EFR summary

Applied Econometrics (Masters)

2023-2024

Lectures weeks 1 to 6

Deloitte.

DeNederlandscheBank
EUROSYSTEEM

Details

Subject: applied econometrics 2023-2024

Teacher: Sacha Kapoor

Date of publication: 12.10.2024

Applied Econometrics – masters course – lecture week 1

Econometric Models

Econometric models have two main ingredients.

They specify a relationship of interests, for example the relationship between Y as a function of X: $Y = F(x)$.

They model the uncertainty of a relationship, whose value if not known by econometricians : $Y = F(x) + e$.

Random variable

This refers to a function X(.) that associates a unique number with every possible outcome of some sort of trial. Random refers to the period before the event takes place. These variables take two forms.

Discrete – one can count values of all the potential outcomes, e.g., only integer values.

Continuous – one cannot count these values, e.g., any number.

Example: flipping a coin. X would be a function that maps heads into the number 1 and tails into 0. X(heads) = 1 and X(tails) = 0.

Random sampling: randomness in the process of information retrieval.
Randomised controlled trials: randomness in the assignment of a "treatment" to a person in the sample.

Probability distribution

The function F(x) marks the probability of our random variable X taking on each of its possible outcomes. In the example below the probability that it is smaller or equal than the value x:

$$F(x) = P(X \leq x)$$

Random sampling

The selection of individual information retrieved of population N is done in a random manner. This means that the information obtained from one individual is independent of the other information we obtain from another, e.g. because Janna was selected tells us nothing about whether her friend Nikita will be selected.

If the true sample is represented by $I_1 * X_1$, $I_2 * X_2$, $I_3 * X_3$, ..., $I_n * X_n$, our sample is a collection of those individuals, where $I_i=1$ if person i is selected and X_i is the information obtained from person i.

In a simple random sample, the probability of each individual being chosen is given by $P(I_i=1) = 1/N$.

From sampling to the population

The reason we use a sample is to gain insight of the overall distribution of a population without gaining information on each individual. From the sample we can calculate the **sample mean** $\overline{X}$ and the **sample variance Var ($\overline{X}$).** This in itself doesn't tell us much about the population; to aid us we have to make assumptions.

$$\overline{X} = \sum_{i=1}^{N} \left(\frac{1}{N}\right) x_i$$

$\text{Var}(\overline{X}) = \frac{Var(X_i)}{N}$; where $\text{Var}(X_i)$ is the population variance (which is sometimes denoted as σ_x^2).

Suppose the **expectation** or **expected value** of the population can be calculated as follows:

$$E(X) = \sum_{i=1}^{m} P(X = x_i) x_i$$

where $P(X=x_i)$ is the ***probability distribution*** of our random variable X. This is a function to calculate the probability of each outcome of x.

With random sampling and a large sample $\overline{X}$ meets established criteria for a "good" estimator $E[X]$. If we collect an infinite number of large samples, measure $\overline{X}$ for each sample, then the fraction of samples with $\overline{X} \approx E[X]$ will tend to 1. It is in this sense that we call $\overline{X}$ a consistent estimator for $E[X]$. Mathematically this relationship is given by:

$$(\overline{X}) = E[X]$$

Recall that $\text{Var}(\overline{X}) = \frac{Var(X_i)}{N}$. Therefore, a smaller sample variance is associated with an increasing sample size or a lower population variance. Generally, it is good if an estimator has a small sampling variance.

There are some properties properties associated with with expectations, which can be used:
$E[X+Y] = E[X] + E[Y]$
$E[aX] = aE[X]$
$E[XY] = E[X]E[Y]$; only when the variables X and Y are independent

From variance to standard error

In practice, we rely on the **standard error** to represent the variability of our data. The standard error of the sample mean is given by:

$$SE(\overline{X}) = \frac{\sigma_x}{\sqrt{N}}$$

As σ_x is the standard deviation of the population, we do not know its value.
Instead, we estimate it using a sample estimate with:

$$S(X_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

Our estimated standard error is then given by:

$$\hat{SE}(\overline{X}) = \frac{S(X_i)}{\sqrt{N}}$$

From our sample we want to test hypotheses about underlying population parameters. Currently this means using $\overline{X}$ to learn about $E(X_i)$. This is done by performing a t-test on the hypothesised value μ. We reject the null hypothesis if it is significantly unlikely to be true, given our sample. The Central Limit Theorem enables this to be done on hand of a standard normal distribution.

$$t(\mu) = \frac{(\overline{X} - \mu)}{\hat{SE}(\overline{X})}$$

Data structures

Cross-sectional: many agents (individuals, firms, households) for whom 1 observation is available (at one moment in time).
Time Series: many observations of only one agent at various points in time.

Panel Data: many agents are observed over various points in time. A collection of cross-sectional observations over time. (Time series and Cross-sectional data are special cases of panel data, existing only in one of the two dimensions).
Clustered Data: many grouped agents where outcomes are correlated within groups. Interdependence is not modelled.
Spatial Data: observations in close physical proximity have correlated outcomes. Unlike clustered data, this interdependence is modelled explicitly.

Joint distributions

So far, we have had one data point per individual from a population characterised by a probability distribution: $F(x) = P(X \leq x)$.

Instead of one data point per individual, often we have 2 or more data points for each individual. Suppose the observations for N individuals are then a collection of pairs (Y_1, X_1), (Y_2, X_2), (Y_3, X_3), … , (Y_N, X_N); where the probability distribution is given by $F(x, y) = P(X \leq x, Y \leq y)$.

Any data pairs taken from this probability function would be characterised by the means $E[X]$ and $E[Y]$, as well as their variances $Var(X)$ and $Var(Y)$. The covariance of two variables is a basic measure to describe their relationship. It tells us whether X and Y deviate from their respective means $E[X]$ and $E[Y]$ together or not.

$$Cov(Y, X) = E[(X - E[X])(Y - E[Y])]$$

To compare relationships between pairs of variables, we can normalise the covariance to get the **correlation**.

$$Corr(Y, X) = \frac{Cov(Y,X)}{\sqrt{Var(X)Var(Y)}} \quad ; \quad \text{where } -1 \leq Corr(Y, X) \leq 1$$

Conditional expectation

This asks what the expected value of Y is conditional on a given value of X. e.g. if Y is discrete with possible outcomes of $y_1, y_2, ..., y_k$ then the expectation is the sum of the probabilities of each outcome occurring multiplied with the value of the outcome.

$$E[Y|X] = y1P(Y = y1|X) + y2P(Y = y2|X) + \ldots + ykP(Y = yk|X)$$

Law of iterated expectation (lie)

$$E[Y] = E[E[Y|X]]$$

This law enables the calculation of the E[Y] using the following formula:

$$E[Y] = E[Y|X = x1]P(X = x1) + E[Y|X = x2]P(X = x2) + \dots + E[Y|X = xm]P(X = xm)$$

Independence

A random variable Y is independent from another variable X if knowing the value of X does not affect your expectation of what Y will be. Mathematically this means the best guess at the expected value of Y given X is simply the expected value of Y.

$$E[Y|X] = E[Y]$$

If **all three** hold, then we can prove that the **Cov(Y, X) =0.**

$$\begin{aligned}
Cov(Y, X) = & \quad E[XY] - E[X]E[Y] & \\
= & \quad E[E[XY|X]] - E[X]E[E[Y|X]] & \text{(by LIE)} \\
= & \quad E[E[X|X]E[Y|X]] - E[X]E[E[Y|X]] & \text{(by independence)} \\
= & \quad E[XE[Y|X]] - E[X]E[E[Y|X]] & \text{(because } E[X|X] = X\text{)} \\
= & \quad E[XE[Y]] - E[X]E[E[Y]] & \text{(by independence)} \\
= & \quad E[X]E[Y] - E[X]E[Y] & \text{(because } E[Y] \text{ is constant)} \\
= & \quad 0 &
\end{aligned}$$

From sample to population

E[Y], E[X], E[Y^2] (or Var(Y)), E[X^2] (or Var(X)), E[XY] (or Cov(X, Y)), E[Y|X] are all **parameters** that describe the population of interest. These parameters are fixed and are invariant to changes in the context or environment being considered. Using statistics, we estimate these true values from the data and determine the informativeness of these estimates.

