 with $1 - \mathscr{P}(x)$ and w_1 with $\mathscr{P}(x)$

The average treatment effect

$$\tau = \mathbb{E}(w_1 - w_0) = \mathbb{E}[\mathbb{E}(w_1 - w_0 \mid X)] = \mathbb{E}(\mu_1 - \mu_0 + \beta_2 X_2 + \epsilon_1 - \epsilon_0)$$

can then be estimated by

$$\mathbb{E}\left(\frac{w\mathcal{D}}{\mathscr{P}(X)}\right) - \mathbb{E}\left(\frac{w(1-\mathcal{D})}{1-\mathscr{P}(X)}\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d_i w_i}{\mathscr{P}(x_i)} - \frac{(1-d_i)w_i}{1-\mathscr{P}(x_i)}\right] \tag{11}$$

Do this for both the analytical propensity score and empirical propensity score, we have the estimated average treatment effect.

For the theoretical propensity score, I get -1.704; while for the empirical propensity score, I get 3.49

3.2.9 Regress on $\mathcal{D}$

3.2.10 Regress with Control Variable

See Table 1

Table 1

	Dependent variable:	
	W	
	(1)	(2)
D	3.570^{***}	3.506^{***}
	(0.002)	(0.002)
x1		0.500^{***}
		(0.001)
x2		-0.190^{***}
		(0.001)
Constant	12.014^{***}	11.668^{***}
	(0.001)	(0.006)
R^2	0.196	0.222
Note:	*p<0.1; **p<0.05; ***p<0.01	