Micro Econometrics

Sol Yates

May 29, 2024

Contents

Lecture 1: Regression

Wed 31 Jan 11:21

Nomenclature:

1. *Estimand* - the quantity of parameter you want to estimate. The theoretical true value that exists in the population (e.g. The average height is the estimand for the average heigh of all people in a country)
2. *Estimator* - a rule, formula, or method used to calculate an estimate of the estimand based on sample data. It is a function of re sample data (e.g. The sample mean (average height of people in a sample) is an estimator of the population mean)
3. *Estimate* - the actual value obtained from the estimator when applied to a specific sample of data. It is a specific numerical value (e.g. If you measured the height of 100 people and calculated their average height to be 175cm - this is the estimate.)

1 Regression

[L1 Regression]

Regression Fundamentals

As an empiricist, differences in economic fortune are hard to explain, as applied econometricians, we believe we can summarise and interpret 'randomness' in a useful way.

Whilst expectations are a population concept, in practice samples rarely consist of the entire population. We use these to make inferences about the population, so the sample CEF is used to learn about the population CEF.

Conceptually, we consider random variables (Y, X, D) from a joint Distribution F: $(Y, X, D) \sim F$ We consider a sample of size n where the Ith draw gives us the variables (Y_i, X_i, D_i) Random sampling where each draw is independent and identically distributed

$$(Y_i, X_i, D_i) \overset{i.i.d.}{\sim} F$$

Notationally, $W_i = (X_i, D_i)$. And the random sampling happens jointly, there's no distinction between Y and W in the sampling process.

Consider the linear predictor

$$E * [Y_i|W_i] = W_i'\beta$$

Where $E*$ denotes the linear predictor such that β are the minimiser of the expected squared error loss $E\left[(Y - E * [Y_i|W_i])^2\right]$ Then, recall that

$$\beta = E\left[W_i W_i'\right]^{-1} E\left[W_i Y_i\right]$$

Which is the traditional OLS estimator where β is a population object, defined based on 2 moments

Recall that the least squares estimator of β is

$$\begin{aligned} b\left(Y_n W_n\right) &= \left(n^{-1}\sum_{i=1}^{n} W_i W_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} W_i Y_i\right) \\ &= \left(W_n' W_n\right)^{-1} W_n' Y_n \end{aligned}$$

Recall that b is a function of random variables Y_n and W_n (an estimator) and also a random variable.

Since we can't directly study $E\left[b\right]$, we study instead $E\left[b|W_n\right]$, which focuses on the conditional distribution of $Y_i|W_i$ (studying $E\left[b\right]$ is difficult due to the non-linearity stemming from the expectation of a ratio is not equal to the ratio of expectations)

If we consider $E\left[b(Y_n, W_n|W_n = w)\right]$, we see

$$E\left[b\left(Y_n, W_n|W_n = w\right)\right] = (w'w)^{-1}w'E\left[Y_n|W_n = w\right]$$

When we are correctly specified, and the conditional expectation $E\left[Y_n|W_n\right] = W_n\beta$, then we have

$$E\left[b\left(Y_n, W_n|W_n = w\right)\right] = \beta$$

Using LIE and fact this is true for any w.

Regression

Studying the relationship between one (dependent) variable y and k other independent variables x_j, $(j = 1, \ldots, k)$

1. Does the Covid vaccine work
2. What are the returns to schooling
3. What is the effect of having internet at home on student's grades
4. Does a job training program decrease the time of getting out of unemployment

We are often interested in a single variable, including other regressors as controls

1.1 Classical Linear Model

Linear regression

- Relies on 5 main Gauss-Markov assumptions
- In small samples is unbiased and BLUE
- In large samples is consistent and asymptotically normal. There is no need for the normality assumption to establish asymptotic distribution

If x contains years of schooling and experience, and the main component of u is innate ability then this assumption implies that ability is uncorrelated with education and experience in the population.

1.2 Multiple Regression

Classical Linear Model

$$\begin{aligned} Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + u_i, i = 1, \ldots, n \\ = x_i\beta + u_i \quad \text{vector notation} \\ Y = X\beta + U \quad \text{matrix notation.} \end{aligned}$$

1. Where β_0 is the intercept, β_j is the parameter (slope) associated with x_j
2. A is the unobserved error term : containing factors other than x_j 's explaining y
3. n is the number of observations

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{1,k} \\ 1 & x_{2,1} & \ldots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{bmatrix} \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Least Squares Estimator

Objective: To estimate the effect of x_j on y, we need to estimate the population parameters $\beta_0, \ldots, \beta_k$.

Ordinary Least squares estimates

β by minimising the sum of squared residuals :

$$\hat{\beta} = argmin \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_k x_{i,k})^2 = \|Y - X\beta\|^2$$

Taking first order conditions

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

It can be shown that

1. Residuals : $\hat{u}_i = y_i - x_i\beta$ with $\sum_{i=1}^{n} \hat{u}_i = 0$
2. Fitted values : $\hat{y}_i = x_i\beta$

It can be shown that in the single regressor model where

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

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2} \\ &= \frac{\hat{\text{Cov}}(x_1, y)}{\hat{v}(x_1)} \\ \text{and} \quad \hat{\beta}_0 &= \bar{y} - \hat{\beta_1}\bar{x_1}\end{aligned}$$

1.3 Gauss Markov Assumptions

Assumption 1. *Linear in parameters (MLR.1)*

The model in the population can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + B_k x_k + u$$

Where $\beta_0, \beta_1, \ldots \beta_k$ are the unknown parameters (constants) of interest and u is an unobservable random error or disturbance term

Assumption 2. *Random Sampling (MLR.2)*

We have a random sample of n observations, following the population model assumption in MLR.1.

- Often referred to as IID assumption
- Ensures that our sample's representative for the population
- Would fail if we observed only part of the population in our sample

Assumption 3. *No perfect colinearity MLR.3*

In the sample (thus pop too), none of the independent variables are constant, and there are no exact linear relationships between the independent variables

- Often referred to as full rank assumption
- Dummy variable trap - not to include a binary for both male and female
- Not to confuse with highly but not perfectly correlated variables (multicollinearity)

Assumption 4. *Zero Conditional Mean MLR.4*

The error u has an expected value of zero given any value of the explanatory variable,

$$E[u|x_1, x_2, \ldots, x_k] = 0$$

- Key to deriving unbiasedness
- If it holds for variable x_j, the variable is exogenous
- Requires at a minimum : all factors in the observed error term must be uncorrelated with the explanatory variables

- Any problem that causes u to be correlated with any of the x_j 's causes this assumption to fail and OLS to be biased !
- Examples for endogeneity : misspecified functional form, omitting important variables, measurement error and any x_j being jointly determined with y

Assumption 5. *Homoskedacity MLR.5*

The error u has the same variance given any value of explanatory variables, in other words

$$V[u|x_1, \dots x_k] = \sigma^2$$

- The variance of the unobserved error u conditional on the explanatory variables is the same for all combinations of the outcomes of the explanatory variables
- If this assumption fails, we speak of heteroskedatic errors
- This assumption is not needed for unbiased/ consistency but for efficiency of OLS
- This also means that $V[y|x] = \sigma^2$

1.4 Small Sample Properties

Unbiasedness of OLS

Under Assumption 1 - Assumption 4, the OLS estimator is unbiased

$$E[\hat{\beta}_j] = \beta_j \quad \text{for } j = 0, \dots, k$$

For any values of the population parameter β_j.

Or, we say the OLS estimators are unbiased estimators of the population parameters

However,

- Might not exactly be the population value
- Deviations from the population value are not systematic
- If we were to repeat the estimation on several random samples the deviations should average out to zero

Variance

Formally, the variance of $\widehat{\beta}$ revolves around the structure of $E\left[e_n e_n' | W_n\right] = \Sigma_n$:

$$\begin{aligned}\mathbb{V}(\hat{\beta} \mid W_n) &= (W_n' W_n)^{-1} W_n' \mathbb{E}(\epsilon_n \epsilon_n' \mid W_n) W_n (W_n' W_n)^{-1} \\ &= (W_n' W_n)^{-1} W_n' \Omega_n W_n (W_n' W_n)^{-1}\end{aligned}$$

To simplify further, we can consider homoscedasticity, where $\Sigma = \sigma^2 I_n$ and simplify our variance to

$$V\left[\widehat{\beta}\right] = \sigma^2 \left(W_n' W_n\right)^{-1}$$

Which, beyond $\text{Cov}(\varepsilon_i, \varepsilon_i) = 0$ assumes $V\left[\varepsilon_i | W_i\right] = V\left[\varepsilon_i\right]$

A feasible estimator for the homoskedatic variance estimand, where k is the number of regressors in β (excluding the constant), follows using empirical analogs:

$$\hat{V}\left(\widehat{\beta}\right)_{\text{homoskedatic}}\left[\sigma^2\left(W_n'W_n\right)^{-1}\right] \quad \widehat{\sigma}^2 = (n-k-1)^{-1}\,\hat{\varepsilon_n}'\hat{\varepsilon_n} \tag{1}$$

Sampling variance of the OLS slope estimators

Under Assumption 1 - Assumption 5, conditional on the sample values of the independent variables the variance is

$$V[\hat{\beta}_j] = \frac{\sigma^2}{SST_j(1-R_j^2)} \quad \text{for } j = 0,\ldots,k$$

Where $SST_j = \sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ is the sum of total sample variation in x_j and R_j^2 is the R-squared from regressing x_j on all other independent variables (including an intercept)

- The standard error formulas make it apparent that we need variation in the regressors to increase precision
- The R_j^2 representation makes it also apparent that high multicollinearity increases the variance of the estimator

Matrix Representation

General formula in matrix form (including the intercept)

$$V[\hat{B}_j] = \sigma^2(X'X)^{-1}$$

The variance of the j-the parameter estimate

$$\sigma^2(X'X)^{-1}_{[j+1,j+1]}$$

Gauss Markov Theorem

Theorem 1. *Gauss Markov Theorem*

Under Assumption 1 - Assumption 5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUE)s of

$$\beta_0, \beta_1, \ldots, \beta_k$$

- If the assumptions hold, we do not need to look for another unbiased estimator since this is the best
- Best meaning the most efficient, with smallest variance

There exists a more general (and asymptotic) version of the GM theorem ([**JW-CS- 14**]). While stating that OLS has the smallest variance in the class of linear, unbiased estimators, it does not allow us to compare OLS to unbiased estimators that are not linear in the vector of observations on the dependent variable

Small Sample Inference Error Variance Estimation

Since we are interested in performing inference, we need : variance (standard error) and the distribution of parameter estimator

Firstly Estimation of the error variance :

$$\sigma^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1}$$

We can show this estimator is unbiased under

Theorem 2. *Unbiased Estimation of* σ^2

Under the GM assumptions (1-5),

$$E[\hat{\sigma^2}] = \sigma^2$$

Recall that our distribution assumptions come from considering the following statistics:

$$T = \frac{\widehat{\beta} - \beta}{\sqrt{\hat{V}}/v}$$

Confidence Intervals And Finite Sample Performance

The feasible estimator for $\widehat{\sigma}^2(x)$ was not made explicit, for the homoskedatic case, we have a simple estimator eq. (1).

95 % confidence intervals based on these asymptotic results:

$$\text{CI} = \left(\beta - t_{0.975}^{n-2} \times \sqrt{\hat{V}}, \beta + t_{0.975}^{n-2} \times \sqrt{\hat{V}}\right)$$

Standard Errors

Is the $\sqrt{\text{variance}}$

$$Sd(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{SST_j(1-r_j^2)}}$$

$$Se(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{SST_j(1-R_j^2)}} = \frac{\hat{\sigma}}{\sqrt{n}sd(x_j)\sqrt{1-R_j^2}}$$

Where $sd(x_j) = \sqrt{n^{-1}\sum_i (x_{ij} - \bar{x}_j)^2}$

Standard errors shrink to zero at the rate $\frac{1}{\sqrt{n}}$ (since in denominator)

Assumption 6. *Normality MLR.6*

The population error u *is independent of the explanatory variables* $x_1, x_2, \ldots, x_k$ *and is normally distributed with zero mean and variance* $\sigma^2 : u \sim \mathcal{N}(0, \sigma^2)$

This is a stronger assumption than 1-5 and means we are necessarily assuming zero conditional mean (4) and homoskedacity (5).

Theorem 3. *Normal Sampling Distributions*

Under Assumption 1-Assumption 6, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, V(\hat{\beta}_j)$$

Where $V[\hat{B}_j] = \sigma^2(X'X)^{-1}$

Therefore,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim \mathcal{N}(0,1)$$

Or, $\hat{\beta}|X \sim MVN(\beta, \sigma^2(X'X)^{-1}$ *(matrix notation)*

1.5 Asymptotic Properties

Consistency

Assumption 7. *Zero Mean* ***and*** *Zero correlation (MLR.4')*

$$E[u] = 0 \quad and \quad Cov[x_j, u] = 0, \quad for \quad j = 1, 2, \ldots, k.$$

- *If we are only interested in consistency : this replace zero conditional mean (MLR.4)*
- *However, zero conditional mean important for finite sample and to ensure that we have properly modelled the population regression function* $E[y|x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- *This gives us the average or partial effects of* x_j *on the expected value of y.*

Theorem 4. *Consistency of OLS*

Under assumptions MLR.1 - MLR.4, (or replacing 4 with 7), the OLS estimator $\hat{\beta}_j$ *is consistent for* β_j *for all* $j = 1, 2, \ldots, k$

Consistency means that when n goes to ∞*, the estimator will recover the population value in probability :*

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_j = \beta_j$$

Essentially, the asymptotic bias shrinks to 0.

For the simple model with one regressor : $y_i = \beta_0 + \beta_1 x_{i,1} + u_i$, to show consistency :

1. Write down the formula for $\hat{\beta}_1$ and plug in y_i:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_{i,1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^n (x_{i,1} - \bar{x}_1)^2} \\ &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (x_{i,1} - \bar{x}_1)(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^n (x_{i,1} - \bar{x}_1)^2}\end{aligned}$$

2. Apply the LLN: $\bar{X}_n \xrightarrow{P} X$ as $n \to \infty$

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1 + \frac{\operatorname{cov}[u, x_1]}{V(x_1)} \qquad = \beta_1$$

Since $\text{Cov}[u, x_1] = 0$ by Assumption 4 (where Assumption 7 is a weaker assumption because assumption 4 implies assumption 7).

