ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE DE L'INFORMATION

ANNÉE 2023/2024

Applied Microeconometrics

Understand better econometrics applied to microeconomy

Auteur :
Cheryl Kouadio
Based on the courses of :
Yutec Sun

Contents

1 Introduction

Applied microeconometrics refers to the empirical analysis of microeconomic data, such as data on individuals, households, or firms, using econometric methods. Its main goal is to estimate the causal effect of economic decisions, and it often deals with samples that aren't exactly the same.

Understand causal effect is important because it helps to understand the impact of economic decisions on individual behavior and to evaluate the effectiveness of policies or interventions. It also helps to identify the factors that influence individual behavior and to predict the effects of changes in economic conditions or policies.

Microeconometric methods include different technique like randomized experiments, fixed-effects models, instrumental or control variables, regression discontinuity designs, matching estimators, and analysis of limited dependent variables. These methods are used to estimate the effects of policies, interventions, or other economic decisions on individual behavior, such as labor supply, consumption, savings, or investment.

Therefore, in this paper we are going to explore the main methods used in microeconometrics and how to apply them in R. This paper is based on classes taught by Sun Yutec at ENSAI.

2 Control variables

2.1 An example to start : causal effect of hospital treatment

Estimating causal effect of events could be a difficult task because of bias that may arise for different reasons. For example, if we want to estimate the causal effect of hospital treatment on health outcomes, we may face the problem of selection bias. Indeed, patients who receive hospital treatment may be different from those who don't receive treatment, and this difference may affect the outcome of the treatment. Suppose we have a dataset with the following variables : Y_i is the health outcome of individual i, D_i is a binary variable indicating whether individual i received hospital treatment. We can write Y_i like :

$$\begin{aligned}
Y_i &= \begin{cases} Y_{1i} & \text{si } D_i = 1 \text{ (i received the hospital treament)} \\ Y_{0i} & \text{si } D_i = 0 \text{ (i did not received the hospital treament)} \end{cases} \\
&= \begin{cases} Y_{1i} + Y_{0i} - Y_{0i} & \\ Y_{0i} & \text{si } D_i = 0 \end{cases} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})D_i
\end{aligned}$$

Using simply the average difference between the health outcomes of individuals who received hospital treatment and those who didn't may lead to biased estimates of the causal effect of hospital treatment. Indeed, the difference between the two groups may be due to differences in the characteristics of the individuals (for example, people who goes to the hospital are the one who have bad health conditions), rather than the effect of the treatment itself. In fact, we can not observe Y_{1i} and Y_{0i} for the same individual.

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{causal effect}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}}$$

To address this issue, one of the solutions is the randomize the treatment (D_i is given independently of Y_i). However, in practice, it is not always possible to randomize the treatment. Another method is to use control variables to adjust for differences in the characteristics of the individuals.

2.2 Control variables

Control variables are variables that are correlated with the treatment variable and the outcome variable, but are not affected by the treatment. By including control variables in the regression model, we can estimate the causal effect of the treatment while controlling for the effect of the control variables.

Let's consider the example of the causal effect of hospital treatment on health outcomes. The causal effect of hospital treatment on health outcomes can be estimated by β (the Average Treatment Effect - ATE) using the following regression model :

$$Y_i = \alpha + \beta D_i + \eta_i$$

The OLS estimation of β would be :

$$\hat{\beta} = \frac{\hat{\text{cov}}(D, Y)}{\hat{\text{var}}(D)}$$

Doing this, we assume that the treatment is the only factor that affects the health outcomes of individuals. However, in practice, there may be other factors that affect the health outcomes of individuals. Hence the OLS estimation of β would be biased. That's why we can include control variables in the regression model to adjust for these factors :

$$Y_i = \alpha + \beta D_i + X_i + \eta_i$$

where X_i is a vector of control variables that are correlated with the treatment variable and the outcome variable, but are not affected by the treatment. A good control variable is a variable that is correlated with the treatment variable and the outcome variable, but is not affected by the treatment. Formally, it respects the conditional independence assumption :

$$\{Y_{1i}, Y_{0i}\} \perp D_i | X_i \Leftrightarrow \eta_i \perp D_i | X_i$$

This implies that that the treatment is independent of the potential outcomes **given the control variables** or D_i only depends on Y_i only through X_i, but also the selection bias is removed. In practice, it is difficult to find control variables that satisfy this assumption. Valid control variables can be formulated from information used for making treatment decisions, or from information that is available before the treatment is assigned :

- $X \Rightarrow D \quad \text{and} \quad X \not\Leftarrow D$: the control variable is used to determine the treatment but not the inverse.
- Controls that happened before the treatment generally work well.

2.3 Bad controls setup

Controls variable that does not satisfy the conditional independence assumption are called bad controls. They are variables that are correlated with the treatment variable and the outcome variable, and are also affected by the treatment. Including bad controls in the regression model can lead to biased estimates of the causal effect of the treatment.