Applied Econometrics – masters course – Lecture week 2

Regression

A regression is a simple manner to represent a relationship of variables. This is typically specified to be a linear model, meaning it is linear in parameters. These have the characteristic that their partial derivatives are independent of any β_i.

$$Yi = \beta 0 + \beta 1\, Xi1 + \beta 2\, Xi2 + \ldots + \beta k\, Xik + ei$$

where:
$\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters we want to know
$Y, X_{i1}, \ldots, X_{ik}$ are **observable random variables**
e_i is an unobservable random variable ("disturbance")
we define: $E[ei \mid Xi1, \ldots Xik] = 0$ & $E[ei] = 0$

To allow marginal effects to vary by other factors, the inclusion of **interaction effects** is possible. This means that the total effect of X_{i1} on Y is not constant, but changes with at least one other variable X_{ik}.

e.g.: If the function for Y(X) is specified as
$Yi = \beta 0 + \beta 1\, Xi1 + \beta 2\, Xi2 + \beta 3\, Xi1Xi2 + ei$

Taking the derivative with respect to X_{i1} yields a marginal increase in Y that is dependent on the value of X_{i2}.

$$\frac{dY_i}{dX_{i1}} = \beta_1 + \beta_3 X_{i2}$$

Matrix notation used in regression

Data on individual i can be written in the vector format, where K+1 is the number of data points collected for each individual.

$$X_i = X_{i0}\, X_{i1}\, X_{i2} \ldots X_{ik}$$

For each individual x_i we add a row to the matrix, so the complete matrix for all individuals' observations of X_k looks as follows:

$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \cdots & X_{NK} \end{bmatrix}$; with dimensions N x K+1 (K+1 columns and N rows)

Note: the first row of 1s allows for the inclusion of the constant β_0 *into our model, which is not dependent on any* X_i.

The observations y of each individual i can similarly be represented using a row matrix.

$$Y = [Y_1 \; Y_2 \; \ldots \; Y_N]$$

Similarly, the coefficients of β_k can be represented in a column-matrix, which is the same for all individuals in the population. This is depicted as:

$$\boldsymbol{\beta} = \begin{matrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{matrix}$$

Their product $\mathbf{X_i}\,\boldsymbol{\beta}$ is merely a compact way of writing a simple regression equation:

$$\mathbf{X_i}\,\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

Identification

This step in the identification-estimation-inference paradigm derives how we can recover unknown parameters. This identification can be done in three ways.

Identification with moment conditions

In our econometric specification we set up two moment conditions:
$E[e_i | \mathbf{X_i}] = 0$
$E[e_i] = 0$

We had previously defined $e_i = Y_i - \mathbf{X_i}\,\boldsymbol{\beta}$. Combining this definition and the assumption that parameters E[Y], E[X], Var(Y), Var(X), E[XY] are known or can be constructed, we can uncover the unknown parameter E[Y|X].

This leads us to the moment conditions:

$$\begin{aligned} E[e_i] &= 0 \\ E[e_i X_{i1}] &= 0 \\ E[e_i X_{i2}] &= 0 \\ E[e_i X_{i3}] &= 0 \end{aligned}$$

$$\vdots$$
$$E\left[e_i X_{iK}\right] = 0$$

Plugging in $e_i = Y_i - \mathbf{X_i}\boldsymbol{\beta}$ yields:

$$E\left[Y_i - X_i\beta\right] = 0$$
$$E\left[(Y_i - X_i\beta)X_{i1}\right] = 0$$
$$E\left[(Y_i - X_i\beta)X_{i2}\right] = 0$$
$$E\left[(Y_i - X_i\beta)X_{i3}\right] = 0$$
$$\vdots$$
$$E\left[(Y_i - X_i\beta)X_{iK}\right] = 0$$

This is a system of K+1 equations and K+1 unknowns. Under certain conditions this system has only one solution.

In matrix form this set of equations can be expressed as:

$E\left[X_i^T(Y_i - X_i\beta)\right] = 0$; where X_i^T is the transpose of X_i

In the general case,

$$E\left[X_i^T(Y_i - X_i\beta)\right] = 0 \qquad \Leftrightarrow \qquad E\left[X_i^T Y_i\right] = E\left[X_i^T X_i\right]\beta$$

Multiplying both sides with the inverse matrix of $X_i^T X_i$ gives the solution

$$\beta = E\left[X_i^T X_i\right]^{-1} E\left[X_i^T Y_i\right]$$

This yields $\beta_k = \frac{Cov\left(Y_i, \widetilde{X_{ik}}\right)}{V\left(\widetilde{X_{ik}}\right)}$; where $\widetilde{X}$ is the residual from the regression of X_k on all other X's.

We can relax

$$E\left[e_i \mid \mathbf{X_{i1}}, \mathbf{X_{i2}}, \ldots, \mathbf{X_{ik}}\right] = 0 \text{ to}$$
$$Cov(e_i, X_{ik}) = 0$$

Under-identification occurs if there are more unknowns than equations. This means the unknowns in the system of equations cannot be identified.

Identification with 'line of best fit'

Here we aim to minimise the error variance, so that the line best represents the data. When we minimize $E\left[(Y_i - X_i\beta)^2\right]$ we get the same equation for β_k as with the moment conditions:

$$\beta_k = \frac{Cov\left(Y_i, \tilde{X_{ik}}\right)}{V\left(\tilde{X_{ik}}\right)}$$

Once we have identified the parameters β_1 through β_k we can use this to find β_0:

$$\beta_0 = \overline{Y} - \beta_1 \overline{X_1}$$

Identification with 'maximum likelihood '

This method maximises the probability of a observing the data, under a specified statistical model. This in turn comes down to the same formula as the other two methods. It is clear to see that maximising the likelihood is equivalent to minimising the variance for the error.

$$\beta_k = \frac{Cov\left(Y_i, \tilde{X_{ik}}\right)}{V\left(\tilde{X_{ik}}\right)}$$

Estimation

After identifying β_k we can estimate this from our sample as $\hat{\beta}_k$.

If $\beta = E\left[X_i^T X_i\right]^{-1} E\left[X_i^T Y_i\right]$ in the population, we can estimate $\hat{\beta}_k$ from the data with:

$\hat{\beta}_k = \left(\sum_{i=1}^{N} \frac{X_i^T X_i}{N}\right)^{-1} \left(\sum_{i=1}^{N} \frac{X_i^T Y_i}{N}\right)$ *where* $\hat{\beta}_k$ *is the true* β_k if it holds that $E\left[X_i^T e_i\right] = 0$

Goodness of Fit estimation

We define $\hat{R^2} = \frac{\hat{V}\left(X_i\hat{\beta}\right)}{\hat{V}\left(Y_i\right)}$ as a measure of the goodness of fit. This quantifies the proportion of the variation in Y that is explained by the estimation $X_i\hat{\beta}$.

a high R^2 does not imply a causal interpretadion

a low R^2 does not imply no causal interpretation

a low R^2 does not preclude precise estimation of marginal effects

Functional forms involving logarithms

Logarithms are used in regressions to express variables as a percentage. It is handy to know how to interpret the beta's with the following log regressions:

Level-log | y = log(x) | $\Delta y = (\beta_1/100)\%\Delta x$

Log-level | log(y) = x | $\%\Delta y = (100 * \beta_1)\Delta x$

Log-Log | log(y) = log(x) | $\%\Delta y = \beta_1 \%\Delta x$

Inference

With a reasonable estimator of $E[Yi|X_i]$ we require more assumptions to test hypotheses about this estimator. From the data we can infer properties of the population from the descriptive statistics of the sample. To gain information of the distribution we require the **variance-covariance matrix.** Deriving this generates the following equation.

$$\hat{Var}(\hat{\beta}) = \left(\sum_{i=1}^{N} \frac{X_i^T X_i}{N}\right)^{-1} \left(\sum_{i=1}^{N} \frac{\hat{e}_i^2 X_i^T X_i}{N^2}\right) \left(\sum_{i=1}^{N} \frac{X_i^T X_i}{N}\right)^{-1}$$

With this variance we can begin hypothesis testing as to how likely a β is given our sample distribution. If the variance is independent of $\mathbf{X_i}$, then the errors are **homoscedastic**. If they are dependent on $\mathbf{X_i}$, then this is called **heteroskedasticity**. In *Stata* this can be controlled by using a 'robust' regression command. Clustering also relaxes assumptions on independence across observations.