Inconsistency of OLS Just as failure of $E[u|x_1, \ldots, x_k]$ causes bias in the OLS estimators, correlation between u and any of $x_1, \ldots, x_k$ generally causes *all* of the OLS estimators to be inconsistent. I.e. If the error is correlated with any of the independent variables, then OLS is biased and inconsistent (bias persists as the sample size grows in fact)

In the simple regression case, we can obtain the inconsistency from the first part of **??** which holds whether or not u and x_1 are uncorrelated. The *inconsistency* is sometimes also called the **asymptotic bias**:

$$\text{plim}\,\widehat{\beta}_1 - \beta_1 = \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}$$

Which, because $\text{Var}(x_1) > 0$ the inconsistency in $\widehat{\beta}_1$ is positive if x_1 and u are positively correlated and vice versa. But we cannot estimate how big the covariate is because u is unobserved.

Notes

- In a single regressor model : $\beta_1 = \frac{Cov(y,x_1)}{V(x_1)}$
- Including more regressor changes this expression for the population estimate β_j but since the effect of the other covariates is partialled out, we still recover β_j
- Multicollinearity only affects the variance of the estimator but not consistency

Theorem 5. *Asymptotic Normality of OLS*

Under the Gauss-Markov assumptions (1-5),

1. $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim \mathcal{N}(\beta_j, \frac{\sigma^2}{a_j^2})$ *where* $(\frac{\sigma^2}{a_j^2}) > 0$ *is the asymptotic variance of* $\sqrt{n}(\hat{\beta}_j - \beta_j)$ *for the slope coefficients,* $a_j^2 = plim\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{ij}^2$ *where the* $\hat{r}_{ij}^2$ *are the residuals from regressing* x_{ji} *on the other independent variables. We can say that* $\hat{\beta}_j$ *is asymptotically normally distributed*
2. $\hat{\sigma^2}$ *is a consistent estimator of* $\sigma^2 = V(u)$
3. *For each j,* $(\frac{\hat{\beta}_j - \beta_j)}{sd(\hat{\beta}_j} \sim \mathcal{N}(0,1)$ *(where sd unobserved)*
4. *For each j,* $(\frac{\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j} \sim \mathcal{N}(0,1)$ *(where se estimated)*

Where $se(\hat{\beta}_j)$ *is the usual OLS estimator*

Matrix Form

$$\sqrt{n}(\hat{\beta} - \beta \sim \mathcal{N}\left[0, \sigma^2(\text{plim}\,\frac{X'X}{n})^{-1}]\right]$$

With plim $\frac{X'X}{n} = E[x'x]$

- For convergence, one needs the asymptotic normalisation $\sqrt{n}$
- However we are interested in the variance of $\hat{\beta}$. For estimation, we use the sample analog of the variance covariance and remove the asymptotic normalisation again by dividing by n
- We obtain the asymptotic variance : $\hat{AV} = \hat{\beta} = \sigma^2(X'X)^{-1}$

This is a very important result for inference. The normality assumption is not needed in large sample. Therefore, regardless of the error distribution, if properly standardised, we have approximate normal standard distributions. We can use the (unobserved) $sd(\hat{\beta}_j$ or the observed $se(\beta_j)$ to achieve this result, where we can estimate the latter since it depends on $\hat{\sigma^2}$

Then because the t distribution approaches the normal distribution for large degrees of freedom, we can also say that $(\beta_j \hat{-} \beta_j)/se(\hat{\beta}_j \sim t_{(n-k-1)}$. But we still need homoskedacity, and with large sample, all the testing issues still apply.

Partialling Out

Intuitively, $\hat{\beta}_1$ measures the sample relationship between y and x_1 after the other regressors have been partialled out

1. Model :
$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \quad E[u, x_1, \ldots, x_k] = 0$$
2. Regress $x_1 \sim x_1 + x_2 + \ldots + x_k$ and compute the residual $\hat{r_{i1}}$
3. Regress $y \sim \hat{r_1}$ which yields the OLS estimate $\hat{\beta}_1$
4. One can show that the resulting OLS estimator from the regression in 3, equals the OLS estimator for β_1 from a regression based on the model in 1

In general form, this is called the **Frisch-Waugh Theorem**

Also giving us the regression anatomy formula :

$$\beta_j = \frac{Cov(y_i, \hat{r}_{i,j}}{V(\hat{r}_{i,j}}$$

1.6 Interpretation And Modelling

1. $y = \beta_0 + \beta_1 x_1 + u$
2. $y = \beta_0 + \beta_1 \log(x) + u$
3. $\log(y) = \beta_0 + \beta_1 x + u$
4. $\log(y) = \beta_0 + \beta_1 \log(x) + u$

Model	Dep. Variable	Ind. Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
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)	$\%\Delta y = (\beta_1)\%\Delta x$

Source: Wooldridge (2020)

Figure 1: Interpretation And Log Transformation Table

Example. *Log-Log model*

```
Call:
lm(formula = LPROD ~ LArea + LLabor + LNPK, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80481 -0.16484  0.05152  0.21016  0.89923

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.66964    0.24870  -6.714 7.98e-11 ***
LArea        0.32976    0.06240   5.285 2.25e-07 ***
LLabor       0.38375    0.06581   5.831 1.28e-08 ***
LNPK         0.28292    0.03993   7.086 8.01e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3302 on 340 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8581
F-statistic: 692.4 on 3 and 340 DF,  p-value: < 2.2e-16
```

Figure 2: R Output - Tut 5

This is a log-log model. A 1 % increase in the labour force increases the rice output by 0.38 %. A 1 % increase in the use of fertiliser increases the rice output by 0.28 %. A 1 % increase in land increases the rice output by 0.32 %

Linear Probability Model

Model $y = \beta_0 + \beta_1 x + u$ such that $y = \{0, 1\}$ is a binary dependent variable

- Since y is binary

$$E[y|x] = P(y = 1|x) = \beta_0 + \beta_1 x$$
$$1 - E[y|x] = P(y = 0|x) = 1 - \beta_0 - \beta_1 x$$

  With marginal effects

$$\frac{\partial E[y|x]}{\partial x} = \beta_1$$

- The predicted values are probabilities of the outcome being equal to 1
- Interpretation : β_1 is the change in the probability that $y = 1$ for a 1 unit increase in x_1 (percentage points)

An Aside on Predicted Probabilities

- Pros are the estimation and interpretation is straight forward
- Cons are for prediction, the predicted probabilities can be outside the interval $[0, 1]$
- Another con is the errors are Heteroskedacity and hence violate the gauss Markov assumption

$$V(y|x) = P(y = 1|x)(1 - P(y = 1|x))$$

Note other binary dependent variable models are probit and logit

Binary Regressors

Model

$$Y + \beta_0 + \beta_1 x + u \quad E[x|u] = 0$$

Where

- $x \in \{0, 1\}$ is binary variable
- Often also referred to as 'dummy' variable
- Example : effect of gender on hourly wage. Let x = 1 if the individual is a woman (x = 0 man)
- Often used to evaluate a treatment effect such as the effect of an intervention, a policy, a program
- OLS results in comparing group averages

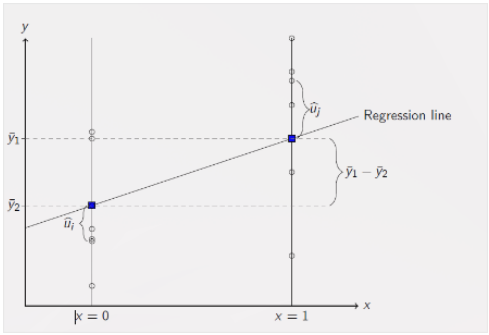


Figure 3: Relationship between y and x when x is binary

The white circles represent typical datapoints and the blue rectangles represent sample averages.

The average effect on y is the difference between the averages of both groups

Population Raw Differential

The CEF for $x = 1$ and $x = 0$:

$$\begin{aligned} \mu_1 \equiv E[y|x=1] &= \beta_0 + \beta_1 + E[u|x=1] \\ \mu_0 \equiv E[y|x=0] &= \beta_0 + E[u|x=0] \\ &\text{under zero conditional mean:} \\ E[y|x=1] &= \beta_0 + \beta_1 \\ E[y|x=0] &= \beta_0 \\ &\text{hence} \\ \beta_1 &= \mathbf{E[y|x=1] - E[y|x=0]} \end{aligned}$$

Where the parameter to be estimated by OLS is the **population raw differential**

Sample Raw Differential

Now we can replace the conditional expectations by their sample analogue, that the conditional expectation of y for women is the mean outcome of women and the conditional expectation of y for men is the mean outcome of men.

- Replacing the population means by their sample averages, we obtain the OLS estimators
- $\hat{\beta}_1$ is the **sample raw differential**

$$\hat{\beta}_1 = \frac{1}{n_1} \sum_{i:x_i=1} y_i - \frac{1}{n_0} \sum_{i:x_i=1} y_i = \bar{y}_1 - \bar{y}_0 \quad , \quad \hat{\beta}_0 = \bar{y}_0$$

Where n_1 (n_0) is the number of observations with x = 1 (x = 0)

Let $\widehat{\beta}_0 = 7$ and $\widehat{\beta}_1 = -2.5$ then men earn on average 7 GBP per hour, and women earn on average 2.5 GBP per hour less than the *average* man

1.7 Timeout: Testing Equal Means

To test whether the population means for 2 sub-samples are the same

$$H_0 = E[y|x = 1] = E[y|x = 0] =\equiv \beta_1 = 0$$

With Test stat :

$$\hat{\beta}_1/se(\hat{\beta}_1 = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\hat{\sigma_1^2}/n_1 + \hat{\sigma_0^2}/n_0}}$$

Where $\hat{\sigma^2}$ are the estimated group specific error variances

log(y) and Binary Regressors

Model : $\log(y) = \beta_0 + \beta_1 x$ for $x \in \{0, 1\}$

$$\log(y) = \begin{cases} \log(y_1) = \beta_0 + \beta_1 + u & \text{if } x = 1 \\ \log(y_0) = \beta_0 + u & \text{if } x = 0 \end{cases}$$

Log Point Interpretation If the resulting $\%\Delta y/100$ is small :

$$\begin{aligned} \beta_1 = \log(y_1) - \log(y_0) = \log(\frac{y_1}{y_0}) = \log\left(1 + \frac{y_1 - y_0}{y_0}\right) \\ = \log(1 + \frac{\%\Delta y}{100}) \approx \%\Delta y/100 \\ 100\beta_1 \approx \%\Delta y \end{aligned}$$

If $\%\Delta y/100$ is small, one can interpret β_1 as the *raw differential*

$$log(wage) = \beta_0 + \beta_1 educ + u$$

(a) Estimate the model using OLS and interpret.

```
Call:
lm(formula = lwage ~ educ, data = dat)

Residuals:
Min       1Q   Median       3Q      Max
-1.94620 -0.24832  0.03507  0.27440  1.28106

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.973063   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
```

Figure 4: Log-Level Model

Here, the effect of education is 0.059839 which is significantly different from zero at the 1% level. Assuming all assumptions are met for OLS, one more year of education increases log wage by 5.9 %.

Exact Interpretation - Percentage Change

$$\frac{\Delta y}{y_0} = \frac{y_1 - y_0}{y_0} = \frac{y_1}{y_0} - 1 = frac\exp beta_0 + \beta_1 + u\exp(\beta_0 + u - 1 = \exp(\beta_1) - 1$$

Then, plugging the estimate into the equation

$$\%\Delta y = 100[\exp(\hat{\beta}_1) - 1]$$

Interpretation - $\frac{\delta y}{y_0} = -0.26$ means that a woman's wage is 26% below that of a comparable man's wage

Categorical Regressors

- Some characteristics such as regions are originally categorical
- We can render categorical variables based on an originally continuous variable, say bins based on firm size
- The solution is to create multi-category dummies d_k to account for differential effects
- Say the effect of law school ranking on median starting salaries, $\ln(y_i)$ where they found better ranked schools result in higher wages

In order to estimate whether there is a differential effect for the different ranks include the categories as dummies, say

$$D_{i,1} = \begin{cases} 1, \ if \ 1 \leqslant rank \leqslant 10 \\ 0, \ otherwise \end{cases} \quad :D_{i,6} = \begin{cases} 1, \ if \ rank > 100 \\ 0, \ otherwise \end{cases}$$

The model

$$\ln(y_i) = \beta_0 + \beta_1 d_{i,1} + \ldots \beta_5 d_{i,5} + x\gamma + u_i$$

Where we exclude one dummy $d_{i,6}$ to avoid perfect-collinearity (dummy var trap)

Interpretation

- Usual interpretation for a binary regressor with respect to base category
- $\beta_j = E[\ln(y)|d_j = 1] - E[\ln(y_i)|d_6 = 1]$ for $j = 1, \ldots, 5$
- β_0 : log median starting salary for the omitted (base) category, the largest rank category
- Estimated percentage change for the first rank bin compared to the largest bin 101.29%

Interaction Terms

Model : $y + \beta_0 + \beta_1 x_1 + \beta_2 d + \beta_3 x_1 \times d + u$

- Where x_1 is continuous
- Then the effect of x_1 is different for each group : $\frac{\partial E[y|x_1,d]}{\partial x_1} = \beta_1 + \beta_3 \times d$
- If d = 1, then $\frac{\partial E[y|x_1,d]}{\partial x_1} = \beta_1 + \beta_3$
- If d = 0, then $\frac{\partial E[y|x_1,d]}{\partial x_1} = \beta_1$
- If d=1 for women, then a unit increase in x_1 leads to a $\beta_1 + \beta_3$ increase for women and an increase for men of β_1. That is, the returns to x_1 are for women β_3 higher.
- If the independent variable is in log, we can interpret the coefficient as 1oo*[parameter]%

Slightly different model, where both regressors are binary, $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_1 \times d_2 + u$

- We can compute expectation for each case:

$$\begin{aligned} E[y|d_1 = 0, d_2 = 0] &= \beta_0 \\ E[y|d_1 = 0, d_2 = 1] &= \beta_0 + \beta_2 \\ E[y|d_1 = 1, d_2 = 0] &= \beta_0 + \beta_1 \\ E[y|d_1 = 1, d_2 = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \end{aligned}$$

- We can now interpret the obtained regression coefficients according to these differentials
- For example, if $d_1 = 1$ for female and $d_2 = 1$ for being married, then the outcome is on average for married women by $\beta_1 + \beta_2 + \beta_3$ higher than for single men

Polynomials

We often use to model non-linear relationships such as the diminishing returns to experience

- Model : $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- Interpretation : $\frac{\partial E[y|x]}{\partial x} = \beta_1 + 2 \times \beta_2 x$
- We can compute the average change in y by a one unit change in x for a specific point in time (say for 0, 1, 2 years of experience)

Econometrics Techniques

- Nonlinear relationships : EG modelling, non-parametric regression
- Standard errors: robust, clustered and bootstrap standard errors

- Addressing homogeneity :

Estimator	Min Data Requirement	Notes
Regression, Matching	Single cross-section	Observables
Instrumental Variables	Single cross-section	Valid instrument
Randomized Controlled Trials	Single cross-section	Program manipulation
Fixed Effects	Panel Data	Only FEs omitted
Random Effects	Panel Data	FEs uncorrelated
Difference-in-Differences	Repeated cross-sections	Common trends
Regression Discontinuity	Single cross-section	Running variable

Figure 5

Lecture 2: Standard Errors

Sun 04 Feb 17:54

[**L2**]

2 Standard Errors

Significance Level (α)	**Two-Tailed Critical Value**	**One-Tailed Critical Value**
10% (0.10)	± 1.645	± 1.28
5% (0.05)	± 1.96	± 1.645
1% (0.01)	± 2.576	± 2.33

Table 1: Critical values for two-tailed and one-tailed tests at common significance levels.