For example, assuming we want to estimate the causal effect of school education D_i on salary Y_i. To control from the effect of the individual's ability, we include the occupation of individuals X_i as a control variable.

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$
$$X_i = D_i X_{1i} + (1 - D_i) X_{0i}$$

Let's assume that D_i is randomly assigned. Hence, it's independent of Y_i and X_i. However, the occupation of individuals may be affected by the level of education of individuals. In this case, the occupation of individuals is a bad control, and including it in the regression model would lead to biased estimates of the causal effect of school education on salary. To see why this is a bad control, let's consider the following regression model :

$$Y_i = \alpha + \beta D_i + \gamma X_i + \eta_i$$

where we could estimate the treatment effect β by :

$$\begin{aligned}\beta &= E[Y_i|X_i = 1, D_i = 1] - E[Y_i|X_i = 0, D_i = 1] \\ &= E[Y_{1i}|X_{1i} = 1] - E[Y_{0i}|X_{0i} = 1] \quad \text{by independence of } D_i \\ &= \underbrace{E[Y_{1i} - Y_{0i}|X_{1i} = 1]}_{ATE} + \underbrace{E[Y_{0i}|X_{1i} = 1] - E[Y_{0i}|X_{0i} = 1]}_{\text{Selection bias}} \quad \text{by linearity of expectation}\end{aligned}$$

Selection bias will still exist even after including the control variable X_i in the regression model, which is a bad control.

3 Difference-in-differences

One of the methods used to estimate the causal effect of a treatment is the difference-in-differences (DID) method. The DID method is used to estimate the causal effect of a treatment by comparing the change in the outcome variable before and after the treatment between the treatment group and the control group. It's essentially used when we have cross-sectional data for many years (for example, ex and ante treatment like it's the case for survey data).

The first DID analysis was conduced by John Snow. He wanted to understand the cause of the cholera outbreak in London, which has caused tens millions of deaths in 1849.

He then collected sample of sub districts on the number of deaths due to cholera before and after the water pump of Lambeth where the other water pump remain at the same place, before and after treatment. It was a randomized experiment because the water pump was removed independently of the cholera outbreak. He found that the cholera death rates decreased significantly after the removal of the water pump in the areas that were served by the water pump, but not in the areas that were not served by the water pump. This led him to conclude that the water pump was the cause of the cholera outbreak.

Formally to understand the DID method, let's consider that the death outcome of cholera is Y_{it} for individual i at time t is generated from :

$$Y_{it} = \alpha + \beta T_t + \gamma T_t D_i + \epsilon_{it}$$

where T_t is indicator for time after treatment, D_i is the treatment indicator (water pump removal), and ϵ_{it} is the error term, which means the permanent difference between treated and control samples.

Then the ATE (γ)is estimated by DID as:

$$\gamma = (E[Y_{it}|D_i = 1, T_t = 1] - E[Y_{it}|D_i = 1, T_t = 0]) - (E[Y_{it}|D_i = 0, T_t = 1] - E[Y_{it}|D_i = 0, T_t = 0])$$

We make difference between the treated and control group before and after the treatment. One of the assumptions of the DID method is the parallel trend assumption, which states that the trends in the outcome variable before the treatment are the same between the treatment group and the control group. This assumption is important because it allows us to estimate the causal effect of the treatment by comparing the change in the outcome variable before and after the treatment between the treatment group and the control group. It depends on the exogeneity of the treatments which could be obtain when treatment is randomly assigned (example natural events or exogeneous policy etc.)

DID fails to work when the parallel trend assumption is violated. In this case, the DID estimate of the causal effect of the treatment would be biased or when the error term is correlated over time like in panel data. It means that the permanent difference between treated and control samples on t time is correlated with the existing difference on t-1 time.

4 Matching estimators

To ensure that treatment decision does not depend on no unobservable characteristics, we can use matching estimators based on non-experimental data. Matching estimators are used to estimate the causal effect of a treatment by matching treated and control units that are similar in terms of observable characteristics. The idea is to create a control group that is similar to the treatment group in terms of observable characteristics **that satisfy the Conditional Independence assumption**, so that the difference in the outcome variable between the treatment group and the control group can be attributed to the treatment. Matching approaches estimates the mean of : $Y_{1i} - \hat{Y_{0i}}$ for individuals i among the treated group $I_1 = \{i : D_i = 1\}$. $\hat{Y_{0i}}$ is the counterfactual outcome for individual i if he had not received the treatment, thus i's hypothetical clone matched on the basis of observed characteristics X_i. It is estimated by the outcome of the control group $I_0 = \{i : D_i = 0\}$ that are similar to individual i in terms of observable characteristics :

$$\hat{Y_{0i}} = \sum_{j \in I_0} W(i,j) Y_{0j}, \quad i \in I_1$$

where $W(i,j)$ is the weight assigned to the control unit j for the treated unit i. The weight is a function of the distance between the treated unit i and the control unit j in terms of observable characteristics. The weight can be calculated using different methods.