If we want to test if the Beta is significantly different from 0, we set up the following hypotheses and t-test:

$$H_0: \beta_k = 0$$
$$H_1: \beta_k \neq 0$$

$$t = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{Var}(\hat{\beta}_k)}}$$

The null-hypothesis is consequently rejected if $\hat{\beta}_k$ significantly deviates from zero, in other words: if it is significantly larger or smaller than zero.

Prediction and control variables

Using regression models as the basis, predictions can be made of the dependent variable. If we have a good estimation of β then we can learn about Y_i before observing it. The **Lasso** method is a method of selecting and fitting variables with the aim of prediction and model selection.

'Least Absolute Shrinkage and Selection Operator' (LASSO)

The methodology is based on the Bias-Variance trade off. For prediction it may be worth trading off some bias for even lower variance, to get a more accurate one.
From previous derivations we know that calculating the "right" β is done by minimising the error variance.

$$E\left[(Y_i - X_i\beta)^2\right] + \lambda \sum_{k=0}^{K} |\beta_k|$$

In the Lasso methodology we minimise this while allowing violating the constraint at a price λ. This is the 'acceptable' trade-off selected when setting up the equation. For values between 0 and λ_{max} we trade off increased bias for less variance in our estimation of β. In practice cross validation selects this for you in Stata.

This procedure is performed on the **Training sample (T)** and is then tested on the **Validation sample (V).** The **Mean Squared Error** (MSE) is used to calculate the prediction error of the Lasso model.

$MSE^V = E\left[\left(Y_i^V - X_i^V \hat{\beta}^T\right)^2\right]$; where $\hat{\beta}^T$ is the betas estimated from the training sample.

Applied Econometrics – masters course – Lecture week 3

There are a number of biases that **compromise identification** of regression models. These are:

1. Data Missing Completely at Random (MCAR)
2. Sample Selection Bias
3. Selection Bias
4. Bad Controls
5. Measurement Error
6. Simultaneity and Reverse Causality
7. Omitted Variable Bias

Data missing completely at random (MCAR)

In a dataset it is possible that there will be missing observations. These are represented by a "." in Stata. Stata will drop these observations in the estimation.

This is not necessarily a bad thing if you are satisfied that the data is missing completely at random (MCAR). The only issue with removing the data is that the sample size is now smaller.

However, removing observations means that there is less statistical power meaning you are less likely to find statistically significant effects. If you are concerned about sample size, the following process can be used:
If you have a random sample $\{Y_i, Xi_1, Xi_2\}$ of N observations and X_{i2} is missing variables. You define $Z_{i2}=(1-M_{i2})\ Xi_2$ where:

$$M_{i2} = \{1,\ if\ X_{i2}\ is\ missing;\ 0,\ if\ X_{i2}\ isn't\ missing$$

This gives you the following observations for Z_{i2}

$$Z_{i2} = \{X_{i2,}\ \ if\ X_{i2}\ is\ observed;\ 0,\ otherwise$$

Following this you can estimate:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_{i2} + \beta_3 M_{i2} + e_i$$

To interpret this, if M_{i2} is 0 we calculate $\beta_0^0,\ \beta_1^0, \beta_2^0$

If M_{i2} is 1 we can calculate β_0^1, β_1^1. B_2 cannot be calculated as when M_{i1} is 1, Z_{i2} is 0 as $Z_{i2}=(1-M_{i1})\ X_{i2}$

The coefficients β_0^0 and β_0^1 are combined (a weighted average is taken) as well as β_1^0 and β_1^1. By doing this it allows us to get the most accurate and statistically significant results for β_0 and β_1 despite the missing observations for X_{i2}.

Sample selection bias

Sample Selection Bias is when a bias exists because of an error in selecting your sample. Two possible ways this can happen are:

1. **Sample Design**; when data is collected non-randomly.
2. **Respondent Behaviour** (Self-Selection); when participants in the survey either do not answer all questions or drop out after answering a few.

There are a few ways that sample selection bias can be avoided.

1. Selecting a sample based on an exogenous variable (a variable determined outside the model)
2. Do not select a sample based on the dependent variable or an endogenous control variable
3. You can use a regression to control for factors driving sample selection

Proof that you can use an exogenous variable

Imagine you want to estimate $E[Y_i|X_i] = X_i\beta$. This is identifiable if $E[e_i|X_i] = 0$
We know that this assumption holds when a sample is selected at random. If you have a non-random let

$$S_i = \{1,\ if\ i\ is\ selected\ 0,\ if\ i\ is\ not\ selected$$

Now we must determine if $E\ [e_i|X_i,\ Si = 1] = 0$
There are two cases where this is obvious:

1. S_i is purely determined by X_i, therefore making S_i redundant
2. S_i is independent of Y_i, X_i and e_i

When this is the case, we can estimate β using. (It is important to understand the following working out, but you do not need to derive it yourself)

$$\hat{\beta}=\left(\sum_{i=1}^{N}\frac{S_i X_i^T X_i}{N}\right)^{-1}\left(\sum_{i=1}^{N}\frac{S_i X_i^T Y_i}{N}\right)$$

this equals

$$\left(\sum_{i=1}^{N}\frac{S_i X_i^T X_i}{N}\right)^{-1}\left(\sum_{i=1}^{N}\frac{S_i X_i^T (X_i\beta+e_i)}{N}\right)$$

$$\left(\sum_{i=1}^{N}\frac{S_i X_i^T X_i}{N}\right)^{-1}\left(\sum_{i=1}^{N}\frac{S_i X_i^T (X_i\beta+S_i X_i^T e_i)}{N}\right)$$

$$\beta + \left(\sum_{i=1}^{N}\frac{S_i X_i^T X_i}{N}\right)^{-1}\left(\sum_{i=1}^{N}\frac{S_i X_i^T e_i}{N}\right)$$

In a large sample $\left(\sum_{i=1}^{N}\frac{S_i X_i^T e_i}{N}\right)$ goes to 0 as it is a consistent estimator in large settings.
This suggests that it is ok to select a sample based on exogenous variables because:
$E[ei|Xi\,, Si\,] \;=\; E[ei|Xi\,] \;=\; E[ei] \;=\; 0$
If the above is not the case, you cannot ignore sample selection.

Example: If you want to find $Y_i = X_i\beta + \beta_{k+1}X_{ik+1} + e_i$. Where Y_i is someone's weekly earnings and we X_{ik+1} is the self-reported ranking of i's happiness out of 10, if we want to select people who are of above average happiness, we make our selection rule.
$Si = \{1,\ if\ X_{ik+1}\ is \geq 7.5\ 0,\ if\ X_{ik+1}\ is\ <\ 7.5$
However respondent behaviour may have an effect on who is selected in this sample. In this case we need to control for people's ego when they self-report their happiness. Not doing so will lead to sample selection bias as people may overstate their happiness because they want to seem better off than they actually are. Controlling for EGO we get the following estimate: $Y_i=X_i\beta + \beta_{k+1}X_{ik+1} + \beta_{k+2}EGO_i + e_i$
We then rely on the assumption that:
$E[e_i|X_i,X_{ik+1},EGO_i,s_i] = E[e_i | X_i,X_{ik+1},EGO_i]=0$
If the assumption holds, then we get a good estimate from this sample.

Selection bias

Selection Bias is when you have biased estimates of β because of underlying differences between the units in the sample. This is different to sample selection bias where you have biased estimates of β because of differences between your sample

and the actual population. Selection bias can occur when some subjects select themselves into different groups. This will generate biased estimates of β as they will be comparing people who should be in the same group leading to inaccuracies. As before with sample selection bias you can control for factors that will lead to selection bias.