2.1 Introduction

- After a point estimate, we want to know the statistical significance to draw conclusions
- This typically requires the standard error and the distribution
- If the standard errors are wrong we cannot use the usual t / F statistics for drawing inference
- We either have too large SEs,
  - Zero might be included in the CI when it should not be
  - There is a risk of not detecting an effect even there was
- Or, too small SEs
  - Zero might not be in the CI when it should be
  - We may claim the existence of an effect when in reality there is none
  - Of course, this is worse - a **wrong** SE can lead to a **wrong** conclusion!

Robust SE

- Traditionally, inference assumes homoskedacity
- But the variance of error terms might be different for different observations depending on their characteristics
- Heteroskedacity robust SE accounts for this

Standard Errors

- Traditional estimation relied on random sampling
- In the case of data with a group structure, the error terms might be correlated
- To account we use clustered SE

Bootstrap

- Bootstrap is a re sampling method that offers an alternative to inference based on asymptotic formulas convenient in cases where the sampling distribution is unknown

2.2 Heteroskedacity - Robust Standard Errors

Heteroskedacity Problems

- Traditional inference assumes homoskedatic errors $V(u|x) = \sigma^2$
- This implies that the variance of the unobserved error a, is constant for all possible values of all the regressor x's
- Since the proofs for unbiasedness and consistency do not depend on this assumption we still obtain unbiased and consistent OLS estimates
- However, if this is not true (σ_i^2) then the errors are called **heteroskedatic** and traditional variance estimators are biased
- Heteroskedacity robust SE specifically in the CS case
- If the degree of heteroskedacity is low ,the traditional variance estimator might be less biased

Example (Returns to education).

Regressing wage ∼ educ

It is reasonable to believe that the variance is unobserved factors hidden in the error term differers by educational attainment

Individuals with higher education : potentially more diverse interests and more job opportunities affecting their wage

Individuals with very low education : fewer opportunities and often must work at the minimum wage, the error variance is typically lower

Variance Estimation With Heteroskedacity

Simple regression : $y = \beta_0 + \beta_1 x + u$

We know $\hat{b}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x} u_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Which is a function of the error terms

Therefore :

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

Where $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2$

- Where σ_i^2 is conditional variance of error term (depending on each individual)
- If $\sigma_i^2 = \sigma^2$ the formula reduces to the traditional (OLS variance) formula : $V(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$
- We have to estimate the conditional variance of the error, we do this by tang the residuals of OLS, squaring them and replacing them in the following formula for the error variance
- This leads to the following heteroskedacity robust estimator (simple regression model) :

$$\hat{V}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \bar{u}_i^2}{SST_x^2}$$

Where u_i^2 are the OLS residuals

Generalisation

The formula generalises to

$$\hat{V}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

Where the σ_i^2 are replaced by residuals soured from OG regression and the $\hat{r_{ij}}$ are the residuals from regressing x_j on all other independent variables.

Where $\hat{r_{ij}}$ is the i-th residual from regressing x_j on all other independent variables and SSR_j the sum of squared residuals from this regression

- Robust to heteroskedacity **of any form** (inc homoskedacity)
- Often also called white, huber, eicker SE
- Sometimes degrees of freedom adjustment by multiplying $\frac{n}{n-k-1}$
- But with **drawback** that it only has asymptotic justification (need large sample for it to be valid)

Matrix Representation - Asymptotic Variance

Model : $y = X\beta + U$

We know $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V)$ Where

$$V = E[E[X'X]]^{-1}[E[X'Xu^2]][E[X'X]]^{-1}$$

with fixed regressors (replace with sample analog)

$$= [\frac{1}{n}X'X]^{-1}[\frac{1}{n}X'\psi X][\frac{1}{n}X'X]^{-1}$$

And the variance-covaraince matrix ψ

$$\psi = \begin{bmatrix} V(u_1|x) & 0 & \dots & 0 \\ 0 & V[u_2|x] & \dots & 0 \\ \dots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & V[u_n|X] \end{bmatrix}$$

Eventually, $\frac{V}{n} = AV(\hat{\beta})$

Matrix Representation - Estimation

- We can then find an estimate the middle term by :

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \hat{x}_i' x_i = \frac{1}{n} X\hat{\psi}X$$

- Where $\hat{\psi} = diag[\hat{u_i^2}, \ldots,]$

$$\hat{V} = [\frac{1}{n}X'X]^{-1}\frac{1}{nX'\hat{\psi}}X[\frac{1}{n}X'X]^{-1}$$

- In order to estimate the Asymptotic Variance (AV) $\hat{\beta}_j$, we need to remove the asymptotic normalisation by dividing by n
- Resulting Estimator :

$$\hat{AV} = n[X'X]^{-1}\frac{\sum_{i=1}^{n}\hat{u}_i^2 x_i' x_i}{n}[X'X]^{-1}$$

- Sometimes corrected by the degrees of freedom $n/n-k-1$ to improve finite sample properties
- SEs : square root of the diagonal elements
- Recall that under homoskedacity, we obtain $\sigma^2(X'X)^{-1}$

Example. *Returns to Education Reg1 = lm(wage ~ educ, data =wage1)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.21158 -0.36393 -0.07263  0.29712  1.52339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
educ        0.082744   0.007567  10.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1843
F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16

> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.5837727  0.0982339  5.9427 5.118e-09 ***
educ        0.0827444  0.0077389 10.6920 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: R regression output

In which we have used coeftest and vcovHC HC1 variance covariance matrix for one form of the robust one. We have obtained the estimates in both cases.

*Comparing, we have the **same estimate**, however the **SE** in the robust case are slightly bigger. This isn't a great example since it doesn't change significance however it shows both estimators can give different SE, but the estimate from OLS remains the same.*

Breusch-pagan Test For Heteroskedacity

- Testing the null hypothesis of constant variance

$$H_0 : V(u|x_1, \ldots, x_k) = E(u^2|x_1, \ldots, x_k) = \sigma^2$$

  Where $V[u|x] = E[u^2|x] - \underbrace{E[u|x]}_{0}$

- Then assuming a linear relationship :

$$u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v, E[v|x_1, \ldots, x_k] = 0$$

- Since we cannot observe the errors (u^2), we replace them with the residuals ($\hat{u}^2$) and estimate the regression:

1. Estimate
$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + error$$
   Recover the $R^2_{\hat{u}}$
2. Hypothesis : $\delta_1 = \ldots = \delta_k = 0$
3. Test statistic:
$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \overset{H_0}{\sim} \mathcal{F}_{k,n-k-1} \quad \text{or}$$
$$LM = nR^2_{\hat{u}^2} \overset{H_0}{\sim} \chi^2_k$$
4. Decision : if the p-value is small enough (typically < 0.05), we **reject** the null of homoskedacity

Exercise 1 (Heteroskedacity with 2 Categories). *Model $y_i = \beta_0 + \beta_1 d_i + u_i$, $i = 1, \ldots, n$ where d_i is a binary variable*

Let $n_1 = \sum_i d_i$, $n_0 = \sum_i (1 - d_i)$, $n = n_1 + n_0$ and $p = \frac{n_1}{n}$ (probability of being trated, share of treated ind in samp / n)

We have seen that $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ and $\hat{\beta}_0 = \bar{y}_0$ (differences in group mean outcomes) ($\hat{\beta}_0$ is intercept, mean of untreated)

Under homoskedacity in small sample conventional t statistic has a t-distribution

Heteroskedacity here means that the variances in the $d_i = 1$ and $d_i = 0$ population are different: the exact small sample distribution for this problem is unknwon

Differences in the standard error formulae depend on how the variance in d_i is modelled (residual as difference between outcome and group mean outcome)

- *Note $\hat{u}_i = y_i = \bar{y}_i$ for $d_i = I$, $I \in \{0, 1\}$*
- *Define $s^2_I = \sum_{i:d=I} (y_i - \bar{y}_I)^2$ (which is the estimated sum of squared residuals in each group)*
- *Under conventional SEs: $\hat{\sigma^2}(X'X)^{-1}$ with estimate of $\hat{\sigma^2} = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u_i}^2$*
- *Where $\sum_{i=1}^n \hat{u_i}^2 = \sum_{i:d=1} \hat{u_i}^2 + \sum_{i:d=0} \hat{u_i}^2 = s^2_1 + s^2_0$ (sum od squared resid = sum of residuals squared for treated and untreated ind)*

- *Hence* $\hat{\sigma}^2 = \frac{s_1^2+s_0^2}{n-2}$ *(is equal to n-2 since have single regressor and intercept)*
- *Now,* $(X'X)^{-1}_{[2,2]} = \frac{n}{nn_1-n_1^2}$ *(if interested in slope, take X and 2,2 element equal to this expression, using this we can take estimator for variance)*
- *Hence* $\hat{V}(\hat{\beta}_1)_c = \frac{n}{n_1 n_0}\frac{s_1^2+s_0^2}{n-2}$ *(conventional variance estimator if replace elements by percentage shares)*
- *It can be shown that* $\hat{V}(\hat{\beta}_1)_c = \frac{1}{np(1-p)}\frac{s_1^2+s_0^2}{n-2}$
- *For robust SEs :* $\hat{\sigma^2}(X'X)^{-1}(X\hat{\psi}Z)(X'X)^{-1} \rightarrow \hat{V}(\hat{\beta}_1)_r = \frac{s_1^2}{n_1^2} + \frac{s_0^2}{n_0^2}$
- *When* $\frac{s_1^2}{n_1} = \frac{s_0^2}{n_0}$*, both estimates coincide (for large n)*
- *When* $n_1 = n_0 = \frac{n}{2}$ *they also coincide, when the data are balanced, the robust SE won't differ much from the traditional one under heteroskedacity*
- *If both groups variances are the same, then both estimates coincide, because then we have homoskedacity*
- *Also if we have the same indiviudals for treated and untreated groups, then they also coincide, so if we have very balanced data (2 cat) the robust SE won't differ much from the traditional one*

BP test

- *Interpretation of the BP test*
- *Recall the regression* $\hat{u_i}^2 = \delta_0 + \delta_1 d_i + v$

$$\hat{\delta}_0 = \frac{\sum_{i:d=0}\hat{u_i}^2}{n_0} = \frac{s_0^2}{n_0}\hat{\delta}_1 = \frac{\sum_{i:d=1}\hat{u_i}^2}{n_1} - \frac{\sum_{i:d=0}\hat{u_i}^2}{n_0} = \frac{s_1^2}{/}n_1 - \frac{s_0^2}{n_0}$$

- *Testing* $H_0 : \delta_1 = 0$ *is equivalent to testing* $\sigma_1^2 = \sigma_0^2$

Example (Housing Price Equation). *Log is sometimes used to get rid of heteroskedacity*

Model : $price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$

Where price is the housing price, lotsize the size of the lot, size size of house in sq ft

We want to estimate teh above regression and test for heteroskedacity and see whether using logs in the dependent variable changes our conclusion

```
Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-120.026  -38.530   -6.555   32.323  209.376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6607
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

Figure 7: Housing Price Equation Output 1

We cannot really learn much about heteroskedacity, although lot and size is statistically significant

Testing for heteroskedacity using BP test, predicting residuals from previous regression and squared them, then we take the squared residuals and regress on independent variables

```
Call:
lm(formula = u.hat2 ~ lotsize + sqrft + bdrms, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -9044  -2212  -1256    -97  42582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.523e+03  3.259e+03  -1.694  0.09390 .
lotsize      2.015e-01  7.101e-02   2.838  0.00569 **
sqrft        1.691e+00  1.464e+00   1.155  0.25128
bdrms        1.042e+03  9.964e+02   1.046  0.29877
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6617 on 84 degrees of freedom
Multiple R-squared:  0.1601,    Adjusted R-squared:  0.1301
F-statistic: 5.339 on 3 and 84 DF,  p-value: 0.002048

>
> # Compute F-stat by hand: recover the R^2:
> R_u2 = summary(reg.res)$r.squared
> df = reg.res$df # n-k-1
> k = 3
>
> # F-stat:
> F = (R_u2/k) / ((1-R_u2)/(df))
> F
[1] 5.338919
```

Figure 8: Housing Price equation output 2

We obtain the f stat, testing for joint normality of parameter estimate, 5.3 with p value <0.05, testing for heteroskedacity using BP test leads us to reject the null of homoskedacity

```
Call:
lm(formula = log(price) ~ log(lotsize) + log(sqrft) + bdrms,
    data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-0.68422 -0.09178 -0.01584  0.11213  0.66899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.29704    0.65128  -1.992   0.0497 *
log(lotsize)  0.16797    0.03828   4.388 3.31e-05 ***
log(sqrft)    0.70023    0.09287   7.540 5.01e-11 ***
bdrms         0.03696    0.02753   1.342   0.1831
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1846 on 84 degrees of freedom
Multiple R-squared:  0.643,     Adjusted R-squared:  0.6302
F-statistic: 50.42 on 3 and 84 DF,  p-value: < 2.2e-16
```

Figure 9

Does our question change if we use logs? Running the regression we obtain the above, not telling us much again, but helps us to predict residuals based on this regression, then we can test for heteroskedacity

```
Call:
lm(formula = u.hat.ln2 ~ log(lotsize) + log(sqrft) + bdrms, data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-0.05601 -0.03011 -0.01687  0.00523  0.40978