4.1 Example of matching

We want to estimate the causal effect of military service D_i on wage Y_i for veterans defined as :

$$E[Y_{1i} - Y_{0i} | D_i = 1]$$

, where $D_i = 1$ for veterans and $D_i = 0$ for non-veterans. For the same reason as the one mentioned in section 2, we can not estimate the causal effect of military service on wage by simply comparing the average wage of veterans and non-veterans. Indeed, veterans may be different from non-veterans in terms of observable characteristics, such as age, education, or experience, and this difference may affect their wage. If there exist observable characteristic X_i that satisfy the Conditional Independence assumption, we can use it as matching estimators to match veterans and non-veterans that are similar in terms of observable characteristics.

Formally, the matching approach estimates the treatment effect on the treated by :

$$\begin{aligned}
\Delta_{TOT} &= E[Y_{1i} - Y_{0i} | D_i = 1] \\
&= E[E[Y_{1i} - Y_{0i} | X_i, D_i = 1] | D_i = 1] \\
&= E[E[Y_{1i} | X_i, D_i = 1] - E[Y_{0i} | X_i, D_i = 0] | D_i = 1] \quad \text{by CI} \\
&= E[\delta_{X_i} | D_i = 1]
\end{aligned}$$

where δ_{X_i} is earnings gap between the treated & control groups with characteristics X. Hence, matching is about finding for each treated its untreated clone with similar characteristics X.

However, matching is straightforward when we have a small number of treated units and a large number of control units. But it becomes more difficult when we have a large number of treated units and a small number of control units. In this case, we can use the propensity score matching, which is a method that estimates the probability of receiving the treatment based on observable characteristics, and then matches treated and control units with similar propensity scores.

4.2 Propensity score matching

Matching estimates of treatment effect

$$\Delta^{PSM} = \sum_x \delta_x W(x)$$

are obtained by separately estimating :

$$\begin{aligned}\delta_x &= E[Y_{1i}|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] \\ W(x) &= P(X_i = x|D_i = 1) \quad \text{or} P(X_i = x)\end{aligned}$$

for sample. Various methods can be used for $W(x)$ estimation (logistic or kernel regression, nearest neighbor, random forest, ...).

But in practice, $W(x)$ estimation can quickly become infeasible as more controls are included in X_i. Propensity score matching is a method that solves this problem by reformulating it into single-dimensional matching.

If the CI assumption holds for X_i, then

$$\{Y_{0i}, Y_{1i}\} \perp D_i|X_i \Leftrightarrow \{Y_{0i}, Y_{1i}\} \perp D_i|P(X_i)$$

where $P(X_i) = E[D_i|X_i]$ is the propensity score.

Hence matching can be based on propensity score $P(X_i)$ instead of X_i.

PSM with $P(X_i)$ works in the same way as matching with X_i. To see this, consider how the PSM estimates the treatment effect on the treated :

$$\begin{aligned}\Delta^{PSM} &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[E[Y_i|P(X_i), D_i = 1] - E[Y_i|P(X_i), D_i = 0]|D_i = 1] \\ &= E[\delta_{P(X_i)}|D_i = 1]\end{aligned}$$

where $\delta_{P(X_i)}$ is the earnings gap between the treated and control groups with similar propensity scores. To estimate it, the PSM need two assumptions :

- Conditional Independence : $\{Y_{0i}, Y_{1i}\} \perp D_i|P(X_i) \Leftrightarrow E[Y_d|P(X), D = 1] = E[Y_d|P(X), D = 0]$
- Common support for propensity score : $0 < P(X_i) < 1$. It implies that matching should include only samples with common X where both treated and control groups can be found.

Formally, the PSM estimator of $\Delta^{PSM} = E[Y_{1i} - Y_{0i}|D_i = 1]$ can be defined as :

$$\hat{\Delta}^{PSM} = \frac{1}{N_1} \sum_{i \in \{i:D_i=1\}} [Y_{1i}(X_i) - \hat{E}[Y_{0i}|P(X_i), D_i = 0]]$$

where N_1 is the number of treated individuals with common support on $P(X)$ and $\hat{E}[Y_{0i}|P(X_i), D_i = 0]$ is the predicted Y_{0i} conditional on $P(X_i)$ for the control group, and can be estimated non-parametrically. Remark that the propensity score need to be estimated befor computing the PSM estimator.