Example: A study on the long-run effects of attending the gym on a person's weight. In this case we have two groups.

$X_{i1} = \{T,\ \ person\ i\ goes\ to\ the\ gym\ C,\ person\ i\ does\ not\ go\ to\ the\ gym$

For everyone they either go to the gym or not. These two outcomes are expressed by $Y_i(T)$ and $Y_i(C)$. To understand the effect of going to the gym we will calculate $Y_i(T)—Y_i(C)$.

*Unfortunately, only **one of these** can be observed at a time. We cannot observe the missing observations for a potential outcome (we do not know the weight of a person who attends the gym had they not attended, and vice versa). This missing outcome is known as a counterfactual.*

These comparisons may therefore be inaccurate as they could be down to other factors that lead people to not attend the gym. This is called selection bias.
In order to get past this selection bias, we must control for it. Firstly, we will model the error for i as a derivation from the average weight of the population if they do not attend the gym

$$e_i = Y_i(C)—E[Y_i(C)]$$

We set the weight of someone who attended the gym to

$$Y_i(T) = Y_i(C) + \beta_1$$

And let

$$B_0 = E[Y_i(C)]$$

Combining the equations gives

$$Y_i(C)=\beta_0 + e_i$$
$$Y_i(T) = \beta_0 + \beta_1 + e_i$$

Then replace T with 1 and C with 0 for X_{i1}

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i$$

Thus, when doing the regression, you get

$$E[Y_i\ |X_{i1} = T] = \beta_0 + \beta_1 + E[e_i\ |X_{i1} = T]$$
$$E[Y_i\ |X_{i1} = C] = \beta_0 + E[e_i\ |X_{i1} = C]$$

Subtracting the C equation from the T equation gives

$$E[Y_i|X_{i1} = T] - E[Y_i|X_{i1} = C] = \beta_1 + E[Y_i(C)|X_{i1} = T] - E[Y_i(C)|X_{i1} = C]$$

The $E[Y_i(C)|X_{i1} = T] - E[Y_i(C)|X_{i1} = C]$ part of this calculation represents the selection bias. We cannot identify β_1 because of the selection bias meaning that the treated group differs from the control group and therefore we cannot get a good estimate for β_1 as:
$E[Y_i(C)|X_{i1} = T] \neq E[Y_i(C)|X_{i1} = C]$

Case study: Dale Krueger solution

Dale and Kreuger (2002) found a solution to solve the problem of selection bias. They wanted to find the effect of going to an elite college on people's earnings. They ran into the problem of selection bias as there were too many non-observable differences between the two groups.

They gathered individual information on people's SAT scores, what type of university they attended and what they applied to as well as their earnings in 1996. From the universities they had information on the type of university they were as well as tuition fees and average SAT scores. Dale and Kreuger realised that 'the colleges that people applied to' could identify hard-to-observe characteristics such as where their parents went to and their ambition. While 'the colleges they were admitted to' showed other unobservable characteristics about the students.

They then compared the earnings of people in public and private universities that were accepted into the same colleges. They found that the average causal effect of going to an elite university on wage was not statistically significant.
The regression equation was given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \sum_{j=2}^{151} \beta_j X_{ij} + X_i\beta + e_i$$

where X_{ij} indicates whether individual i belonged to group j while X_i includes the observable characteristics of i such as SAT scores etc. The estimates of β are interpreted as causal if conditional independence holds i.e. if e_i is independent of X_{i1} once the different groups are controlled for.

Bad controls

If the right controls are included, the estimated coefficients can be interpreted as causal. However, it is not good to include all possible variables as controls. A 'kitchen sink' regression refers to one in which all possible controls are thrown into; some

controls are simply bad controls. These can **threaten identification**. Secondly, adding controls takes out variance of our key explanatory variables; meaning the estimation becomes trickier.

If the control variable is determined **before** the variable of interest (treatment), there is no problem with identification. In contrast, if the control is determined **after** the variable of interest, then controlling for it distorts the full picture. These types are called mechanisms and should not be included, else they hinder efforts to uncover the coefficient of interest.

Proxy controls are potential bad controls that researchers intentionally include in an attempt to control for an important missing variable.

Measurement error (ME)

Measurement errors exist in two forms. It can be in the **explanatory** variables or in the **dependent** variables. This occurs when there is a clearly defined quantitative measure of the target variable, yet this is inaccurately measured.

ME in the dependent variable

If the measurement error is given by:

$$u_i = Y_i - Y_i^*$$

and the 'true' model is given by:

$$Y_i^* = X_i\beta + e_i$$

Then the model run on the measured values is: $Y_i = X_{i\beta} + e_i + u_i$

In this model, β can only be identified if $Cov(X_{ik}, u_i)=0$. In both cases, inference is affected by the measurement error. The variance and consequently the $\hat{SE}(\hat{\beta})$ becomes larger which decreases the likelihood of a statistically significant result.

ME in the explanatory variable

If the independent/explanatory variable has a measurement error ánd the error covaries with the reported values, then the estimates of the regression will lead to a

lower bound **conservative** estimate of the variable of interest. This means that the 'true' β will be larger than the observed β.

Simultaneity bias

Simultaneity bias refers to a situation in which X determines Y, but Y is also at least partly determined by X. This is related to **reverse causality.**
Mathematically this can be shown by the following relationships.

$$Y_i = \alpha X_{i1} + \beta X_{i2} + e_i$$

Yet,

$$X_{i1} = \gamma Y_i + Z_i + e_i$$

As these are simultaneously determined, and $\gamma \neq 0$, simply running the first regression will not return the true causal effect of X1 on Y, but a biased estimator will be determined instead.

Omitted variable bias (ovb)

The problems mentioned so far boil down to a false estimate being made due to the exclusion of variables, leading to a bias. In the basic form, the estimate of the effect of X_1 on Y then includes the effect that is in fact due to X_2 but is not controlled for.

Basic form: $Y_{i1} = \beta_0 + \beta_1 X_{i1} + e_i$

Full form: $Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

The estimate of β can be simplified to:

$$\hat{\beta} = \beta_1 + \beta_2 \frac{Cov(X_{i2}, X_{i1})}{Var(X_{i1})}$$

OVB

From this, one can easily see how the omitted variable X_2 has the potential to bias the estimator of X_1. The following table shows the effect of the ovb on β_1:

	$\beta_2 > 0$	$\beta_2 < 0$
$Cov(X_{i1}, X_{i2}) > 0$	overestimate β1	underestimate β1
$Cov(X_{i1}, X_{i2}) < 0$	underestimate β1	overestimate β1

Applied Econometrics – masters course – Lecture week 4

Causal inference

This week the methods to deliver a causal interpretation are discussed. These are:

1. Experiments
2. Instrumental Variables (IV)
3. Regression in Discontinuity Design
4. Differences in Differences

Fundamental problem of causal interference

If X_{i1} can have two values, T and C, the target parameter can be described as:
$\beta_{i1}=Y_i(T)—Y_i(C)$

However, it is only possible to observe one of these outcomes at any given time. This is called the **fundamental problem of causal inference**.

To show that it is not possible, the target can be redefined as
$E[Y_i(T)—Y_i(C)]$
This is the average treatment effect (ATE) for our population. This can be rewritten as
$E[Y_i|X_{i1}=T]— E[Y_i|X_{i1}=C]$
Which equals

$$\underbrace{\mathbb{E}[Y_i(T)|X_{i1}=T]-\mathbb{E}[Y_i(C)|X_{i1}=T]}_{\mathbb{E}[Y_i(T)-Y_i(C)] \text{ by Independence}}+\underbrace{\mathbb{E}[Y_i(C)|X_{i1}=T]-\mathbb{E}[Y_i(C)|X_{i1}=C]}_{\mathbb{E}[Y_i(C)]-\mathbb{E}[Y_i(C)] \text{ by Independence}}$$

And then simplifies to $E[Y_i(T)—Y_i(C)]$

Experiments

Selection bias can be eliminated through randomisation. We will run a randomised control trial (RCT). This assigns people into the treatment group. This means that the treatment and control groups are the same on average except for the treatment assignment.