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.509994   0.257857   1.978   0.0512 .
log(lotsize) -0.007016   0.015156  -0.463   0.6446
log(sqrft)   -0.062737   0.036767  -1.706   0.0916 .
bdrms         0.016841   0.010900   1.545   0.1261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07309 on 84 degrees of freedom
Multiple R-squared:  0.04799,   Adjusted R-squared:  0.01399
F-statistic: 1.411 on 3 and 84 DF,  p-value: 0.2451
```

Figure 10

Doing the same as before (without logs) we take our residuals, square the, then regress on individual variables. Giving us f stat of 1.4, which given the p val of 0.2 leads us to failing to reject the null of homoskedacity . Thus our initial SE werent very useful, but using the logs we can assume homoskedacity.

Heteroskedacity Conclusion

- Use robust SE when heteroskedatic errors
- But there is a danger of small sample bias from robust SE (arising from asymptotic justification)
- Under homoskedacity or little heteroskedacity, it might be preferable to use the traditional OLS variance estimator
- It is recommended to report both the robust and conventional standard error and suggest to take the maximum of both for inference
- White test for heteroskedacity includes the squares and cross-products of the independent variables
- LPM : built in heteroskedacity → need to compute robust SEs
- Using logs in the dependent variables has been seen to improve in terms of heteroskedacity in many applications

2.3 Clustered Standard Errors

Illustration of Moulton Problem

The Moulton Problem - biased standard errors where observations are not independent within groups, but the regression model incorrectly specifies that they are.

The clustering of data can lead to an underestimation of standard errors → statistical tests likely overly optimistic about the significance.

Closely related to correlation over time in DiD, state average employment rates are correlated over time[**ME-AP**]

- Pillar assumption is random sampling
- There is potential dependence of data within a group structure
  - Exam grades of children from same class or school : grades are correlated because of the same school, teacher and background / class environment
  - Health outcomes in the same village, Errors are correlated because of the same medical and food supply and similar cultural background
  - Earning in the same region might be correlated because of the same industrial structure
  - Analysing workers in firms (earnings, tenure, promotion) will suffer from common firm effects

The Moulton Problem

- Illustration using a simple model with a group structure
- Intuitively, effect of a macro variable on an individual level outcomes
  - Effect of school-type on exam-grades
  - Effect of regional unemployment on individuals' wages
- Model

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$$

  - With $g = 1, \ldots, G$ and $i = 1, \ldots, n$
  - y_{ig} is the outcome for individual I in group g
  - Here x_g varies only at the group level

Data Structure

i	g	y_{ig}	1	x_g
1	1	y_{11}	1	x_1
2	1	y_{21}	1	x_1
⋮	⋮	⋮	⋮	⋮
n_1	1	$y_{n_1 1}$	1	x_1
1	2	y_{12}	1	x_2
2	2	y_{22}	1	x_2
⋮	⋮	⋮	⋮	⋮
n_2	2	$y_{n_2,2}$	1	x_2
⋮	⋮	⋮	⋮	⋮
1	G	y_{1G}	1	x_G
2	G	$y_{2,G}$	1	x_G
⋮	⋮	⋮	⋮	⋮
n_G	G	$y_{n_G,G}$	1	x_G

- Recall $E[e_{ig}] = 0$ & $V[E_{ig}] = \sigma_e^2$
- Recall correlation:

$$\rho_e = \frac{Cov(e_{ig}, e_{ig})}{sd(e_{ig})sd(e_{ig})}$$

- Likely: for individual i and j from the same group g:

$$Cov[e_{ig}, e_{jg}] = \rho_e \sigma_e^2 > 0$$

Additive Random Effects

- Group correlation often modelled using additive random effects, assume $e_{ig} = \nu_g + \eta_{ig}$
- ν_g : group specific error term which captures all the within-group correlation with $E[\nu_g] = 0$ & $V(\nu_g] = \sigma_b^2$
- η_{ig} : individual level specific error term with $E[\eta_{ig}] = 0$ & $V(\eta_{ig}) = \sigma_\eta^2$
- Assuming ν_g and η_{ig} are uncorrelated
- We note that ν_{ig} and η_{jg} are uncorrelated

$$Cov(e_{ig}, e_{jg}) = E[(v_g, n_{ig})(v_g + n_{jg})] = E[v_g^2] = \sigma_v^2$$
$$V[e_{ig}] = E[(v_g + n_{ig})^2] = E(v_g^2 + n_{ig}^2) = \sigma_v^2 + \sigma_n^2$$

Intraclass Correlation Coefficient

- The intraclass correlation coefficient as the proportion of variation in $(v + n)$ due to v :

$$\rho_e = \frac{Cov(e_{ig}, e_{jg})}{sd(e_{ig})sd(e_{jg})} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_n^2}$$

- When the regressor of interest varies only at group level, then this error structure can increase standard errors sharply
- By how much is the conventional variance of the OLS estimate inflated?
- Let $V_c(\hat{\beta}_1)$ denote the conventional OLS variance expression and $V(\hat{\beta}_1)$ be the correct sampling variance with this error structure
- Depending on the data structure. There are various versions to quantify $\frac{V(\hat{\beta}_1)}{V_c[\hat{\beta}_{s1:1}]}$

For the following data structure :

- Nonstochastic regressors fixed at the group level (that is, all regressors are the same for each individual in a group)
- Equal group sizes $N = n_1 = \ldots = n_G$ with total sample size $n = G * N$

$$\frac{V[\hat{\beta}_{s1:1}]}{V_c[\hat{\beta}_{s1:1}]} = 1 + (N-1)\rho_e$$
$$\text{Moulton Factor } \tau = \sqrt{1 + (N-1)\rho_e}$$

Which quantifies how much we over estimate precision by ignoring intraclass correlation

Remark.
- *Conventional standard errors become increasingly misleading as group size N and / or ρ_e increase*
- *If there is no error correlation ($\rho_e = 0$) , there is no overestimation*
- *If $\rho_e = 1$ (or $n_{ig} = 0$), then within a group, all 's are the same : the conventional variance is scaled up by (N-1) since we copy each information N times without generating new information*
- *With the total sample size fixed, increasing the group sizes N just decreases the number of clusters which leads to less independent information*
- *The Moulton factor can be very big even with a small correlation. Assume 100 observations per group and a $\rho_e = 0.1$ leads to a Moulton factor of 3.3. The conventional standard errors are only roughly $\frac{1}{3}$ of what they should be*

Generalisations

The most general form where x varies by g and I with variations in g :

$$\frac{V[\hat{\beta}_{s1:1}]}{V_c[\hat{\beta}_{s1:1}]} = 1 + [\frac{V(N_g)}{\bar{n}_g} + \bar{n}_g - 1]\rho_r\rho_x$$

where rho_x is the within cluster correlation coefficient for x:

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j}(x_{ig} - \bar{x})(x_{jg} - \bar{x})}{V[x_g] \sum_g n_g(n_g - 1)}$$

- ρ_x is a generic measure of the correlation of the regressors within the group. If this correlation is zero, the Moulton effect disappears
- Clustering has a bigger impact on standard errors with variable group sizes and when ρ_x is large
- If the group size is fixed but x varies by g and I, the Moulton factor becomes the square root of $1 + (N - 1)\rho_E\rho_x$

Moulton Problem - Solutions

Model $y = \beta_0 + \beta_1 x_{ig} + e_{ig}$ with $g = 1, \ldots, G$

1. Parametric approach
   - Fix the conventional standard errors using the general formula for the Moulton factor by estimating the intraclass correlations ρ_e and ρ_x
2. Cluster standard errors
   (a) Generalisation of white's robust covariance matrix

$$\hat{AV}(\hat{\beta}_{s1:1}) = (X'X)^{-1}(a \sum_{g=1}^{G} X'_g \hat{e}_g \hat{e}'_g X_g)(X'X)^{-1}$$

   (b) Where $\hat{e}_g$ is a n_g x 1 vector of redials for observations in the g-the cluster and X_g is a n_g x k matrix of regressors for observations in the g-the cluster
   (c) Typically, there is a degrees of freedom adjustment $a = \frac{G(n-1)}{(G-1)(n-k)}$
   (d) Consistent if number of cluster is large but not consistent with fixed number of groups (even when group sizes tend to ∞)
   (e) No assumptions about within-group correlation structure (not just parametric such as in the additive error structure)
   (f) If each individual is his own group (I = g and G = n) then the formula collapses back the robust estimator
3. Use group averages instead of microdata
   (a) Model : $y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$, $g = 1, \ldots, G$
   (b) We estimate $\bar{y}_g = \beta_0 + \beta_1 x_g + \bar{e}_g$ by weighted least squares using n_g as weights
   (c) However, neglecting heteroskedacity unless the group sizes are equal

   (d) Relying on asymptotics for group number, not group sizes
   (e) With modest group sizes, it is expected to have good finite sample properties of regressions with normal errors
   (f) And is likely to be more reliable than clustered standard errors with few clusters
   (g) But does not work if x varies within groups and ignores any other micro-level covariates
   (h) But there exists a 2 step approach to include micro level covariates (A & P)
4. Block bootstrap
   (a) To be discussed
5. GLS or Max Likelihood approaches

   GLS :
   (a) In some cases is possible to estimate GLS or maximum likelihood model
   (b) Requires a model for error structure

Example (Star Experiment). *Krueger (1999) uses IV to estimate the effect of class size on students' achievements y_{ig} is the test score of student I in class g and class size x_g*

Students were randomly assigned to each class but data are unlikely to be independent across observations.

Test scores in the same classes are correlated because students in the same class share background characteristics and are exposed to the same teacher and classroom environment

It is likely for students I and j from the same class g :

$$E[e_{ig}, e_{jg}] = \rho_r \sigma_e^2 > 0$$

The estimation strategy is for now not in our focus, thought we can compare the different standard error estimates

Standard errors for class size effects in the STAR data (318 clusters)

Variance Estimator	Std. Err.
Robust (HC_1)	.090
Parametric Moulton correction (using Moulton intraclass correlation)	.222
Parametric Moulton correction (using Stata intraclass correlation)	.230
Clustered	.232
Block bootstrap	.231
Estimation using group means (weighted by class size)	.226

Notes: The table reports standard errors for the estimates from a regression of kindergartners' average percentile scores on class size using the public use data set from Project STAR. The coefficient on class size is −.62. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

Figure 11: Robust standard errors after correcting for clustering

[Moulton Derivation]

Note. *Lecture : if we estimate a model parameter consistently, why do we care about inference?*

- *We would like to investigate a problem*
- *Policymaker would like to know whether to implement school building program*
- *But what is decision rule? Typically, think about Statistical significance and sufficient magnitude then the policymaker wants to adopt the program, if not then not adoptable.*
- *We need CI or at least a statistical test. For this we need SE and distribution*
- *Need to estimate SE correctly to get correct CI, if we have too large CI (SE wrong), the implication/ error is that we risk not detecting an effect, when there is*
- *But the other way around too small SE (forgot to cluster), might think building schools help and invest a lot of money, but the effect is 0*
- *This is a danger and the problem is incorrect standard errors lead to incorrect confidence intervals*

Note. *Lecture : Once we have accounted for clustering using the Moulton approach compared to the standard errors, is it more likely that the clustered standard errors are larger or smaller than the OLS*

Note.

Larger

Note. *Lecture : what solutions exist to account for clustering*

- *Group averages (only valid for regressors that don't vary within each individual within a group)*
- *Parametric - estimate the Moulton factor*
- *Clustering SE*
- *Block bootstrap*

Lecture 3: Bootstrap and IV

Tue 13 Feb 15:20

[L3]

2.4 Basic Intro to Bootstrap

Bootstrap is an effective alternative to distributional assumptions on T. It is a versatile tool for constructing confidence intervals, particularly useful in linear regression. The process involves re sampling the data with replacement and re-estimating the model, simulating multiple 'made-up' datasets.

The bootstrap is generally reliable except when estimating the minimum or maximum values or under non-Gaussian distributions. It is a good choice for asymptotically normal situations.

Omitted Variable Bias (OvB) creates bias only if there is correlation with regressors. If the variable is non-correlated, it's fine. However, including too many regressors can add noise and include variables correlated with the regressor.

Attenuation Bias occurs with measurement error, which persists as long as variance exists. Reducing measurement error variance helps, provided the error isn't systematic

- Another method for estimating variance, CI and dist on statistic
- Often used when exact distribution is unknown
- Different versions but non parametric most common

Non-parametric Bootstrap

Resampling n observations of (Y_i, W_i) from the data with replacement, then re-estimating the model. After B samples are re-estimated, this will give us a distribution of the parameter and we can describe the statistical properties of this distribution (e.g. The 95 % interval of this distribution)

However, the non-parametric bootstrap can have issues if the sample is small or the regressors are skewed as the additional noise introduced by resampling creates worse distributional approximates. Thus wild bootstrap.

- X is distributed according to some distribution F: $X \sim F$
- $x = (x_1, \ldots, x_n)$ represents an iid sample from this variable
- Suppose we want to estimate the variance and the distribution of a statistic $T_n = g(x_1, \ldots, x_n)$
- Ultimately interested in variance of distribution of this statistic $T_n = g(x_1, \ldots, x_n)$

NP Bootstrap - Variance

- Let V_F denote the variance of T_n where the subscript F indicates that the variance is a function of F

- If we knew F, we could compute the variance
- For example for $T_n = \frac{1}{n}\sum_{i=1}^{n} x_i$,

$$V_F(T_n) = \frac{V(x)}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}$$

  Where $dF(x)$ is the pdf in integral form (2nd term)
- Which is a function of F
- Idea is to estimate $V_F(T_n)$ with $V_{\hat{F}}(T_n)$
- Or, *use a plug in estimator* of the variance
- Since $V_{\hat{F}}(T_n)$ may be difficult to compute, we approximate it with a simulation estimate denoted by v_{boot}

Key Idea

Put the initial sample $x = (x_1, \ldots, x_n)$ into an urn

1. Draw n observations from x with replacement
   - Each observation has the probability of $\frac{1}{n}$ of being drawn
   - Gives each bootstrap sample $x_1^* = (x_{11}^*, \ldots, x_{n_1}^*)$
2. Based on the single bootstrap sample, we estimate (compute bootstrap statistic)

$$T_{n_1}^* = g(x_{11}^*, \ldots, x_{n_1}^*)$$

3. Repeat steps 1 and 2 B times to get $T_{n_1}^*, \ldots, T_{nB}^*$ where :

$$T_{nb}^* = g(x*_{1b}, \ldots, x_{nb}^*) \; for \; b = 1, \ldots, B$$

   Where B is the number of bootstrap replications

$$V_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T_{nb}^* - \frac{1}{B}\sum_{r=1}^{B} T_{nr}^*)$$

   (then take sample analog of variance)

Then by the law of large numbers $v_{boot} \xrightarrow{a} V_{\hat{F}(T_n)}$ as $B \to \infty$ (bootstrap variance tends to variance of stat we were after)

Need to reiterate quite often, in real world we have initial sample from true distribution F which gives us stat T_n in bootstrap world we have bootstrap sample which comes from resampling our initial sample which gives us our bootstrap stat : T_n^*

Imagine in real world, initial sample with 4 observations, giving us stat which is a function of these 4 observations (say the average over these 4 abs). In order to get into bootstrap world, we place sample in urn, we draw b times 4 observations each time with replacement

$$
\begin{array}{rll}
1^{st}\ draw: & x_1^* = \{1,3,1,2\} & \rightarrow g(1,3,1,2) = T_{n1}^* \\
2^{nd}\ draw: & x_2^* = \{1,4,4,4\} & \rightarrow g(1,4,4,4) = T_{n2}^* \\
\ldots & & \\
b^{th}\ draw: & x_b^* = \{2,4,1,1\} & \rightarrow g(2,4,1,1) = T_{nb}^* \\
\ldots & & \\
B^{th}\ draw: & x_B^* = \{1,3,2,4\} & \rightarrow g(1,3,2,4) = T_{nB}^*
\end{array}
$$

Figure 12: Bootstrap World

When we do the 1-st draw we get 1, then since draw with replacement, it could happen we draw this again, second is 3, we also put this back, we do this even further then we got again observation with 1.

Then eventually we get the observation with 2 then we can compute the bootstrap statistic b times to obtain bootstrap samples

Use of the Bootstrap

- The empirical distribution of the B bootstrap samples gives us the approximated distribution / moments of T_n
- EEG standard Errors : $\hat{se} = \sqrt{v_{boot}}$
- Approximate the CDF of T_n. Let $G_n(t) = (T_n < t)$ be the CDF of T_n
- The bootstrap appropriate to G_n is

$$\hat{G}_n^*(t) = \frac{1}{B}\sum_{b=1}^{B} 1_{\{T_{nb}^* \leqslant t\}}$$

  Where the binary variable obtains probability
- Confidence intervals based on SE or quantiles
- Normal interval :

$$T_n \pm z_{\frac{a}{2}}\hat{se}_{boot}$$

- Where $\hat{se}_{boot}$ is the bootstrap estimate for the SE
- Where $z_{\frac{a}{2}}$ is the $\frac{a}{2}$ quantile of the standard normal distribution
- The interval is not accurate unless the distribution of T_n is close to normal