There are two ways of implementing matching estimator depending on the data :

1. **Cross-sectional data**
   - Single snapshot of population among which some individuals are treated.
   - Propensity score matching
2. **Repeated cross-sectional data**
   - Multiple snapshots before and after the treatment occurs for some individuals.
   - Difference-in-differences

4.3 Matching and regression

Regression estimates the treatment effect δ_R in the equation :

$$Y_i = \sum x d_{ix} \alpha_x + \delta_R D_i + \epsilon_i$$

where $d_{ix} = \nVdash(X_i = x)$ is the dummy variable for each level of X_i. The regression estimates the treatment effect δ_R by :

$$\begin{aligned}
\delta_R &= \frac{Cov(Y_i, \tilde{D_i})}{Var(\tilde{D_i})} \\
&= \frac{E[(D_i - E[D_i|X_i])Y_i]}{E[(D_i - E[D_i|X_i])^2]} \\
&= \frac{E[(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]]}{E[(D_i - E[D_i|X_i])^2]} \quad \text{by iterated expectations.} \qquad (1)
\end{aligned}$$

Regression of Y_i on D_i & X_i is equivalent to the regression of CEF $E[Y_i|D_i, X_i]$.
Since we have:

$$E[Y_i|D_i, X_i] = E[Y_i|D_i = 0, X_i] + \delta_X D_i,$$

the numerator in Equation (1) becomes:

$$\begin{aligned}
&E\left[(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\right] \\
&= E\left[(D_i - E[D_i|X_i])E[Y_i|D_i = 0, X_i]\right] + E\left[(D_i - E[D_i|X_i])D_i\delta_X\right] \\
&= E\left[(D_i - E[D_i|X_i])D_i\delta_X\right] \quad \text{since } D_i \perp\!\!\!\perp E[Y_i|D_i = 0, X_i] \\
&= E\left[(D_i - E[D_i|X_i])^2\delta_X\right],
\end{aligned}$$

where the last equality follows from:

$$\begin{aligned}
E\left[D_i E[D_i|X_i]\delta_X\right] &= E\left[E\left[D_i E[D_i|X_i]\delta_X|X_i\right]\right] \\
&= E\left[E[D_i|X_i]^2\delta_X\right] \quad \text{since } E[D_i|X_i]\delta_X \in \sigma(X_i).
\end{aligned}$$

By substituting the numerator, Equation (1) becomes:

$$\begin{aligned}
\delta_R &= \frac{E\left[(D_i - E[D_i|X_i])^2\delta_X\right]}{E\left[(D_i - E[D_i|X_i])^2\right]} \\
&= \frac{E\left[E\left[(D_i - E[D_i|X_i])^2|X_i\right]\delta_X\right]}{E\left[E\left[(D_i - E[D_i|X_i])^2|X_i\right]\right]} \\
&= \frac{E\left[\sigma_D^2(X_i)\delta_X\right]}{E\left[\sigma_D^2(X_i)\right]},
\end{aligned}$$

where $\sigma_D^2(X_i) = E\left[(D_i - E[D_i|X_i])^2|X_i\right]$ is the conditional variance of D_i given X_i. Since D_i is binary, $\sigma_D^2 = \Pr(D_i = 1|X_i)\left[1 - \Pr(D_i = 1|X_i)\right]$.

How does δ_R compare with δ_{TOT}?

4.3.1 Difference between matching and regression

Define $Pr(d|x) = Pr(D_i = d|X_i = x)$. Then the treatment effects are calculated as follows:

Regression:

$$\delta_R = \frac{\sum_x \delta_x Pr(1|x)[1 - Pr(1|x)]Pr(X_i = x)}{\sum_x Pr(1|x)[1 - Pr(1|x)]Pr(X_i = x)}$$

Matching:

$$\begin{aligned}\delta_{\text{TOT}} &= \sum_x \delta_x Pr(X_i = x|D_i = 1) \\ &= \sum_x \delta_x Pr(1|x)Pr(X_i = x) \div Pr(D_i = 1) \quad \text{by Bayes' rule} \\ &= \frac{\sum_x \delta_x Pr(1|x)Pr(X_i = x)}{\sum_x Pr(1|x)Pr(X_i = x)}\end{aligned}$$

Difference:

- δ_R puts more weight on groups of x with high variance of treatment.
- δ_{TOT} puts more weight on groups of x with high likelihood of treatment.

Both estimations imply a common support, which means they only include groups of x where both treated and control groups can be found.

5 Endogeneity

Suppose a structural causal model given by :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$

where u is the **unobservable** random error term, y, x_1, ..., x_K the **observable** random variables.

Under the zero conditional mean assumption, the regression can be obtained and the OLS can consistently estimate $\beta_j, \text{for} j = 0, \ldots, K$.