RCT is very expensive and difficult to do in a credible way.

Estimation and inference

If the RCT has one treatment and one outcome (Y) you can either do the regression.

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i$$

Or calculate

$$\overline{Y}^{X_{i1}=1} - \overline{Y}^{X_{i1}=0}$$

It is better to use regression analysis as that simultaneously generated the t-statistics.

If the RCT has multiple treatments and/or multiple outcomes (many Ys), the researcher should adjust for this. If you were to run an experiment with 4 treatments and 5 outcomes. We would then get this set of equations for estimation.

$$\begin{aligned}
Y_i^1 &= \beta_0^1 + \beta_1^1 X_{i1} + \beta_2^1 X_{i2} + \beta_3^1 X_{i3} + \beta_4^1 X_{i4} + e_i^1 \\
Y_i^2 &= \beta_0^2 + \beta_1^2 X_{i1} + \beta_2^2 X_{i2} + \beta_3^2 X_{i3} + \beta_4^2 X_{i4} + e_i^2 \\
Y_i^3 &= \beta_0^3 + \beta_1^3 X_{i1} + \beta_2^3 X_{i2} + \beta_3^3 X_{i3} + \beta_4^3 X_{i4} + e_i^3 \\
Y_i^4 &= \beta_0^4 + \beta_1^4 X_{i1} + \beta_2^4 X_{i2} + \beta_3^4 X_{i3} + \beta_4^4 X_{i4} + e_i^4 \\
Y_i^5 &= \beta_0^5 + \beta_1^5 X_{i1} + \beta_2^5 X_{i2} + \beta_3^5 X_{i3} + \beta_4^5 X_{i4} + e_i^5
\end{aligned}$$

Omnibus treatment

Omnibus treatments are experiments with a lot of broad based interventions that have a number of things being manipulated at one time. The advantage of these is that they help people discover if something works, unfortunately it does not show why something works and is difficult to link back to theory.

Experiments – summary

Experiments are expensive. They can be risky as something may go wrong in the question or design. Another issue is that it is hard to determine whether the outcome will work in all cases (can be generalised) or only in the conditions and time of the experiment.

Instrumental variables (IV)

Identification

A different way of finding causal effects is to use instrumental variables. These are "outside forces" that effect only X_i and not Y_i.
Suppose our model of interest is

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i$$

This is called the second stage regression.
As $Cov(X_{i1}, e_i) \neq 0$, we can't find β_1 using $\frac{Cov(Y_i, X_{i1})}{Var(X_{i1})}$

IV aims to find a variable Z_i that changes X_i but is not correlated with e_i. This can be implemented into the following 'first stage' regression:

$$X_{i1} = \pi_0 + \pi_1 Z_{i1} + u_i; \text{ where } \pi_1 \neq 0, \; and \; Cov(Z_i, e_i) = 0$$

The estimation of coefficients then follows previous methodology yields the **IV formula**:

$$\frac{Cov(Y_i, Z_{i1})/Var(Z_{i1})}{Cov(X_{i1}, Z_{i1})/Var(Z_{i1})} = \beta_1$$

There are a number of assumptions that allow us to get β_1 via the reduced form and first stage.

1. $\pi_i \neq 0$. This is what makes it a meaningful first stage.
2. $Cov(Z_{i1}, e_i) = 0$ says Z_{i1} is exogenous relative to unobservables affecting Y_1.
   Independence. IV is randomly assigned and unrelated to omitted variable bias in the reduced.
   Exclusion Restriction. IV affects the outcome only through the effects it has on X_i.

If the independence and exclusion restriction hold then it is said to be a **valid IV**.
The final assumption needed is:

3. Monotonicity. The instrument does not cause people to select into the treatment or for others to select out. There should be no 'defiers'.

Estimation

When you have one instrument Z_{i1} and endogenous variable X_{i1}, the population moments are:

$$E\left[Y_i - \beta_0 - \beta_1 X_{i1}\right] = 0$$

$$E\left[(Y_i - \beta_0 - \beta_1 X_{i1})Z_{i1}\right] = 0$$

This equals

$$\beta_0 + E\left[X_{i1}\right]\beta_1 = E\left[Y_i\right]$$
$$E\left[Z_{i1}\right]\beta_0 + E\left[X_{i1}Z_{i1}\right]\beta_1 = E\left[Y_i Z_{i1}\right]$$

Given a random sample we substitute in the sample averages.

$$\hat{\beta}_0 + \left(\sum_{i=1}^{N} \frac{X_{i1}}{N}\right)\hat{\beta}_1 = \left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)$$
$$\left(\sum_{i=1}^{N} \frac{Z_{i1}}{N}\right)\hat{\beta}_0 + \left(\sum_{i=1}^{N} \frac{Z_{i1}X_{i1}}{N}\right)\hat{\beta}_1 = \left(\sum_{i=1}^{N} \frac{Y_i Z_{i1}}{N}\right)$$

β_1 can be estimated by doing the following:

1. Estimate the first stage
$$X_{i1} = \pi_0 + \pi_1 Z_{i1} + u_i$$
2. Calculate the fitted value
$$\hat{X}_{i1} = \hat{\pi}_0 + \hat{\pi}_1 Z_{i1}$$
3. Replace X_{i1} with $\hat{X}_{i1}$ in the second stage equation
$$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + e_i$$
4. Estimate the last equation using least squares regression

It is better to let STATA get the estimates for you using commands such as ivreg, ivreg2, xtivreg, and xtivreg2, so that the standard errors will be correct.

When more IVs are included than endogenous variables another method must be used to find β. The generalised method of moments (GMM) approach is used, which finds β while minimising the following equation.

$$\left(\sum_{i=1}^{N} Z_i^T (Y_i - X_i \beta)\right) W_N \left(\sum_{i=1}^{N} Z_i^T (Y_i - X_i \beta)\right)$$

Where the following condition is set.

$$W_N = \left(\sum_{i=1}^{N} \frac{Z_i^T Z_i}{N}\right)^{-1}$$

Inference

The standard error for the 2SLS estimator is

$$SE\left(\hat{\beta}_1\right) = \frac{S(e_i)}{\sqrt{N}\, S(\hat{X}_{i1})}$$

The standard error will be larger than the OLS standard error

$$SE\left(\hat{\beta}_1\right) = \frac{S(e_i)}{\sqrt{N}\, S(X_{i1})}$$

because X_{i1}^{*} will vary less than X_{i1} from sample to sample.

Important remarks on IV

Check for a weak first stage. The first stage t statistic should not be lower than 3.3 and the F statistic should not be lower than 10. Do balancing tests. If you see that there is an imbalance check how your estimates change when using imbalanced controls. One should note that IV estimator is consistent but generally not unbiased.

Regression discontinuity design (RDD)

Identification

There are often well-defined rules that define whether someone is in a treatment (T) or control (C) group. One example is whether someone graduates cum-laude or not. The grade would represent the variable which determines what group the individual is in.

$$Assignment_i = \left\{T \;\; X_{i1} \geq x \; C \;\; X_{i1} < x\right\}$$

We can then use this discontinuity to identify the effect of the treatment at the threshold x.

$$E\left[Y_i(T) - Y_i(C) | X_{i1} = x\right]$$

To identify β_1 we need to allow h to be a positive number as small as possible to calculate the treatment effect within a certain threshold just above or below x.

$$E\left[Y_i | x \leq X_{i1} \leq x + h\right] - E\left[Y_i | x > X_{i1} \geq x - h\right]$$

We can show that this will give the treatment effect as h approaches 0.

$$\left(E\left[Y_i | x \leq X_{i1} \leq x + h\right] - E\left[Y_i | x > X_{i1} \geq x - h\right]\right) = E\left[Y_i(T) - Y_i(C) | X_{i1} = x\right]$$

For this identification to work $E\left[Y_i(T) | X_{i1} = x\right]$ and $E\left[Y_i(C) | X_{i1} = x\right]$ must be continuous at x.