```
> set.seed(1000)
> n = 100
> mu = 5
> sigma2= 2
> X = rnorm(n,mu,sd = sqrt(sigma2))
> #Tn = mean(X)
>
> B = 500
> TnStar = c()
> for (i in 1:B){
+   XStar = sample(X, size = n, replace = TRUE)
+   TnStar[i] = mean(XStar)
+ }
> # Bootstrap Mean:
> mean(TnStar)
[1] 5.031754
> # True mean: 5
>
> # Bootstrap Variance:
> var(TnStar)
[1] 0.02003969
> #True Variance:
> V = sigma2/n
> V
[1] 0.02
```

Figure 13: Boostrap Code

Set seed to ensure RV is same on diff computers, then 100 obs, mean = 5, variance = 2

We want 500 bootstrap replications, then we initiate an empty vector t_n^* to collect bootstrap replications, then iterate over 500 i's. For each I in 1:500 we sample from our initial vector, with replacement 100 observations, giving us bootstrap sample, then we take mean to obtain bootstrap mean

T_n^* has 500 bootstrap means, then we take mean over these 500 and compare to true expectation. 5.03 is very close to the true mean,

We proceed the same to estimate variance based on bootstrap replications, we are also close to variance also.

Practically, it depends on the situation to normalise test stat (demean or standardise in order to ensure normal distribution)

Regression Estimates

Procedure quite similar, but with at least 2 characteristics for each individual parameters

Instead of drawing directly from RV, we Draw pairs of $\{y_i, x_i\}$ to

- Sometimes called the *pairs bootstrap*
- Instead of drawing directly from the random variable, you would sample the indices of the observations

Empirical, non parametric, standard, pairs.

Wild Bootstrap

Relies on Assumption That Error Term at Disposal

- Model $y_i = \beta_0 + \beta_1 x_i + u_i$ (one regressor)
- Preserves heteroskedatic behaviour since don't destroy link between x's and error terms
- Initial sample $z = [(y_1, x_1) \dots (y_{n,x_n})]$ with outcome and regressors for each individual

Methodology

Quite similar but main difference that it is residual bootstrap but keep regressors fixed

1. Estimate $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ for all $i = 1, \ldots, n$
2. Randomly create a bootstrap residual (weights)

$$\text{weights:} \quad w_i = \begin{cases} 1\ , & \text{with probability } \frac{1}{2} \\ -1\ , & \text{with probability } \frac{1}{2} \end{cases}$$

   Bootstrap residuals $\hat{u}_i^* - w_i \hat{u}_i$, for all $i = 1, \ldots, n$
3. Compute the bootstrap dependent variables (essentially changed sign of original residual)

$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

   For all $i = 1, \ldots, n$ Gives : a single bootstrap sample : $z_1^* = [(y_{11}^*, x_1), (y_{n1}^*, x_n)]$

Repeat 1-3 B times to obtain B wild bootstrap samples : $z_b^*[(y_{1b}^*, x_1), \ldots, (y_{nb}^*, x_n)]$ for $b = 1, \ldots, B$

Alternative formulation

1. Estimate the linear model $\widehat{\beta}$ and obtain the residuals $\hat{\varepsilon}_i$
2. In each bootstrap step b:
   (a) For each observation i, the X_i is fixed, along with $\widehat{\beta}$ and $\hat{\varepsilon}_i$. We then draw a binary variable $U_{i,b}$ that is either 1 or -1 with equal probability. We set $Y_{i,b} = \widehat{\beta} X_i + U_{i,b} \hat{\varepsilon}_i$
   (b) With the new dataset, we re-estimate the model and construct a t-statistic $t_b^1 = \left(\widehat{\beta}_b - \widehat{\beta}\right) / \sqrt{\hat{V}_b}$ where $\sqrt{\hat{V}_b}$ is the standard
3. With the full set of t_b, we can construct a confidence interval by calculating the $q_{0.95}(|t_b|)$, the 0.95 quantile of $|t_b|$, and using it in place of our usual critical value:

$$CI_{WILD} = \left(\widehat{\beta} - q_{0.95}\left(|t_b|\sqrt{\hat{V}}\right), \widehat{\beta} + q_{0.95}\left(|t_b|\right)\sqrt{\hat{V}}\right)$$

Clustered Bootstrap

Under similar randomisation, you do not preserve the dependence (unobserved factors relating to micro-data) structure in the data

To fix this, we draw blocks of data defined by the groups g. Say block bootstrap by re sampling entire classes instead of individual students, to keep structure of correlation intact.

Can also have cluster 1 bootstrap, maybe you have stratified sampling such that while you sampled you made sure have say gender quota or certain subset, we would need to do bootstrap for this.

The way you sample data structure, try to mimic through the bootstrap exactly this structure. That is, replicate DGP as close as you can (provided we know about it)

Exercise 2. *Algorithm to obtain SE using clustered bootstrap Set the seed If we have 1 village, should we have randomly different households in village 1? No Village 1, 2, 3 and households a, b, c (1), def, (2), ghi (3). Everything within the villages is kept fixed We then draw 1 (abc) , 3(ghi) , 3 (ghi) for 1-st bootstrap Then 2, (ghi) , 1 (abc) , 1 (abc) for 2-nd bootstrap Need large sample. Based on practice not*

theory Advantages - as opposed to random draw, forget about cluster, end up doing randomly a,d,f,c and estimate OLS $\hat{\beta}_1$ and g,h,c,f and estimate $\hat{\beta}_2$, you have no attachment to group and lose correlation within each group Need to re merge together 'blocks' into single data set since we have individually sampled blocks Have to store estimator so on

Exercise 3. *Effect of schooling on wages, use father educ as instrument for years of educ Conditions of good instrument? Exclusion restriction (orthogonality - instrument cannot be correlated with error term), Relevance restriction ($Cov(x_1, z) = 0$) Maybe since there is push to education, maybe with time this effect is fading, but likely still relevant, but maybe in other countries this is deterministic and is something we can test Exclusion restriction - 1. We can control for this, if this is not part of the model. 2. Might be violated if we can argue ability for singers, parents can sing, inherit singing talent so opera hires, this might be correlated with number of years taking singing lessons but choosing to take singing lessons due to natural talent suggests violation of exclusion criteria*

Exercise 4. *Is month of birth good instrument for years of education Exclusion Restriction - it is pretty random when you are born, there is no reason to believe the error term left over when explaining wages is correlated with when you are born Some spikes of the births (seasonality etc) though, could this be an issue? Relevance condition - Structure of Education System : cutoff for year of schooling in the year, therefore matters for when say leave school at 16 Usa : start in September, able to leave when 16, people born earlier get extra months of schooling Does extra month of education have strong effect? No F stat - very low in this case, but we wont a large F-stat If we have a small f stat we have very low relevance, but also $\beta = cov(y, x)/cov(x, z)$ Bias end up having depends on the covariance between y and u and x and z If low relevance, then $\frac{cov(z,u)}{cov(x,z)}$ 'explodes' To test, there is no proper overall applicable test for exclusion restriction, it is something you have to argue for.*

3 Instrumental Variables

[L3-IV]

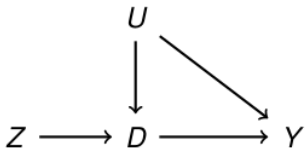
Motivation

Let's say we are interested in identifying the causal effect of years of schooling on wage, we estimate the model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

But in many examples, we've worried about estimating the effect of some treatment D_i n Y_i, but concerned that this estimate will be biased. When we have strong ignore ability (say due to randomisation), we felt confident that this was not a concern.

But what about when this isn't the case, we need to create 'as-if' random variation in D_i

What is an instrumental variable?



We have a variable Z which can identify two effects:

1. Z on D
2. Z on Y

What is the content of this instrumental variable?

1. It affects Y (**relevance**)
2. It only affects Y through D (**exclusion**)

Without further assumptions, it won't be possible to identify the effect of D on Y using this, but it highlights the features of an IV

Example. *Canonical Setup With an IV:*

$$Y_i = D_i\beta + W_i\gamma_1 + \varepsilon$$
$$D_i = Z_i\pi + W_i\gamma_2 + u_i$$

with W_i a set of exogenous controls

There are a couple of notable features about this setup:

- *We've assumed a very parametric model for Y_i*
- *In particular, we've assumed a constant effect of D_i on Y_i*

The necessary assumptions to identify D_i in this setting are straightforward

1. *Relevance $\pi \neq 0$ ($E[D_iZ_i]$)*
2. *Exclusion $E[\varepsilon_iZ_i|Wi] = 0$ ($E[Z_i\varepsilon_i]$)*

However, the exclusion restriction can be slightly opaque, ε_i captures the set of "other" things that can happen. But it can be harder to map into a counterfactual way of discussing outcomes

Useful result Let $Z*_i = Z_i - E[Z_i|W_i]$, then the exclusion restriction can be viewed as saying that $E[\varepsilon_iZ*_i] = 0$ - the variation in Z_i above and beyond W_i has to be exogenous for ε_i However, there are two issues with this setup

We have assumed homogeneous effects e.g. β is the same for all individuals - though we can fix this in the model, we want to have a clear idea on what we want to estimate

- One of the key assumptions for unbiasedness is the *homogeneity of regressors* : $E[u|x_1, \ldots, x_k] = 0$
- Indeed, the problem arises when the regression error is correlated with a regressor : i.e. $E[u|x_k] \neq 0$
- There are three broad reasons for *Endogeneity* :
  1. Omitted variable bias
  2. Measurement error
  3. Simultaneous equations
- In our example, x_k is said to be endogenous, meaning the years of schooling might be correlated with innate and unobserved ability
- The OLS estimator β_k is *biased* and *inconsistent*
- One approach to deal with this issue is to use instrumental variables

3.1 Forms of Endogeneity

- Omitted variables
  1. Arises in cases when one fails to control for a regressor that is correlated with other regressors
  2. Often due to self selection : if an agent chooses the value of the regressor, this might depend on factors that we cannot observe
  3. That is, *unobserved heterogeneity*
- Measurement Error
  1. Occurs when we can only observe an imperfect measure of a variable
  2. Depending on how the observed and true variable are related, we might have endogeneity
- Simultaneity
  1. Occurs when dependent and independent variables are simultaneously determined
  2. If x is partially determined by y, then the error might be correlated with x

Though this is not to say there exist sharp distinctions

Example. *Effect of alcohol consumption on worker productivity (measured by wages)*

Alcohol usage correlated wuth unobserved factors such as family background, which may also have an effect on wage. Leading to an Omitted variable problem

Alcohol demand can depend on income, leading to the simultaneity problem

There also exists possibility of mismeasurement of alcohol consumption

Omitted Variable Bias

- Long regression - true model is : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
- Then, assuming we cannot observe x_2 but only x_1
- Or, in short : $y = \delta_0 + \delta_1 x_1 + u$ with $u = \beta_2 x_2 + e$
- We know the population parameter can be expressed as :

$$\delta_1 = \frac{cov(y, x_1)}{v[x_1]}$$

$$\begin{aligned} &\text{replacing y from the true model:} \\ &= \frac{Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + e, x_1)}{V[x_1]} \\ &= \beta_1 v[x_1] + \beta_2 Cov(x_2, x_1) + cov(e, x_1)/V[x_1] \\ &= \beta_1 + \beta_2 \frac{cov(x_2, x_1)}{v(x_1)} \end{aligned}$$

Defining τ_1 as the parameter in the population model that relates x_1 to x_2 :

$$X_2 = \tau_0 + \tau_1 x_1 +' error'$$

We therefore estimate $\delta_1 = \beta_1 + \beta_2\tau_1$

However, our OLS estimate is *inconsistent* (asymptotically biased)

$$\operatorname*{plim}_{n\to\infty} \hat{\delta_1} - \beta_1 = \beta_2 \frac{Cov(x_1, x_2)}{v(x_1)} = \beta\tau_1$$

Where we can show

$$Bias(\hat{\delta_1}) = E[\hat{\delta_1}] - \beta_1 = \beta_2\hat{\tau_1}$$

Where thinking about the direction of the correlation helps us think about the direction of the bias

Essentially, if the omitted variable is related to the included regressor, then the parameter in the short regression will not identify the parameter in the long regression.

With more regressors, the formula changes but the principle remains the same

Example. *omitted variable bias*

Let y be the wages, x_1 years of education and x_2 ability

Regressing wages on years of education alone delivers a biased estimate $\hat{\delta_1}$

We would expect both years of education and ability to have a positive impact on average earnings (that is $\{\beta_{1/2} > 0, \}$)

But we also expect both regressors to be positively correlated as individuals with more innate ability tend to choose / acquire more education ($\tau_1 > 0$)

Therefore, $\hat{\tau_1}$ likely overestimates the value of education, since in our education regressor we have not controlled for the correlation with ability and thus include more than the effect of education in this estimate, here thinking about the direction of the correlation has helped us to identify the sign of the bias

Measurement Error in y

Situation 1 : Measurement error in the dependent variable (y)

True model $y = \beta_0 + \beta_1 x_1 + \varepsilon$, $E[\varepsilon|x_1] = 0$

We can only observe $\tilde{y}$ which measures the unobserved y with an error $\tilde{y} = y + e$

We regress $\tilde{y} = \tilde{\beta_0} + \tilde{\beta_1} x_1 + \tilde{\varepsilon}$

$$\tilde{\beta_1} = \frac{Cov(\tilde{y}, x_1)}{V[x_1]} = \frac{Cov(y + e, x_1)}{V[x_1]} = \beta_1 + \frac{Cov(e, x_1)}{V(x_1)}$$
$$\tilde{\beta_0} = E(\tilde{y}) - \beta_1 E(x_1) = E[y] + E[e] - \beta_1 E[x_1] = \beta_0 + E[e]$$

However, this can cause bias and inconsistency. Although it vanishes if the measurement error is statistically independent of each explanatory variable. We note the usual OLS inference procedures are asymptotically valid.

Situation 2 : Measurement error in the regressor (x)

True model $y = \beta_0 + \beta_1 x_1 + \varepsilon$, $E(\varepsilon|x_1) = 0$

Where we can only observe $\tilde{x_1}$, a measure of the unobserved x_1 with an error $\tilde{x_1} = x_1 + \varepsilon$

We then regress $y = \tilde{\beta_0} + \tilde{\beta_1}\tilde{x_1} + \tilde{\varepsilon}$

$$\tilde{\beta_1} = \frac{Cov(y, \tilde{x_1})}{V[x_1]} = \frac{Cov(\beta_0 + \beta_1 x_1 + e, x_1 + e)}{V[\tilde{x_1}]} = \beta_1 \frac{V(x_1)}{V(\tilde{x_1})}$$
$$\tilde{\beta_0} = E(\tilde{y}) - \beta_1 E(x_1) = E[y] + E[e] - \beta_1 E[x_1] = \beta_1 \frac{V(x_1)}{V(x_1) + V(e)} = \beta_1 \lambda$$

With the key assumptions that $\text{Cov}(\mathbf{e}, \mathbf{x_1}) = \mathbf{0}, \text{Cov}(\mathbf{e}, \varepsilon) = \mathbf{0}$, and $\mathbf{E}[\mathbf{e}] = \mathbf{0}$

In which we can show

$$plim_{n\to\infty}\hat{\tilde{\beta}}_1 = \beta_1 \lambda$$

where $\lambda \in \{0, 1\}$: $\hat{\tilde{\beta}}_1$ underestimates β_1, this is attenuation bias.

Though as $V[e]$ shrinks relative to $V(x_1)$, the attenuation bias disappears.

In the general model, it is not the variance of the true regressor that affects the consistency but the variance in the true regressor after netting out the other explanatory variables

Simultaneity / Reverse Causality

Problem : y_1 and y_2 are simultaneously determined.

$$Y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1, E[z_1|u_1] = 0$$
$$Y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2, E[z_2|u_2] = 0$$

A classic example is when y_1 is price and y_2 is the quantity and both equations are demand and supply. But note the intercept is suppressed for simplicity.

Focusing on the 1-st equation, to show that $Cov(y_2, u_1) \neq 0$:

$$Y_2 = \alpha_2[\alpha_1 y_2 + \beta_1 z_1 + u_1] + \beta_2 z_2 + u_2$$
$$= \frac{\beta_1 \alpha_2}{1 - \alpha_1 \alpha_2} z_2 + \frac{\alpha_2}{1 - \alpha_1 \alpha_2} u_1 + \frac{1}{1 - \alpha_1 \alpha_2} u_2$$

Assuming that $\alpha_1, \alpha_2 \neq 0$, $Cov(u_1, u_2) = 0$ and $cov(z_2, u_1) = cov(z_2, u_2) = 0$, thus violating exogeneity

$$Cov(y_2, u_1) = \frac{\alpha_2}{1 - \alpha_1 \alpha_2} V(u_1 \neq 0)$$

Although, without additional controls (z_1), it can be shown that the consistency has the same sign as $\frac{\alpha_2}{1-\alpha_1\alpha_2}$

3.2 IV Estimator

Potential outcomes and mapping it to IV:

Imbens And Angrist (1994) Simplest of cases : binary instrument Z with binary treatment D and no controls.

Extend the potential outcomes framework to allow an instrument Define $Y_i(D_i(Z_i), Z_i)$ and $D_i(Z_i)$ as two forms of potential outcomes.

The exclusion restriction here is that $Y_i(D_i(Z_i), Z_i) = Y_i(D_i(Z_i))$, e.g. Z_i only has an effect on Y_i through D_i And the relevance is that $P(w) = E[D_i|Z_i = w]$ varies across w

The key point is the $Y_i(1) - Y_i(0)$ can be different for every individual, unlike in the structural models previously where β was constant. Further we assume that Z is completely randomly assigned relative to the potential outcomes of Y and D.

Consider $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$ where $P(1) > P(0)$

$$\begin{aligned} &E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \\ &= E(D_i(1)Y_i(1) + (1 - D_i(1))Y_i(0)|Z_i = 1) - E(D_i(0)Y_i(1) + (1 - D_i(0))Y_i(0)|Z_i = 0) \end{aligned}$$

$$= E((D_i(1) - D_i(0))(Y_i(1) -$$

$$= Pr(D_i(1) - D_i(0) = 1) \times E(Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1) - Pr(D_i(1) - D_i(0) = -1) \times E(Y_i(1) - Y_i(0)|D_i(1) - D_i(0)$$

While we assumed that the propensity score was increasing, it does not imply that it's increasing for everyone. And second, we are only identifying the effect of $D(Y_i(1) - Y_i(0))$ for those individuals who's behaviour shifted due to the change in Z. Third, without restrictions on Y_i, this effect can be zero or even negative, even if the true causal effect is positive → those shifted into participating by Z could be exactly cancelled by those who shift out

Two solutions:

1. In a constant effects world, this problem does not exist
2. If there exists an instrument such that $\mathrm{P}(D_i(1) - D_i(0) = -1) = 0$, then this is okay (one sided compliance)

Key innovation : with Monotonicity, we can identify the local average treatment effect

Assumption 1. *Monotonicity*

$$D_i(1) \geqslant D_i(0) \quad \textit{for all } i \textit{ (or vice versa)}$$

All effects must be monotone in the same direction, however this is fundamentally untestable (suffering from the fundamental problem of causal inference)

Conditional on assuming monotonicity, the Wald ration estimates the LATE:

$$\begin{aligned} \tau_{LATE} &= \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} \\ &= \frac{Pr(D_i(1) - D_i(0) = 1) \times E(Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} \\ &= E(Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1) \end{aligned}$$

Thus, meaning that an IV strategy only identifies (non-parametrically) the effect of a treatment for those who respond to the treatment (sine monotonicity ensures that the responders go in one direction)

Single Regressor Model

Properties of an Instrument to Overcome Endogeneity

Model : $y = \beta_0 + \beta_1 x_1 + u$ with $E[u|x_1] \neq 0$

We need an instrumental variable (IV) z with the following properties

- **Exclusion Restriction** : $Cov(u, z) = 0$
- **Relevance** : $Cov(x_1, z) \neq 0$

The instrument **cannot be correlated** with any of the omitted variables for egg, but it does need to be correlated with the endogenous regressor

1. We can test the relevance assumption
2. However we cannot generally test the exclusion restriction. One needs to carefully apply common sense and economic theory to convince the audience about the validity of the instrument

Testing for Relevance

First stage regression

$$X_1 = \pi_0 + \pi_1 z + v$$

Where we regress endogenous regressor on instrument, π_1 is equal to covariance of instrument and endow regressor / variance of endow regressor

Recall, $\pi_1 = \frac{Cov(z,x_1)}{V(x_1)}$

Test whether covariance is zero or not:

$$H_0 = \pi_1 = 0 \;\; vs \;\; H_0 :\neq 0$$

Where we should be able to reject at a small significance level and be confident that the relevance condition holds

Example. *number of siblings as an instrument for education in log wage equation we test the relevance condition by the first stage regression*

$$educ \; = \delta_0 + \delta_1 \; sibs \; + \nu$$