Hence, we define an explanatory variable x_j as **endogeneous** if $E[u|x_1, \ldots, x_K]$. In other term, x_j is correlated with u. This typically occurs when there is one (or all) of these problems:

- omitted variables problems
- measurement error
- simultaneity : x_j is simultaneously determined by y (eg. demand and supply)

To correct for endogeneity bias, we can use instrumental variables (IV), control function approach, limited information maximum likelihood (LIML). These approaches are based on reduced-form model but we can also apply specified structural model.

5.1 Causes of endogeneity bias

5.1.1 Omitted variables

Consider the **real** structural model of the log wage y specified as :

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \nu$$

For estimating the model, suppose that we omit the variable x_2 and use the model :

$$\log(y) = \beta_0 + \beta_1 x_1 + u$$

Our estimation of the error u would be : $u = \beta_2 x_2 + \nu$
Hence, our OLS estimate β_1 will be biased because :

$$\begin{aligned}\hat{\beta_1} &= \frac{Cov(\log(y), x_1)}{Var(x_1)} \\ &= \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var x_1} \\ &\neq \beta_1 \quad \text{if } Cov(x_1, x_2) \neq 0\end{aligned}$$

5.1.2 Measurement error

Suppose we want to estimate the effect of alcohol consumption x on wage y but we can only measure alcohol consumption with measurement error $\tilde{x} = x + u$. The measurement error is i.i.d with $E[u|x] = 0$. Then the model can be expressed as :

$$\begin{aligned}y &= \beta x + \epsilon \\ &= \beta(\tilde{x} - u) + \epsilon \\ &= \beta\tilde{x} - \beta u + \epsilon\end{aligned}$$

The OLS estimate of β will be biased because of the measirement error. Indeed, the OLS estimate of β will be :

$$\hat{\beta} = \beta - \beta \frac{Var(u)}{Var(\tilde{x})}$$

5.1.3 Simultaneity

Suppose that wage y and alcohol consumption x are simultaneously determined as :

$$\begin{aligned}&y = \beta x + z_1 + u \quad (*) \\ &x = \gamma y + v, \\ &x = \gamma(\beta x + z_1 + u) + v \Rightarrow x = \frac{\gamma z_1 + \gamma u + v}{1 - \gamma\beta}\end{aligned}$$

Thus, x is endogenous to the model ($Cov(x, u) \neq 0$).

5.2 Correct for endogeneity bias

5.2.1 Instrumental variables (IV)

Consider that we estimate the model :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where x_2 is endogenous. We can use an instrumental variable z that is correlated with x_2 but not with u to estimate the model. The IV method relies on an observable variable z_1 satisfying the following conditions :

- Exlusion restriction : $Cov(z_1, u) = 0$, this implies that z_1 affect y only through x_2
- Full rank : z_1 is correlated with x_2 : $Cov(z_1, x_2) \neq 0$

Some existing methods can also be viewed as different kinds of IV. For example :

1. Random experiment
   - Randomly assigned treatment satisfies the exclusion restriction and rank conditions because the treatment itself is the IV (no endogeneity exists).
2. Quasi-experiment (natural experiment)
   - Endogenous treatment is shifted by an exogenous event that satisfies the exclusion restriction and rank conditions.

For a single endogenous variable x_K , only one instrument z_1 is sufficient. Such case is called a **just identified** model since it has just the right number of IVs to identify the parameters $\beta_0, \ldots, \beta_K$. The model with multiple instruments $(z_1, \ldots, z_M$) is called an **overidentified** model since it has more than enough IVs to identify the model parameter.

For a model with multiple instruments, we used the two-stage least squares (2SLS) method to estimate the model. The 2SLS method consists of two stages :

1. First stage : regress the endogenous variable x_K on the instruments $z_1, \ldots, z_M$ to obtain the predicted value $\hat{x}_K$.
2. Second stage : regress the dependent variable y on the predicted value $\hat{x}_K$ and the other exogenous variables $x_1, \ldots, x_{K-1}$ to estimate the model.

To test for rank condition, we use the F-test in the first stage regression. The null hypothesis is that the instruments are not correlated with the endogenous variable. If the null hypothesis is rejected, then the instruments are valid.

The 2SLS bias depends on the 1st-stage F-statistic, which measures the strength of the IVs.

- When IVs are weak, F tends to 0, leading to a large bias in 2SLS.
- Weak IVs also tend to fail the rank condition.

Hence, it is important to test for the weak IVs in practice, via the F test at least.

6 Unobserved heterogeneity in panel data

Panel data is a type of data that contains observations on multiple individuals over multiple time periods. Panel data is useful for estimating the causal effect of a treatment because it allows us to control for unobserved heterogeneity that may affect the outcome variable.

Unobserved heterogeneity refers to the differences in the characteristics of individuals that are not observed in the data, but may affect the outcome variable.