Estimation

With a very large sample, it is possible to estimate.

$$E\left[Y_i | x \leq X_{i1} \leq x + h\right] - E\left[Y_i | x > X_{i1} \geq x - h\right]$$

Using

$$\overline{Y}^{+} - \overline{Y}^{-}$$

This means the average outcome just below the threshold is subtracted from the average just above.

Inference

With a very large sample the standard error for the mean difference is.

$$\hat{SE}\left(\overline{Y}^{+} - \overline{Y}^{-}\right) = S(Y_i)\sqrt{\frac{1}{N^{+}} + \frac{1}{N^{-}}}$$

$S(Y_i)$ is the standard deviation for people who are very close to the cut-off point for x.

Estimation and inference

However, rarely are there enough data points at x to get a good estimation or inference. Researchers will therefore increase the bandwidth around x to try and reduce the sampling variance. But this can make our estimator inconsistent/biased.

Using the STATA package rdrobust will allow you to get the optimal bandwidth
So, in order to estimate it for smaller samples let: $D_i = \left\{1 \quad X_{i1} \geq x \; 0 \quad X_{i1} < x\right\}$
The specification for this RDD is

$$Y_i = \beta_0 + \beta_{rd} D_i + \beta_1 X_{i1} + \beta_2 D_i X_{i1} + e_i$$

$\mathbb{E}[Y_i | x \leq X_{i1} \leq x + h] - \mathbb{E}[Y_i | x > X_{i1} \geq x - h]$ is measured by β_{rd} with h as our bandwidth. X_{i1} is for differences between people below x. $D_i X_{i1}$ measures the differences between people above and at x. $Cov(D_i, e_i)=0$

The assumptions of RDD can be tested. The main one is that units that are very close to x are effectively randomised whether they are in the treatment or control groups.

These tests are:

1. **Balancing Test-** it shows that demographics are similar either side of the cut-off
2. **McCrary Test-** it is used to check if the number of units vary smoothly across the threshold.

Fuzzy RDD

A sharp RDD was described above. This is when there is a deterministic and discontinuous jump. This essentially means that it is well defined and known whether someone is above or below the threshold i.e. whether someone is tall enough to ride a rollercoaster.

In some cases the threshold only represents a change in the probability of treatment. One example of this is the legal drinking age. It means that someone older than that age has an increased probability of drinking alcohol however we do not know if they were drinking before that age.
To do a fuzzy RDD you should combine the RDD with an IV.
Define D_i as

$$D_i = \{1\ i\ is\ treated\ 0\ i\ is\ control\}$$

After this estimate

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_{i^1} + e_i$$

Then instrument for D_i using

$$Z_i = \begin{cases} 1 & x \leq X_{i1} \leq x + h \\ 0 & x - h \leq X_{i1} < x \end{cases}$$

You have to make sure that both RDD assumptions and IV assumptions are satisfied.

Differences-in-differences (DD)

Identification

DD uses the fact that some treatments send certain people down a different path and gives them a different outcome then before.

An example of this is if there were two cities with similar crime rates and a new policy came in place in one of them. By looking at their trends and comparing changes

after the new policy has been brought in you can get a causal effect of the policy on crime rates.

In this example we will use two cities (A and B) and two years (1 and 2). City B brings in a new policy to tackle crime in year 2.

The treatment group is defined below

	1	2
A	No New Policy (C)	No New Policy (C)
B	No New Policy (C)	New Policy (T)

The target parameter is $Y_i(d,t)$; where Y is dependent on city or district d and time t.

Let $Y_i(d,t) = \gamma(d) + \lambda(t) + \beta_{T^1}(B,2) + e_i(d,t)$

We are interested in β_T, which is the effect of the intervention. The expectation yields

$$E\left[Y_i(d,t)|d,t\right] = \gamma(d) + \lambda(t) + \beta_{T^1}(B,2) + E\left[e_i(d,t)|d,t\right]$$

Their new policy generates the following four potential outcomes:

	1	2
A	$E\left[Y_i(A,1)|A,1\right]$	$E\left[Y_i(A,2)|A,2\right]$
B	$E\left[Y_i(B,1)|B,1\right]$	$E\left[Y_i(B,2)|B,2\right]$

This overview enables the extraction of the DD formula

$$\left(E\left[Y_i(B,2)|B,2\right] - E\left[Y_i(B,1)|B,1\right]\right) - \left(E\left[Y_i(A,2)|A,2\right] - E\left[Y_i(A,1)|A,1\right]\right)$$

If we want to look at the city with the policy change, we examine

$$E\left[Y_i(B,2)|B,2\right] - E\left[Y_i(B,1)|B,1\right]$$

We can show that this equals

$$\lambda(2) - \lambda(1) + \beta_T + E\left[e_i(B,2)|B,2\right] - E\left[e_i(B,1)|B,1\right]$$

$E\left[e_i(B,2)|B,2\right] - E\left[e_i(B,1)|B,1\right]$ gives us the change in crime rates in city B had the new policy NOT been brought in. We call it $\text{trend}_{B.}$

City A was not treated,

$$E\left[Y_i(A,2)|8,2\right] - E\left[Y_i(A,1)|A,1\right]$$

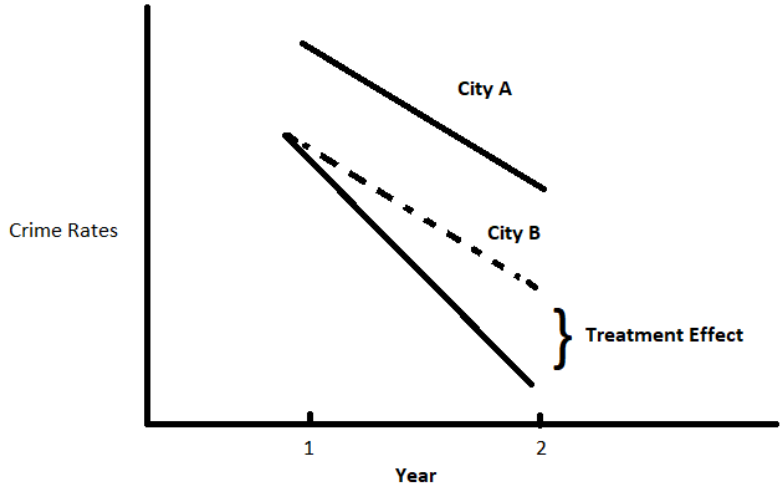
This is

$$\lambda(2) - \lambda(1) + E\left[e_i(A,2)|A,2\right] - E\left[e_i(A,1)|A,301\right]$$

We refer to $E\left[e_i(A,2)|A,2\right] - E\left[e_i(A,1)|A,1\right]$ as trend_A.

Therefore our DD formula is
$\beta_T + \text{trend}_B - \text{trend}_A$

If the common trends assumption holds $\text{trend}_B = \text{trend}_A$ meaning that we are left with β_T in our DD formula (the treatment effect).
The common trends assumption is that the two cities follow the same trend before and in the counterfactual after the intervention. This is vital to a causal interpretation of DD.



The dashed line represents the trend for City B had they not introduced the policy, the difference between the endpoint of the dashed line and the solid line is the treatment effect of the new policy.
If you do not have common trends between the two, one option is to equalise the trend using control variables.

Estimation and inference

We will estimate the following formula:

$$Y_{idt} = \beta_0 + \beta_1 X_d + \beta_2 X_t + \beta_3 (X_d * X_t) + X_{idt}\beta + e_{idt}$$

With c=1 if it is city A and 0 when it's city B, t=1 when t≥2 and 0 when t≤1.
Essentially you should add a control for the year and city, you can find the treatment effect by interacting the city and the year as seen above.

Key assumptions

Experiments: Randomisation is done correctly
Regression: Controls allow for an apples-to-apples comparison
Instrumental Variables (IV): Relevant first stage ($\pi_i \neq 0$), Independence, Exclusion Restriction, Monotonicity
Regression Discontinuity Design: Research subjects cannot control running variable perfectly
Differences in Differences (DD): Common Trend Assumption

Applied Econometrics – masters course – Lecture week 5

This week introduces new ways to tackle the fundamental problem of causal inference. It does so by making use of panel data and repeated observations over time of one individual.