```
Call:
lm(formula = educ ~ sibs, data = dat)

Residuals:
Min      1Q Median      3Q     Max
-5.139 -1.683 -0.683  1.931  6.140

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.13879    0.11314 124.969  < 2e-16 ***
sibs        -0.22792    0.03028  -7.528 1.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.134 on 933 degrees of freedom
Multiple R-squared:  0.05726,Adjusted R-squared:  0.05625
F-statistic: 56.67 on 1 and 933 DF,  p-value: 1.215e-13
```

Figure 14: Regression Output

Where the effect of sibs on education is $\hat{\delta}_1 = -0.228$ *and significantly different from zero with a p value* < 0.01 *and a t-stat of -7.528. Thus, one sibling more decreases the years of education by 0.22792 years and thus the relevance condition seems to hold*

How to find IVs?

- It might be hard to think of a valid instrument and or to have data on them
- Instruments come from
  - Economic theory
  - Exogenous sources of variation in the endogenous regressor arising from a random phenomenon such as whether events, or exogenous policies (cutting class sizes to increase grades)

Example. *Wages Where ability causes the regressor of years of educating to be endogenous One could think of*

- *Family background variables*
- *Proximity to school / college*
- *Month of birth*

As potential Ifs. Whether these would work requires scrutiny

Identification

Model : $y = \beta_0 + \beta_1 x_1 + u$, with $E(u|x_1) \neq 0$ Instrument : $Cov(x_1, z) \neq 0$ and $Cov(z, u) = 0$ (satisfying exclusion and relevance)

Identification in this context : we can write β_1 (parameter of interest) in terms of population moments that can be estimated

We write β_1 in terms of population covariances

$$\text{Cov}(z, y) = \beta_1 \text{Cov}(z, x_1) + \text{Cov}(z, u) \tag{2}$$

$$\textit{Since } \text{Cov}(z, u) = 0 \textit{ and } \text{Cov}(z, x_1) \neq 0 \text{ by assumption} \tag{3}$$

$$\beta_1 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x_1)} \tag{4}$$

However, this fails if $\text{Cov}(\mathbf{z}, \mathbf{x_1}) = \mathbf{0}$, that is *the relevance condition doesn't hold.* This is an expression we can estimate using a random sample.

IV Estimator

- Given random sampling, we estimate the moments by the sample analogs :

$$\widehat{\beta}_1^{IV} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_{i,1} - \bar{x}_1)}, \text{ and } \widehat{\beta}_0^{IV} = \bar{y} - \widehat{\beta}_1^{IV} \bar{x}_1$$

- When $z = x_1$ then the IV estimator reduces to the OLS estimator
- Using the Law of Large Numbers, we can show that $\widehat{\beta}_1^{IV}$ is consistent under the assumptions :

$$\text{plim}_{n \to \infty} \widehat{\beta}_1^{IV} = \beta_1$$

- However, the IV estimator is biased
- Requiring large samples

Identification We note that if we divide the denominator and numerator of eq. (4) by $V[z]$:

$$\beta_1 = \frac{\text{Cov}(z, y)/V(z)}{\text{Cov}(z, x_1)/V(z)}$$

Where β_1 is the ratio of the population regression of the reduced form over the first stage.

Wald Estimator - Binary Instrument

Recalling that for a regression on a binary variable, the resulting slope estimate is the difference between both groups averages. Then, under the IV assumptions, the β_1 can be represented as the ratio of the two OLS estimands, in case of a binary IV:

$$\beta_1 = \frac{\text{Cov}(y, z)V(z)}{\text{Cov}(x_1, z)/V(z)} = \frac{E[y|z = 1]] - E[y|z = 0]]}{E[x_1|z = 1]] - E[x_1|z = 0]]}$$

Then, taking the sample analog : $\widehat{\beta}_1^{IV} = \frac{\bar{y_1} - \bar{y_0}}{\bar{x}_{1,1} - \bar{x}_{1,0}}$ Where $\bar{y_1}$ and $\bar{x}_{1,I}$

Lecture 4: IV continued

Tue 20 Feb 16:07

[L4-IV]

Recap

- Causal question of interest: effect of years of schooling on wage
- Model : $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$
- One of the key assumptions for unbiasedness is the exogeneity of regressors $E[u|x_1 \ldots x_k] = 0$
- Intuitively, the problem arises when the regression error is correlated with a regressor, e.g. $E[u|x_k] \neq 0$
- There are three board reasons for endogeneity : omitted variable bias, measurement error and simultaneous equations
- The consequences are:
  - x_k is said to be endogenous. In our example the years of schooling might be correlated with innate and unobserved ability
  - OLS estimator β_k is biased and inconsistent
- One approach is to use IV

3.3 2SLS

One of the more intuitive instrumental variable estimators is the 2SLS

Example. *example to illustrate explaining some of the IV intuition*

Suppose we have a sample of data on Y, S and Z. For each observation we issue the data is generated according to:

$$Y_i = \alpha + \delta S_i + \varepsilon_i$$
$$S_i = \gamma + \beta Z_i + \varepsilon_i$$

Where $C(Z, \varepsilon) = 0$ *(exclusion restriction) and* $\beta \neq 0$ *(non-zero first-stage)*

We can write out the IV estimator as (using $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$)

$$\begin{aligned}\hat{\delta} &= \frac{C(Y,Z)}{C(S,Z)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(S_i - \bar{S})} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})Y_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})S_i}\end{aligned}$$

Then substituting the true model for Y, we get

$$\begin{aligned}\hat{\delta} &= \frac{1}{n}\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})\{\alpha + \delta S + \epsilon\}}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})S_i} \\ &= \delta + \frac{1}{n}\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})S_i} \\ &= \delta + \text{“small if } n \text{ is large”}\end{aligned}$$

Returning to our first description of $\hat{\delta}$ as the ratio of the two covariances. With some simple algebraic manipulation we get:

$$\begin{aligned}\hat{\delta} &= \frac{C(Y,Z)}{C(S,Z)}\frac{C(Z,Y)}{V(Z)} \\ &= \frac{C(Y,Z)}{C(S,Z)}\frac{1}{C(Z,S)}V(Z)\end{aligned}$$

Where the denominator is equal to $\widehat{\beta}$, we can rewrite $\widehat{\beta}$ as

$$\begin{aligned}\hat{\beta} &= \frac{C(Z,S)}{V(Z)} \\ \hat{\beta}V[Z] &= C(Z,S)\end{aligned}$$

Then we rewrite the IV estimator and make a substitution:

$$\begin{aligned}\hat{\delta}_{IV} &= \frac{C(Z,Y)}{C(Z,S)} \\ &= \frac{\hat{\beta}C(Z,Y)}{\hat{\beta}C(Z,S)} \\ &= \frac{\hat{\beta}C(Z,Y)}{\hat{\beta}^2V(Z)} \\ &= \frac{C(\hat{\beta}Z,Y)}{V(\hat{\beta}Z)}\end{aligned}$$

Notice now that the fitted values from the first stage are inside the parenthesis ($\widehat{\beta}Z$)

2SLS Estimation of β_0 Involves replacing $u_{y2|z}$ by its predicted value based on an estimated version of

$$y_{2,i} = z_1'\delta_{1,0} + z_{2,i}'\delta_{2,0} + u_{2,i} \tag{5}$$

Two step estimating procedure:

1. Stage 1: estimate eq. (5) via OLS to obtain the predicted value of $y_{2,i}, \hat{y}_{2,i}$
2. Stage 2: estimate the model
$$Y_{1,i} = z_{1,i}'\gamma_0 + \alpha_0\hat{y}_{2,i} + \text{ "error"}$$
via OLS to obtain $\widehat{\beta}_{2SLS}$ which is known as the 2SLS estimator of β_0

Formally: IV

Model $y = x\beta + u$ with x a vector of k exogenous and endogenous regressors and z a vector of m IV's (including the exogenous variable)

1. M = k : the model is just identified ,we have an instrument for each endogenous variable $\Rightarrow$ use IV
2. M < k : the model is not identified, we do not have enough IVs
3. M > k : the model is over-identified $\rightarrow$ we have too many IVs. Use GIVE / 2SLS

Case : Length(z) = Length(x)

Model : $y = x\beta + u$, $x = (q, x_2, \ldots, x_k)$ and $z = (1, x_2, \ldots, x_{k-1}, z_1)$. We know $Cov(x_j, u) = 0$ for $j = 2, \ldots, k-1$ and $Cov(x_k, u \neq 0)$

We have an instrument for x_k :

- The instrument must be uncorrelated with the error term (Exogenous) i.e. $Cov(z_1, u) = 0$
- The instrument must be partially correlated with the endogenous variable x_k (Partial Correlation) meaning $\theta_1 \neq 0$ in the equation $x_k = \delta_1 + \delta_2 x_2 + \ldots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k$

The moment conditions for IV estimation imply :

$$E[z'u] = E[z'(y - x\beta)] = 0$$

That the instruments z are orthogonal to the error term u $z \perp u$

We have one instrument at our disposal for this endogenous regressor, we include the constant and all the exogenous regressors because they can be used for instruments for themselves.

Partial correlation best seen by regressing endogenous regressor x_k on all exogenous variables plus the instrument for x_k and we need the parameter on the instrument θ_1 not to be 0, thus partial correlation, the correlation cannot be 0 after the other effects have been 'netted' out. Different to simple case where sufficient to have covariance between endogenous regressor and instrument $\neq 0$

Exogeneity leads to above expression, plugging in expression for u.

Formally: IV

Multiplying the model : $y = x\beta + u$, through with z', taking expectation and using the moment condition :

$$\begin{aligned} E[z'y] &= E[z'x]\beta \\ \text{if rank } \; E[z'x] &= k \\ \beta &= [E[z'x]]^{-1} E[z'y] \end{aligned}$$

There is a *unique solution only under full rank* since it implies $E[z'x]$ is invertible. We solve for β be rearranging to $\beta = [E[z'x]]^{-1} E[z'y]$

And it can be shown that if we rule out perfect collinearity in z, full rank holds iff $\theta_1 \neq 0$ *Estimation* Given a random sample, we can estimate consistently :

$$\hat{\beta}^{IV} = \left(\frac{1}{n} \sum_{i=1}^{n} z_i' x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} z_i' y_i \right) = (Z'X)^{-1} Z'Y$$

Where Z and X are $n \times k$ data matrices and y is $N \times 1$

Given the assumptions, this estimator is consistent

Case: Length(z) > Length(x):

The idea is to used the fitted values from the first stage regression of the endogenous regressor on all the exogenous variables (including the instruments) and use them as "instruments" in the IV estimator

$$Z = (1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_l) - m = k + I \quad \text{vector for} \quad x_k$$

1. Fitted values from the first stage $\hat{x_i} = (1, x_1, \ldots, x_{k-1}, \hat{x}_k)$

$$\hat{x}_{ik} = \hat{\delta_0} + \hat{\delta_1} x_{i1} + \ldots + \hat{\delta}_{k-1} x_{i,k-1} + \hat{\theta}_1 z_{i1} + \ldots + \hat{\theta}_I z_{il}$$

$$\hat{x_i} = z_i \left(\sum z_i' z_i \right)^{-1} z_i' x_i$$

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

For this endogenous regressor you have several potential instruments at your disposal, you would then regress on exogenous variables from initial model and instruments. That gives you a vector of instruments that is equal to (including all potential instruments).

Then you start with obtaining fitted values from First Stage (FS) regressing x_k on exogenous regressors δ and instruments z

Using the fitted values as instruments :

$$\hat{\beta}^{IV} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{x_i}' x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{x_i} y_i \right) = (\hat{X}'X)^{-1} \hat{X}'Y$$