In panel data analysis, one of the most widely used models looks like :

$$y_{it} = \beta_0 + x_{it}\beta + c_i + u_{it}$$

where c_i is the effect of unobserved heterogeneity that is constant over time for individual i, and u_{it} is the error term. The unobserved heterogeneity c_i may be correlated with the treatment variable x_{it}, which can lead to biased estimates of the causal effect of the treatment.

Failing to account for unobserved heterogeneity in panel data analysis can lead to :

- inefficient estimator
- endogeneity bias
- poor prediction

These problems can be addressed by using fixed effects models or random effects models and hierarchical Bayesian methods.

Let's consider estimating a 2-period panel model :

$$y_{it} = \beta_0 + x_{it}\beta + c_i + u_{it}, \quad t = 1, 2$$

The necessary condition for the OLS estimator to be consistent is that $E(u_{it}|x_{it}, c_i) = 0$ (mean independence condition).

This condition is difficult to satisfy in panel data analysis. But as we have panel data, we can use differencing to remove the unobserved heterogeneity c_i.

Let $\Delta y = y_2 - y_1, \Delta x = x_2 - x_1, \Delta u = u_2 - u_1$, c can be differenced out as :

$$\Delta y = \Delta x\beta + \Delta u$$

The condition for the OLS estimator to be unbiased is $E(\Delta u|\Delta x) = 0$.

The estimator of first difference is a special case of fixed effect models that is used to estimate the causal effect of a treatment in panel data analysis. The fixed effect model is used to estimate the effect of a treatment by controlling for unobserved heterogeneity that is constant over time for individual i. Attention should be paid to the fact that the unobserved heterogeneity can be variable over time for individual i (random effect model).

6.1 Fixed effect model

In the fixed effect model, the unobserved heterogeneity c_i is treated as a fixed effect that is constant over time for individual i.

The fixed effect model is used to estimate the causal effect of a treatment by controlling for unobserved heterogeneity that is constant over time for individual i. By differencing the data, we can remove the unobserved heterogeneity c_i from the model, and estimate the causal effect of the treatment by comparing the change in the outcome variable before and after the treatment between the treatment group and the control group. It is consistent if c_i and x_i are correlated under strict exogeneity condition $E(u_{it}|x_{i1}, \ldots, x_{iT}, c_i) = 0$ for $t = 1, \ldots, T$.

Another method is the within estimator that is used in two-step. First of all, we take the mean of the variable for each individual on th T-period :

$$\bar{y}_i = \beta_0 + \bar{x}_i'\beta + a_i + \bar{u}_i$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$.

Secondly, we susbtract this mean to the initial equation :

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i')\beta + (u_{it} - \bar{u}_i)$$

Hence, the within estimator is obtained by OLS on the second equation. It is unbiased if $E(u_{it} - \bar{u}_i | x_{it} - \bar{x}_i) = 0$, which is the case when the strict exogeneity condition is satisfied. The hypothesis is not satisfied in models with "feedbacks", for example when x_{it} is correlated with u_{it-1}. Also, in models when y at the time t influences y at the time $t+1$.

6.2 Random effect model

The random effects(RE) approach accounts for the unobserved heterogeneity by allowing for heteroscedasticity in the unobservables to improve estimation efficiency when unobserved heterogeneity does not generate endogeneity bias.

In addition to the strict exogeneity condition, we need the exogeneity of the unobserved heterogeneity c_i with the regressors x_{it}, which is $E(c_i | x_{i1}, \ldots, x_{iT}) = 0$. It implies that the unobserved heterogeneity is not correlated with the regressors.

6.2.1 Which to choose between RE and FE?

If x_{it} is exogenous to c_i, both RE and FE are consistent. But RE is more efficient than FE. If x_{it} is endogenous to c_i, FE is consistent but RE is not. In this case, we should expect to see a difference between RE & FE.

Hausman proposes a test whether to choose between RE and FE. The null hypothesis is that the RE and FE are consistent (but the RE might be more efficient), while the alternative hypothesis is that the FE is consistent. Hence, if both estimator are different, we should choose the FE estimator (reject H_0).

6.3 Hierarchical Bayesian model

Often, we want to estimate each individual's heterogeneity to improve model prediction. Some examples of question are "which movie to recommend to individual Netflix viewer" etc.

How can we estimate each individual's β_i in the model $y_{it} = x_{it}\beta_i + u_{it}$?

Classical approach allows limited heterogeneity in β_i as a function of observed demographics. Hierarchical Bayesian revolutionized the way to estimate rich heterogeneity distribution of each β_i.

Model is based on the following assumptions :

- $y_{it} = x_{it}\beta_i + u_{it}$, where $u_{it} \sim N(0, \sigma^2)$
- $\beta_i \sim N(\mu_i, \sigma_i^2$
- $\mu \sim N(\mu_0, \sigma_0^2)$
- $\sigma_i^2 \sim IG(a, b)$

HB model makes it possible to estimate posterior for each individual β_i. The priors $(\mu_i, \sigma_i)_i$ are grouped by similar individuals automatically. It is a powerful tool to estimate individual heterogeneity.