Fixed effects

Regression analysis allows for control variables to be included in the model, which decreases potential biases. However, characteristics of individuals are **often unobservable** (*think of intrinsic motivation or natural ability)*, which makes controlling for them (even through a proxy) tricky at best. Fixed effects allows a researcher to control for ALL time invariant characteristics.

Sticking to the thought of intrinsic motivation or natural ability, we can reasonably assume that an individual has a constant level of these. For simplicity we will call this '*ability*'.
Including *ability* in a regression for an individual's income level generated the following basic regression equation:

$$Y_{it} = \beta_0 + \beta_1 X_{itl} + A_i + e_{it};$$

where Y_{it} is income of i at t, X_{it} is the variable of interest of i at t and A_i is individual i's ability, which is constant over time.

Estimating β_1 is only possible if $Cov(X_{it1}, A_i) = 0$

This is a very strong assumption that often does not hold, in which case the coefficient cannot be identified. In contrast, fixed effects make use of repeated observations of an individual to generate a within group specification. By subtracting the previous period's values or as is the case in FE, the mean, only values that change over time remain. Ability, which is constant over time is removed, leaving a model in which β1 can be identified. In this way, without ever knowing the values of the unobservable characteristics it is 'washed out'. **Demeaning** the basic equation generates the following equation. Estimation thorough this is using the **fixed effects estimator.**

$$\left(Y_{it} - \overline{Y}_i\right) = \beta_1\left(X_{it1} - \overline{X}_{i1}\right) + \left(e_{it} - \overline{e}_i\right)$$

In stata this is performed using the commands

```
xtset person year
xtreg Y X, fe robust
```

where *fe* demeans the data before running an OLS regression on the demeaned data. Robust allows heteroskedasticity and serial correlation to be present in the error terms. Allowing for individual specific time invariant factors to be controlled for reduces a lot of potential biases without ever knowing the true values, however there can still be factors that affect the whole population at any given time t. The complete model most often seen in econometrics allows for this by including a dummy variable γ_t for each year in the specification.

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + X_{it1}\beta \ + A_i + \gamma_t + e_{it}$$
$$\text{where } X_{it1}\beta \ = \beta_2 X_{it2} + \beta_3 X_{it3} + \ \dots \ + \beta_k X_{itK}$$

Random effects

If $Cov(X_{it1}, A_i) = 0$ then β_1 can be identified **without FE**. In this case A_i does not have to be controlled for and can be disregarded from the regression equation and bundled into the error term.

$$Y_{it} = \beta_0 + \beta_1 X_{i1} + v_{it}$$

where $V_{it} = Ai + e_{it}$

This essentially runs a simpler OLS regression on the data. However, as A_i is an individual specific constant, putting this into the error term induces correlation in the error term. For this reason, the **standard errors** need to be **adapted** accordingly.
Once the estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_e^2$ have been made, the variance-covariance matrix can be constructed. Stata deals with this automatically using the command: *xtreg Y X, re*

Correlated random effects

The assumption necessary for RE specification is not always a reasonable one. In the case where it is possible to estimate A, this can be done by including individual specific averages of control variables over time.

If the regression model

$$Y_{it} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{it} + A_i + e_{it}$$

is implemented, it is possible to include $\overline{X}_i = \sum_{t=1}^{T} \frac{X_{it}}{T}$ (the average of X_{it} for individual i) and define Ai as $A_i = \beta_4 \overline{X}_i + r_i$. This relaxes the assumption $\text{Cov}(X_{it}, A_i) = 0$ to only r_i being uncorrelated with the X variables. The identification strategy is therefore open to more cases. As more important variable averages are included, this assumption becomes more reasonable. The averages represent exactly the factors that make up Ai. As more and more control variable averages are included, the sum of these values would tend to the value of Ai. These are exactly the factors that in FE were just 'washed out' but in RE should have been controlled for.
Imagine a world in which all individual specific time-invariant criteria are quantifiable. In this world, as the number of included variables increases, more and more of all possible attributes are being added. A_i is nothing more than the weighted average of all of these factors.

A causal interpretation of Beta still relies on certain assumptions. These have to be assessed case by case for how realistic they are.

Dynamic Panels

Often when regressing with panel data in fixed effects we control using lags of the dependent variable. We usually do this to:

Include covariates that remove bias to help gain a causal interpretation for other explanatory variables
Help the identification of the dynamics of the dependent variable (ie. economic growth)

This only works with long data sets, with inpersistent data, as persistent data leads to bias.

Identification

Let's assume we have a model:

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + \beta_2 Y_{it-1} + A_i + \gamma_t + e_{it}$$

Where A is the fixed effects for i and γ is the fixed effects over time. To identify all the parameters we would assume conditional independence:

$$\mathrm{E}[e_{it}|X_{it1}, Y_{it-1} A_i, \gamma_t] = 0$$
$$Cov(e_{it}, X_{it1}) = 0$$
$$Cov(e_{it}, Y_{it-1}) = 0$$
$$Cov(e_{it}, A_i) = 0$$
$$Cov(e_{it}, \gamma_t) = 0$$

We do run into an issue, as errors over time in some models are covariant. For example, when measuring country GDP over time, it is reasonable to assume that the residuals are covariant in different years. This is known as **serial correlation** in the errors and can be due to culture, institutions or country attitudes.

$$Cov(e_{Netherlands,1974}, e_{Netherlands,2010}) \neq 0$$

This serial correlation can be expressed in an equation, for instance:

$$e_{it} = \gamma e_{it-1} + u_{it}$$

Where u_{it} expresses the new error term. This is known as an autoregressive process of order 1 (as it has 1 lag). Here lagged errors are allowed to clump together, while u_{it} are random shocks. Adding more lags into the model clumps together the errors more.

In our model we have the term Y_{it-1}, which is equal to:

$$Y_{it-1} = \beta_0 + \beta_1 X_{i(t-1)1} + \beta_2 Y_{it-2} + A_i + \gamma_t + e_{it-1}$$

This show that Y_{it-1} is dependent on e_{it-1} which means it is covariant with the error term and that causes a failure in conditional independence.

$$Cov(e_{it}, Y_{it-1}) \neq 0$$

How do we fix this, by adding more lags of the dependent variable to the model, this will remove some serial correlation. We keep adding lags till the serial correlation is statistically insignificant.

To identify the equation, we can use fixed effects if we have many periods, as $\overline{e_{it}} = 0$.

However, in short panels with less than 20 time periods we cannot use this, instead we will use the Arellano-Bond Estimator. This exists out of taking the first difference of our model:

$$(Y_{it} - Y_{it-1}) = \beta_1 (X_{it1} - X_{i(t-1)1}) + \beta_2 (Y_{it-1} - Y_{it-2}) + (e_{it} - e_{it-1})$$

However this does not satisfy our conditions as the $\text{cov}(Y_{it-1} - Y_{it-2}, e_{it} - e_{it-1}) \neq 0$, since Y_{it-1} is dependent on e_{it-1}. Thus we use an instrumental variable to fill in for $(Y_{it-1} - Y_{it-2})$. We can use Y_{it-2} or any late lag as the IV as these do not depend on the error terms. This will give us a causal regression

Estimation

In Stata we can use **xtreg Y X, fe robust** to regress panels with data over a longer period of time, and **xtabond** for short panels.

When estimating regressions, we can have issues with high persistence, which means that the change in Y just a cumulation in random shocks:

$$Y_{it} = \xi_{it} + \xi_{it-1} + \xi_{it-2} + \ldots + Y_{it-100}$$

This would mean that Y changes in no clear directions and results in an estimation that is not causally interpretable. For example if X is also very persistent it could result in a very large β_1 even if there is no causal effect, due to the random walk of both variables.

Applied Econometrics – masters course – Lectures 11&12 - week 6

This week discusses limited dependent variables (LDV) models. These are models where the dependent variable is a binary or limited value. An example of such a value could be the colour of a bike (yellow, blue or grey) or whether a book is available in a library (yes or no). So we model conditional probabilities instead of a conditional mean.