Then, using calculus, we can show that $\hat{X}'X = \hat{X}'\hat{X}$ and hence

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

Which is the GIVE / 2SLS estimator since it equals the OLS estimator on the fitted values from the first stage

Where we have simply replaced with fitted values, and then replaced in matrix form. We can essentially show this is the OLS estimator on the FS using the fitted values. Then we use this to plug in the IV estimator to obtain the OLS estimator on the fitted values

And the 2SLS estimator can also be represented as

$$\hat{\beta}^{IV} = \left[\left(\sum_{i=1}^{n} x_i' z_i \right) \left(\sum_{i=1}^{n} z_i' z_i \right)^{-1} \left(\sum_{i=1}^{n} z_i' x_i \right) \right]^{-1} \left(\sum_{i=1}^{n} x_i' z_i \right) \left(\sum_{i=1}^{n} z_i' z_i \right)^{-1} \left(\sum_{i=1}^{n} z_i' y_i \right)$$

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

$$\widehat{\beta}^{IV} = \left[(X'Z)(Z'Z)^{-1} Z'X \right] (X'Z)(Z'Z)^{-1} Z'Y$$

β Can be Obtained By:

1. First Stage: Obtain the fitted values $\hat{x}_k$ from the regression x_k on $1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_l$
2. Second Stage: Run the OLS regression: y on $1 + x_1, \ldots, x_{k-1} + \hat{x_k}$

In practice

- However, omitting the exogenous regressors in the first stage is easily done and will lead to inconsistency
- And, SE obtained from the second step are incorrect

Testing for rank condition : $H_0 : \theta_1 = \ldots = \theta_l = 0$ vs at least one θ_s for $s - 1, \ldots, l$ is non zero.

3.4 Properties

Here x is 1 x k and generally includes unity, several elements of x may be endogenous, while z includes any exogenous variable

Assumption 1. *2SLS.1*
For some 1 x m-vector z, $E[z'u] = 0$

Assumption 2. *2SLS.2*

1. rank $E[z'z] = m$
2. rank $E[z'x] = k$

Under the above 2 assumptions, the 2SLS estimator obtained from a random sample is *consistent* for β

Proof

$$\hat{\beta}^{IV} = \beta + \left[\frac{1}{n}\left(\sum_{i=1}^{n} x_i' z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i' x_i\right)\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i' z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i' u_i\right)$$

Asymptotic Normality

Assumption 3. *2SLS.3* $E\left[u^2 z'z\right] = \sigma^2 E[z'z]$ *where* $\sigma^2 = E\left[u^2\right]$ *for which it is sufficient to assume* $E\left[u^2|z\right] = \sigma^2$

But note, when x contains endogenous elements it makes no sense to make assumptions about $V[u|x]$

Assumption 4. *Asymptotic Normality of 2SLS Under assumptions 1-3,* $\sqrt{n}(\widehat{\beta} - \beta)$ *is asymptotically normally distributed with mean zero and variance matrix*

$$\sigma^2\left(\left[E(x'z)\right]\left[E(z'z)\right]^{-1}\left[E(z'x)\right]\right)^{-1} = \sigma^2\left[E(x^*x^*)\right]^{-1}$$

Where $x* = z\prod$ *is the 1 x k vector of linear projections*

Inference To predict the residuals $\hat{u}_i = y_i - x_i\widehat{\beta}^{IV}$ Then estimate $\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i$

Under homoskedacity :

$$AV\left(\widehat{\beta}^{IV}\right) = \widehat{\sigma}^2\left(\hat{X}'\hat{X}\right)^{-1}$$

Under heteroscedasticity and Assumption 4 and Assumption 3

$$AV\left(\widehat{\beta}^{IV}\right) = \left(\hat{X}'\hat{X}\right)^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2\hat{x}_i\hat{x}_i\right)\left(\hat{X}'\hat{X}\right)^{-1}$$

Sometimes, the heteroscedasticity-robust estimator is multiplied by $\left(\frac{n}{n-k}\right)$ to adjust for degrees of freedom. But standard errors and t-stats are still constructed the same way.

3.5 Weak Instruments

In the single regressor / single instrument setting: $y = \beta_0 + \beta_1 x_1 + u$ with instrument z_1

It can be shown that

$$\text{plim}\,\widehat{\beta}_1^{IV} = \beta_1 + \frac{\text{Cov}(z_1, u)}{\text{Cov}(z_1, x_1)} = \beta_1 + \frac{\text{Cov}(z_1, u)}{\text{Cor}(z_1, x_1)} \frac{\sigma_u}{\sigma_x}$$

Where σ_u and σ_x are the standard deviations of u and x_1 in the population respectively

Recall that one of the key assumptions for our estimates was relevance $\pi_1 \neq 0$ or $\text{Cov}(Z_i, D_i|W_i) \neq 0$

However, consider the 2SLS estimator for β_{IV} when W_i just includes a constant :

$$\widehat{\beta} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

If $\text{Cor}(z_1, x_1)$ is small:

- There is no issue if the exclusion restriction holds
- However, even a slight violation of the exclusion restriction may lead to a large bias if the instrument is weak
- The implication being that it might better to use OLS than IV when there is low relevance
- But is an active area of research

Note

$$AV\left(\widehat{\beta}_1^{IV}\right) = \frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2}$$

With σ_x^2 the population variance of x and $\rho_{x,z}$ the population correlation between z_1 and x_1

Furthermore

$$\hat{AV}\left(\widehat{\beta}_1^{IV}\right) = \frac{\widehat{\sigma}^2}{SST_x R_{x,z}^2}$$

Where $R_{x,z}^2$ is the r-squared from the first stage (but be warned this leads to large asymptotic standard errors

Example. *If* $\text{Cov}(D_i, Z_i) = 0$ *Then the 2SLS estimator is obviously undefined, what what about if its very small? - small variations in it will move around* $\widehat{\beta}$ *in a large way - thus statistical uncertainty.*

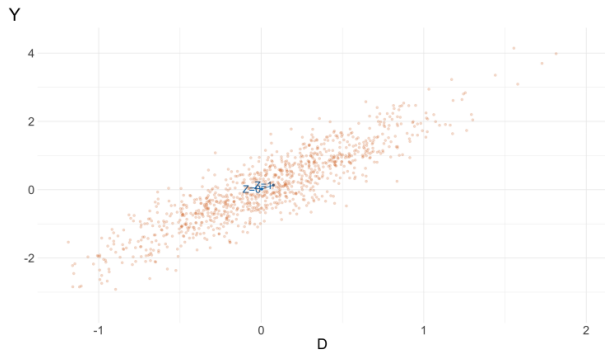


Figure 15

With a simple 2SLS simulation, with a binary instrument. The first stage coif = 0.1, true beta = 2 The estimation on the x-axis comes from variation in the first stage - the larger, the stronger the first stage

However, if the first stage is weak, this interval is quite short even if the variation in D stays the same.

Compared to 0.5, a first stage coefficient of 0.1 means its harder to distinguish the points (overall variance of D is fixed here to keep the correct comparison)

Given that the model is correctly specified, with enough data it should converge to the right β, but small shifts in the x-axis will massively swing the estimate

For simplicity, assume the following variables are demeaned (mean zero) and there are no additional controls (e.g. No constant). Hence,

$$\begin{aligned} Y_i &= D_i\beta + \varepsilon_i \\ D_i &= Z_i\pi + u_i \\ \rightarrow Y_i &= Z_i \underbrace{\pi\beta}_{\delta} + u_i\beta + \varepsilon \end{aligned}$$

The 2SLS estimator (for single endogenous variable) can then be written as:

$$\widehat{\beta}_{2SLS} = \frac{D'Z\left(Z'Z\right)^{-1}Z'Y}{D'Z\left(Z'Z\right)^{-1}Z'D} = \frac{D'P_ZP_Z'Y}{D'P_ZP_Z'D} = \frac{\hat{D}'\hat{Y}}{\hat{D}'\hat{D}} = \frac{\hat{\pi}\hat{Q}\hat{\delta}}{\hat{\pi}'\hat{Q}\hat{\pi}} = \underbrace{\frac{\hat{\delta}}{\hat{\pi}}}_{\text{single instrument}} \qquad \hat{Q} = Z'Z$$

Where intuitively, the 2SLS estimate is just the ratio of the reduced form and the first stage. Given a large enough sample, $\hat{\pi} \xrightarrow{d} \pi$ and we will consistently estimate β

However, in a finite sample $\hat{\pi}$ is noisy, and if the standard error is relatively large, this can cause weird behaviour in $\widehat{\beta}$

We want to account for a lack of normality, we can either:

1. Test for a sufficiently strong first stage so we can ignore the issue
2. Use an approach that is robust to both

Pretesting - just check if the F-stat is large enough that these problems are not an issue. But the key assumption is homoscedasticity, this is a strong assumption.

Robust confidence intervals - with a just-identified single endogenous regressor, Anderson-Rubin confidence intervals are valid, irrespective of the weakness of the first stage.

Note. *Many Instruments Even many instruments create bias:*

$$E\left(\hat{\beta}_{2SLS} - \beta\right) \approx \underbrace{\frac{\sigma_{u\varepsilon}}{\sigma_u^2}}_{OVB}\left[\frac{\underbrace{E\left(\pi'Z'Z\pi\right)/K}_{First\ Stage\ F\ statistic}}{\sigma_u^2} + 1\right]^{-1} \tag{6}$$

Due to overfitting in the project of 2SLS. This is solvable via jackknife IV (which leaves out the own observation)

Solutions

1. Pretesting

A natural solution to this is just to check if the f-statistic is large enough that these highlighted problems are not an issue

3.6 Testing

The Wu-Durbin test is used to detect endogeneity in a regression model

(Wu-Durbin), Test for endogeneity:

H_0 : *exogeneity* vs. H_1 : *endogeneity*

$$\text{Under } H_0: \left.\begin{array}{l} \hat{\beta}^{OLS} \xrightarrow{H_0} \beta \\ \hat{\beta}^{IV} \xrightarrow{H_0} \beta \end{array}\right\} \hat{\beta}^{OLS} - \hat{\beta}^{IV} \xrightarrow{H_0} 0$$

$$\text{Under } H_1: \left.\begin{array}{l} \hat{\beta}^{OLS} \overset{H_1}{\not\longrightarrow} \beta \\ \hat{\beta}^{IV} \xrightarrow{H_1} \beta \end{array}\right\} \hat{\beta}^{OLS} - \hat{\beta}^{IV} \overset{H_1}{\not\longrightarrow} 0$$

Under H_0:

$$\left[\hat{\beta}^{IV} - \hat{\beta}^{OLS}\right]' \left[\hat{AV}(\hat{\beta}^{IV}) - \hat{AV}(\hat{\beta}^{OLS})\right]^{-1} \left[\hat{\beta}^{IV} - \hat{\beta}^{OLS}\right] \overset{H_0}{\sim} \chi^2$$

Hausman Test for Endogeneity

The Hausman test evaluates the null hypothesis that an estimator β_1 is consistent and efficient, against the alternative hypothesis that β_2 is consistent (but inefficient if β_1 is consistent).

- Need a generalised inverse except when the variance is singular
- Not robust to heteroscedasticity
- Assumes that we have a valid instrument

Practical Hausman test

With only a single suspected endogenous variable : focus on the coefficient for this variable

- e.g. If this variable is x_2 with parameter β_2

$$\frac{\widehat{\beta}_2^{IV} - \widehat{\beta}_2^{OLS}}{\sqrt{\hat{AV}\left(\widehat{\beta}_2^{IV}\right) - \hat{AV}\left(\widehat{\beta}_2^{OLS}\right)}}$$

- Still holds under homoscedasticity

Practical Hausman test 2

- Model $y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$; $y_2 = z\pi_2 + v_2$

- Where $z = [z_1, z_2]$ comprises m included and excluded exogenous variables and $x = [z_1, y_2]$ comprises k covariates
- All the IVs are uncorrelated with both error terms:

$$E\left[z'u_1\right] = E\left[z'v_2\right] = 0$$

- Test: $H_0 : \text{Cov}(y_2, u_1) = 0 \leftrightarrow \text{Cov}(v_2, u_1) = 0$
- Which is the same as testing $\rho_1 = 0$ in $u_1 = \rho_1 v_2 + e_1$
- And we replace in the model : $y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1$

1. Regress $y_2 = z\pi_2 + v_2$ compute $\hat{v}_2$
2. Regress $y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1\hat{v}_2 +$ error and test $H_0 : \rho_1 = 0$
3. Note 1 : under the null, if $\rho_1 = 0$, we do not need to worry about generated regressor problem, hence use usual standard errors (OLS / robust)
4. Note 2 : we obtain the 2SLS estimate for δ_1 and α_1 unless $\rho_1 = 0$ in which case they are OLS (control function approximation)

Example. *Hausman test*