7 Discrete choice models

A large number of economic decisions are a choice among finite alternative, hence called discrete choice. For example, the choice of a consumer between different products, the choice of a worker between different jobs, the choice of a student between different schools, etc.

Discrete choice models allow us to estimate the causal effects of treatment when outcomes are discrete choices.

7.1 Framework

The set of alternatives to choose have one of the properties :

- Mutually exclusive : the individual can choose only one alternative
- Exhaustive : the individual must choose one alternative
- Ordered : the alternatives are ordered
- Unordered : the alternatives are not ordered

For example, the choice of a consumer between different products is a mutually exclusive and exhaustive choice. The choice of a worker between different jobs might be an ordered choice. In economics and marketing, discrete choice models are used to estimate consumer demand.

Suppose that decision maker i chooses outcome $Y_i \in \{1, \dots, J\}$ from a set of J alternatives that gives the highest utility U_{ij} :

$$Y_i = j \Leftrightarrow U_{ij} > U_{ik} \quad \forall k \neq j$$

In the random utility framework, the utility of alternative j for individual i is given by :

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

where V_{ij} is the systematic component of utility that depends on the characteristics of the alternatives and the decision maker, and ϵ_{ij} is the random component of utility that captures the unobserved factors that affect the decision maker's choice (individual taste, random utility shock).

In the data, individuals' choice Y_{ij} can be observed, but the utility U_{ij} is not observed.

Under the utility framework, the choice probability is defined as:

$$\begin{aligned} P_{ij} &= P(Y_i = j) \\ &= P(U_{ij} > U_{ik} \quad \forall k \neq j) \\ &= P(V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik} \quad \forall k \neq j) \\ &= P(\epsilon_{ij} - \epsilon_{ik} < V_{ij} - V_{ik} \quad \forall k \neq j) \\ &= \int I(\epsilon_{ij} - \epsilon_{ik} < V_{ij} - V_{ik} \quad \forall k \neq j) F(d\epsilon_i) \end{aligned}$$

where $F(d\epsilon_i)$ is the distribution of the random utility shocks ϵ_i.

Ultimately, we estimate $\{V_{ij}\}_{i,j}$ which matches the choice probabilities $\{P_{ij}\}_{i,j}$ observed in data. Discrete choice models are categorized by the distribution F (e.g., logit, nested logit, probit, mixed logit, hierarchical Bayesian).

In discrete choice models, there is a limit on what model parameters can be estimated. It is important to understand what parameters can and cannot be estimated in discrete choice models. A discrete choice model is **identified** if there exists a unique utility V that best fits choice probabilities P observed in data $\mathcal{D}$.

We aim to find $\{V_{ij}\}_{i,j}$ that match $\mathcal{D}$ most closely. But are they unique?

Consider the definition of choice probability P_{ij} :

$$P_{ij} = \int I(\epsilon_{ij} - \epsilon_{ik} < V_{ij} - V_{ik} \quad \forall k \neq j) F(d\epsilon_i)$$

It depends only on relative differences $(V_{ij} - V_{ik})_{k \neq j}$ but not on absolute levels V_{ij}. Hence, from P_{ij}, we can estimate latent utilities V_{ij} only up to relative levels.

It implies that with J alternatives, we can estimate $J-1$ parameters. The model is not identified if we estimate J parameters.

7.1.1 Implication

Consider utilities of transportation between car (c) and bus (b):

$$\begin{aligned} U_c &= \alpha T_c + \beta M_c + k_c^0 + \epsilon_c, \\ U_b &= \alpha T_b + \beta M_b + k_b^0 + \epsilon_b, \end{aligned}$$

where T_j and M_j are time and monetary costs for $j \in \{c, b\}$, respectively. k_j^0 is the fixed effect for each j.

The choice probabilities (P_c, P_b) remain unchanged for any other (k_c', k_b') if:

$$k_b' - k_c' = k_b^0 - k_c^0$$

Hence, there exist no unique (k_c, k_b) for (P_c, P_b). But if one of the constants is normalized to $k_c = 0$,

$$\begin{aligned} U_c &= \alpha T_c + \beta M_c + \epsilon_c, \\ U_b &= \alpha T_b + \beta M_b + k_b + \epsilon_b, \end{aligned}$$

and we can find unique $k_b = k_b^0 - k_c^0$ for the observed (P_c, P_b).

$\Rightarrow$ With J alternatives, up to $J-1$ constants can be identified in discrete choice models.