Linear probability model(LPM)

The linear model has linear parameters, so we still estimate a model that could look like the following model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i$$

The interpretation of these models is easy, in the one above for example, an increase of 1 in X is associated with a β_1 %-point increase in the chance that Y=1. This means that for lower values of X, the marginal effect is the same as for high values of X. So the partial effects do not depend on **Xi**. A disadvantage is that if X is then really high (or low), the value of Y could be higher than 1 (or lower than 0), which is a strange outcome, since chances are always between 0 and 1.

Nonlinear probability models

To solve this problem of outcomes higher than 1 and lower than 0, we could use a nonlinear probability model, which looks like the following:

$$P(Yi = 1|X_i) = F(X_i\beta)$$

Where in the model above **Xi** is a vector that represents all X's in the regression. Now the interpretation is less easy, since the effect of β_1 depends on how large X is. We could instead use the marginal effect at the mean of **Xi** or the average effect for all X's to still have an easy but less precise interpretation.

Another problem is the modelling of our ignorance. In the case of classification or prediction of Y, the problem is not extremely important, but if we want to have causal interpretations, we need to specify the error term. We have 3 ways to do this:

1. Probit, where ei is normal with a mean of 0, a variance of 1 and a cumulative distribution function. $F\left(X_i\beta\right) = \Phi(X_i\beta)$
2. Logit, where ei is logistically distributed. $F\left(X_i\beta\right) = \frac{exp\left(X_i\beta\right)}{1+exp\left(X_i\beta\right)}$, where exp is an exponential function
3. LPM, where ei is uniformly distributed over the interval [0,1]. $F\left(X_i\beta\right) = X_i\beta$, where if $X_i\beta \leq 0$, $F\left(X_i\beta\right)$ is 0, if $X_i\beta \geq 1$, $F\left(X_i\beta\right)$ is 1

These models are usually estimated by maximum likelihood. This method changes the parameters in our model in such a way that the estimated parameters have the highest chance of observing the data we have used.

In this model we cannot estimate the errors in the traditional way, therefore we can use either the pseudo R^2 (which looks like the normal R^2), or the likelihood ratio index (LRI). The LRI is calculated in the following way:

$$LRI = 1 - \frac{l\left(\hat{\beta}\right)}{l_0}$$

Where $l\left(\hat{\beta}\right)$ is the log likelihood estimation of $\hat{\beta}$, and l_0 is the log likelihood of 0. When then the $l\left(\hat{\beta}\right)$ is closer to being 0, LRI is closer to zero, while if $l\left(\hat{\beta}\right)$ is further away from l_0, so the coefficients are less likely to be 0, LRI is closer to 1.

Utility maximization

We could use these models to estimate the utility of people, where people make a decision, for example, to buy or not buy a good. We can then exploit this binary choice to build a model to estimate peoples utility.

Sample selection corrections

Sometimes we miss observations that we need for our regression to say something about the whole population. A classic example is the wage of people that do not work. This wage cannot be observed, but without these data we cannot say anything about the whole population, only about the people that work and receive a wage.

We can restore identification by using the inverse mills ratio, and use this as a control variable in our regression. We can do this in the following way. Suppose Si is a dummy that is 1 if I is sampled and 0 if I is not sampled.

First, we estimate a probit regression on Si to estimate the chance that Si=1:

$$P(S_i = 1|X_{i1}, X_{i2}) = \Phi(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2})$$

Then we plug these estimates of γ_0, γ_1 and γ_2 into a probability distribution function (pdf) and into a cumulative distribution function (cdf). The inverse mills ratio (IMR) is then equal to:

$$IMR = \hat{\lambda}_i = \frac{pdf\ estimation}{cdf\ estimation}$$

Then we can use $\hat{\lambda}_i$ as a variable in our main regression to gain consistent estimators of βi. Be careful that it is better to have at least one variable in the estimation of Si that is excluded in the main regression.

Poisson models

Sometimes Y is not in between 0 and 1, but is a nonnegative integer. In this case a linear model could give strange results as well. Therefore we can use the Poisson distribution with parameter λ. Under a Poisson distribution Y is a nonnegative integer, where the chance of taking one of these values is:

$$f(Y) = \frac{exp(-\lambda)\lambda^Y}{Y!}$$

Where Y! is the factorial of Y, and where $\lambda = Var(Y) = E(Y)$. This means that a poisson model is heteroskedastic by assumption. f(Y;λ) is now defined as the likelihood function for our sample.

In this model we assume that:

$$E(Y_i|X_i) = exp(X_i\beta)$$

Where Xi is again a vector, so that:

$$f(Y_i|X_i;\beta) = \frac{exp\left(-exp(X_i\beta)\right)exp(X_i\beta)^{Y_i}}{Y!}$$

Here we use again the maximum likelihood to estimate β. When X is discrete, 100β1 is the percentage change in $(Y_i|X_i)$ caused by a change in X. If X is continuous, 100β1 is the percentage change $(Y_i|X_i)$ is caused by a small change in X. If X is then measured in levels, 100β1 is interpreted as a semi-elasticity, while if X is measured in logs, 100β1 is interpreted as an elasticity.
The variance of a poison model has sometimes consequences on our assumptions, therefore relax them and let:

$$Var(Y_i|X_{i1}) = \sigma^2 E(Y_i|X_{i1})$$

Where σ^2 is a parameter that we estimate. However if it is larger than 1, it implies overdispersion. Overdispersion arises when the observed variance is higher than the variance of a theoretical model, in this case the Poisson model. This overdispersion arises when we observe many 0's in our Yi.

Econometrics of gravity equation

In international trade the Poison model is used to estimate the gravity equation:

$$T_{ij} = \beta_0 Y_i^{\beta 1} Y_j^{\beta 2} D_{ij}^{\beta 3}$$

where we would like to measure the β1, β2 and β3 and where:

- Tij is the trade flow from country i to country j
- Yi is GDP of country i and Yj the GDP of country j
- Dij is the distance between i and j, either cultural, geographical or another important factor that describes distance

We model:

$$E(T_{ij}|Y_i, Y_j, D_{ij}) = \beta_0 Y_i^{\beta 1} Y_j^{\beta 2} D_{ij}^{\beta 3}$$

Where the error is:

$e_{ij} = \frac{T_{ij}}{E(Y_i, Y_j, D_{ij})}$, so that: $E(e_{ij}|Y_i, Y_j, D_{ij}) = 1$

In this case, the conditional variance looks like this:

$$Var(e_{ij}|Y_i, Y_j, D_{ij}) = E(e_{ij}^2|Y_i, Y_j, D_{ij}) - 1$$

Here the squared deviation depends on Yi, Yj and Dij, because it is a less restrictive assumption than when it would be constant, and because different values could have different variances. Some small countries are far away from each other and the

trade flows between these countries is very low, probably even zero. These pairs will have a different variance than all other country pairs that do have a reasonable trade flow between them.

The traditional way to then estimate the regression would be to use logs and estimate the following linear model:

$$ln(T_{ij}) = ln(\beta_0) + \beta_1 ln(Y_i) + \beta_2 ln(Y_j) + \beta_3 ln(D_{ij}) + ln(e_{ij})$$

However, the error is measured in the following way:

$$E[ln(e_{ij})|Y_i, Y_j, D_{ij}] = 0$$

But because we now use a log linear model, it will give inconsistent estimators if we have heteroskedasticity. That is because the error, $ln(e_{ij})$, is a nonlinear function.

Another problem is that by taking a log, we lose all country pairs that have a trade flow of 0 (log of 0 does not exist). We could therefore set the trade flow to ln(Tij+1), to use these observations in our regression. The heteroskedasticity problem remains unsolved.

We could use a Poisson model to overcome both problems. In Stata we could use the command:

poisson Tij Dij Yi Yj covariates, robust

References

Kapoor, S. (2023). Week 36 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037

Kapoor, S. (2023). Week 37 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037

Kapoor, S. (2023). Week 38 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037

Kapoor, S. (2023). Week 39 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037

Kapoor, S. (2023). Week 40 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037

Kapoor, S. (2023). Week 41 [PDF Slides]. Retrieved from: https://canvas.eur.nl/courses/44037