7.2 Logit model

Suppose that a decision maker i receives utility U_{ij} from alternative j that is composed of observable systematic component V_{ij} and a random unobservable component ϵ_{ij} :

$$U_{ij} = V_{ij} + \epsilon_{ij} \quad \forall j$$

The logit model assumes that ϵ_{ij} is i.i.d as a type I extreme value, which has density function :

$$f(\epsilon_{ij}) = e^{-\epsilon_{ij}} e^{-e^{-\epsilon_{ij}}}.$$

Hence, the cdf is $F(\epsilon_{ij}) = e^{-e^{-\epsilon_{ij}}}$. We remark that the variance of ϵ_{ij} is $\pi^2/6.$, which implicitly normalizes the scales of utility. The unobserved utility ϵ_{ij} is idiosyncratic (independently distributed) and unrelated to any other alternatives.

Under the logit model assumptions, the choice probability takes a simple closed form :

$$P_{ij} = P(Y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

In statistics, the same model is used by multinomial logistic regression for classification problems.

For any two alternatives $j\&k$, the ratio of the logit choice probs is :

$$\frac{P_{ij}}{P_{ik}} = \frac{\frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}}{\frac{e^{V_{ik}}}{\sum_{k=1}^{J} e^{V_{ik}}}} = e^{V_{ij}-V_{ik}}$$

The ratio is independent of $\{V_l\}_{l\neq j,k}$ (alternatives other than $j\&k$), which implies that what happens to V_i for $i \notin (j,k)$ does not change $\frac{P_{ij}}{P_{ik}}$. It is called the independence of irrelevant alternatives (IIA) property.

7.2.1 Implication

Consider the 1st-round vote shares in the 2022 presidential elect in France : Mélenchon (1/3-ϵ), Macron (1/3), Le Pen (1/3). Share of their 1st-round votes are equally divided :

$$P_{\text{Mélenchon}} = P_{\text{Macron}} = P_{\text{Le Pen}} = \frac{1}{3}$$

.

Let's assume Mr. Mélenchon lost by $\epsilon \approx 0$. When Mélenchon is out of the race, the logit choice probs must satisfy the IIA property :

$$\frac{P_{\text{Macron}}}{P_{\text{Le Pen}}} = 1$$

Hence, the logit model predicts the 2nd-round vote shares as (when assuming zero abstention) :

$$P_{\text{Macron}} = P_{\text{Le Pen}} = \frac{1}{2}$$

The logit model permits proportional substitution between alternatives. It is a strong assumption that may not hold in reality. For Mélenchon's voters, they may not vote for Le Pen or Macron depending on their individual level preferences.

For example, let $y_{ij} = 1$ if individual i chooses alternative j and 0 otherwise. Suppose parameter β enters the model as $V_{ij} = x_{ij}\beta$.

The logit model is estimated by maximizing the log-likelihood function :

$$\mathcal{L}(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log P_{ij}$$

The FOC of the log-likelihood function gives the MLE estimator of β. We get :

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} x_{ij}y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} x_{ij}P_{ij}$$

Hence, the MLE estimator of β matches averages on the LHS and RHS based on model prediction P_{ij} (RHS=average of x_{ij} weighted by P_{ij}, LHS=average of x_{ij} weighted by y_{ij}).

Suppose that x_{ij} is $k-$specific indicator. Then the FOC becomes :

$$\frac{1}{N}\sum_{i=1}^{N} y_{ik} = \frac{1}{N}\sum_{i=1}^{N} P_{ik}$$

It implies that in this logit model, the share of people choosing k is equal to the share predicted at $\hat{\beta}_{MLE}$.

7.2.2 Estimation of logit choice model in aggregate-level data

Often individual choice data are aggregated at choice levels $j = 0, 1, \ldots, J$ to estimate the logit model. The choice probabilities are estimated by the share of people choosing each alternative :

$$P_j = \frac{1}{N} \sum_{i=1}^{N} P_{ij}, \quad j = 0, 1, \ldots, J$$

To simplify, let's assume that $P_{ij} = P_j$ for all i. Then we can obtain the log odds ratio as :

$$\log \frac{P_j}{P_0} = x_j \beta + \xi_j$$

where x_j is the characteristic of alternative j and β is the MLE estimator of the logit model and ξ_j is the utility unobserved to econometrician.

Then the regression method can be applied (often with IV) to estimate the causal effect of x_j.

7.3 Probit model

Consider again the utility :

$$U_{ij} = V_{ij} + \epsilon_{ij}, \quad \forall j$$

The probit model assumes that the errors $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ are normally distributed.

Probit allows for more flexible substitution than IIA, by allowing for correlation between the errors ϵ_{ij}. P_{ij} has no closed form and needs to be numerically computed but computational burden becomes too heavy as the choice set gets larger.

8 To go further

You can find more details explanation of all the notion discussed in this document in the book ”Introduction to Econometrics with R” by Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer. It is a very good book to understand the econometric concepts and the intuition behind them.