```
data:  lwage ~ educ + black + hisp + I(exper^2) + married + union +  ...
chisq = 26.361, df = 10,
p-value = 0.003284
alternative hypothesis: one model is inconsistent
```

Figure 16: Hausman Test R Output - Tutorial 5

The Hausman Test rejects the null that the errors are uncorrelated with the regressors with a p-value of 0.003284. This would be in favour of the fixed effects approach. In practise a rejection can mean that the RE and FE are sufficiently close or the sampling variation is so large in the FE estimates that one cannot conclude practically significant differences.

Testing For Over Identification

Sargan Test If we have $m > k$: more instruments than endogenous regressors

- Test the over identifying restrictions to see whether the remaining $m-k$ instruments are correlated with the error
  1. Estimate by 2SLS, using all instruments

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$$

     obtain the residuals $\hat{u}_1$
  2. Regress residuals on all the exogenous variable $z = [z_1, z_2]$ and obtain R_u^2
  3. Test $H_0 : E\left[z'u_1\right] = 0$

$$nR_u^2 \ \chi^2_{m-k}$$

- A rejection of the null : at least one of the instruments is not valid but we do not know which

- If we have 2 instruments and we reject the null, one instrument is not valid but does not necessarily mean that both aren't valid
- Maintains homoscedasticity assumption

Testing For Weak Instruments

Testing the first stage

- For a single instrument : t-test
- For several instruments : test the joint significance of the instruments in the first stage (excluding the other exogenous variables and intercept in the test)

It has been found that weak instruments lead to distorted statistical inference : even with a large sample, the 2SLS can be biased and the distribution very different from the standard normal

The rule of thumb under homoscedasticity : first stage f-stat (single endogenous regressor) : $F > 10$ or t-stat $|t| > 3.2$

Under heteroscedasticity : more stringent, $F > 20$ suggested (ongoing research)

Group Mean Estimator

[**L4-GM**] In some situations, have instruments that can be changed into 2 groups, water (of birth/ financial year). 'Chop instrument into groups' like Moulton problem/structure.

i	z_g	y_{ig}	1	x_{ig}
1	1	y_{11}	1	x_{11}
2	1	y_{21}	1	x_{21}
⋮	⋮	⋮	⋮	⋮
n_1	1		1	
1	2	y_{12}	1	x_{12}
2	2	y_{22}	1	x_{22}
⋮	⋮	⋮	⋮	⋮
n_2	2		1	
⋮	⋮	⋮	⋮	⋮
1	G	y_{1G}	1	x_{1G}
2	G	y_{2G}	1	x_{2G}
⋮	⋮	⋮	⋮	⋮
n_G	G		1	

It can be shown that group mean estimator is IV, a weighted least squares regression, where it is sufficient to know size of groups and means, do regression and obtain estimator that is equivalent to an IV estimator, that is consistent despite the fact we have an endogenous variable.

- Model : $y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig}$, $g = 1, \ldots, G$
- Where x_{ig} is endogenous
- Let z_g be an IV : a variable with G distinct values representing G groups
- We have g *moment conditions* $E\left[u_{ig} | i : z_g = I\right] = 0$, $I = 1, \ldots, G$
- if this is IV, we know exogeneity must hold, whether group 1 or 2, the error term conditional on this group needs to be equal to 0, this must hold for all groups. Essentially, we have g different groups this is really $E[y_{ig}|z_g = I] = \beta_0 + \beta_1 E[x_{ig}|i : z_g = I]$
- To estimate an expectation, we replace with an average, since this is conditional, to estimate the expectation for the first group (born in the 1-st quarter), we take the average for the first group (conditional average by restricting to the first group) - taking the means of all the groups
- We do the same for $\bar{x}$ and intuitively obtain

$$\bar{y_g} = \beta_0 + \beta_1 \bar{x_g} + \bar{u}_g$$

Aggregated Data

- If we only have aggregated data available, one needs the group sizes for estimation
- If we define a dummy variable for each single group, it can be shown that GLS applied to the aggregated model is identical to 2SLS
- Using m dummies, defined by $z_g : d'_g = 1[z_g = l]$ for $I = 1, \ldots, G$
- Meaning the group means estimator is consistent

Essentially, doing this is the same as using dummy variables for quarter of birth in 2SLS regression, **thus** group means are consistent.

When G = 2:

$$E\left[y_{ig}|1_{z_g=g}=1\right] = \beta + \beta_1 E\left[x_{ig}|1_{z_g=g}=1\right]$$

$$E(y_{ig}|z_g = 1) = \beta_0 + \beta_1 E(x_{ig}|1_{z_g=g} = 1) \text{ and}$$
$$E(y_{ig}|z_g = 2) = \beta_0 + \beta_1 E(x_{ig}|z_g = 2)$$
$$\text{solving for } \beta_1$$
$$\beta_1 = \frac{E(y_{ig}|z_g = 2) - E(y_{ig}|z_g = 1)}{E(x_{ig}|z_g = 2) - E(x_{ig}|z_g = 1)}$$
$$\text{with Sample Analogs :}$$
$$\widehat{\beta}_1 = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}$$

Which results in the Wald estimator

Derivation

- $Y_{ig} = x_{ig}\beta + u_{ig}$, $g = 1, \ldots, m$, $i = 1, \ldots, n_g$
- Use $z_{ig} = (d_g^1, \ldots, d_g^m)$ as instruments such that $d'_g = 1$ if $i = g$
- Apply 2SLS
    1. Predict the first stage
    2. Apply OLS on the predicted values
    3. Show that this equal OLS on the grouped model : $\bar{y}_g = \bar{x}_g\beta + \bar{u}_g$, $g = 1, \ldots, m$

1. regressing $x_{ig} \sim d_g^1 + \ldots + d_g^m + v_{ig}$ gives $\bar{x}_{ig} = \bar{x_g}$ the fitted values are the sample mean of the dependent variable for each category
2. Regress y_{ij} on the fitted values

$$\left(\sum_i x'_{ig}x_{ig}\right)^{-1}\left(\sum_i x'_{ig}y_{ig}\right) = \left(\sum_g n_g \bar{x}'_g \bar{x}_g\right)^{-1}\left(\sum_g \bar{x}'_g \sum_{i:d_g=1} y_{ig}\right)$$

$$= \left(\sum_g n_g \bar{X}'_g X_g\right)^{-1}\left(\sum_g n_g \bar{X}'_g y_g\right) = \left(\sum_g \frac{n_g}{n} \bar{X}'_g X_g\right)^{-1}\left(\sum_g \frac{n_g}{n} \bar{X}'_g y_g\right)$$

Which is WLS with weights $\frac{n_g}{n}$. Hence 2SLS on the micro data is WLS in grouped means

Exercise 5. *Group mean estimator - what happens if the number of groups = 2 If we have dummy variable, we obtain the Wald estimator (last week). Our instrument, we obtain the same expression In order to derive,*

1. *Regress x on dummies, x can only belong to 1, so fitted values are sample means of dependent variable x_{ig}. Fitted values $\hat{x_{ig}}$ are means $\bar{x}_g$*
2. *Apply OLS on this after we have found fitted values, the predicted values are our sample means*

Exclusion Restrictions

Even with a variable that is near-random in its allocation, the exclusion restriction is not always satisfied, and worse yet is is untestable.

Using an IV requires thinking about how one can justified this and what it implies

Angrist, Imbens and Rubin (1996) : the larger the complier group is, the less the bias from violation in the exclusion restriction.

Notes

Correctly done, IV gives a very internally valid estimate, but external validity is worrisome

Where $\Delta = Y_i(1) - Y_i(0)$:

1. $\Delta^{ATE} = E(\Delta)$
2. $\Delta^{ATT}(D = 1) = E(\Delta \mid D = 1)$
3. $\Delta^{LATE}(P(z), P(z')) = \frac{E(Y|p(z)) - E(Y|p(z'))}{p(z) - p(z')}$

I.e. Is the range $[\mathrm{P}(z'), \mathrm{P}(z)]$ special? Is it informative?

Table 1. Causal Effect of Z on Y, $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$, for the Population of Units Classified by $D_i(0)$ and $D_i(1)$

		$D_i(0)$	
		0	1
$D_i(1)$	0	$Y_i(1, 0) - Y_i(0, 0) = 0$ Never-taker	$Y_i(1, 0) - Y_i(0, 1) = -(Y_i(1) - Y_i(0))$ Defier
	1	$Y_i(1, 1) - Y_i(0, 0) = Y_i(1) - Y_i(0)$ Complier	$Y_i(1, 1) - Y_i(0, 1) = 0$ Always-taker

Figure 17: Compliers vs Non-compliers

3.7 Applications

Angrist and Krueger (1991)

- Does compulsory school attendance affect schooling and earnings
- Using *quarter of birth* as an instrument
    1. Exclusion : Season of birth is a natural experiment and hence unrelated to innate ability, motivation or family connections
    2. Relevance : In the US, children were allowed to drop out at 16. Since the age of starting school differs, children have different lengths of schooling when they turn 16

- Potentially weak instrument and potential reasons why quarter of birth might be somewhat correlated with the error

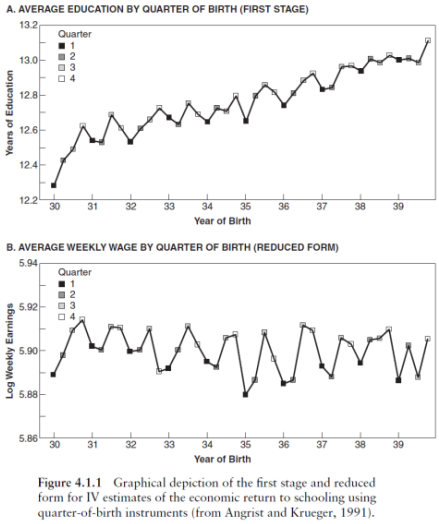


Figure 4.1.1 Graphical depiction of the first stage and reduced form for IV estimates of the economic return to schooling using quarter-of-birth instruments (from Angrist and Krueger, 1991).

Figure 18: First stage and reduced form

Men born in early quarters have a lower schooling on average, and they also earn on average less than those born later in the year

TABLE 4.1.2
Wald estimates of the returns to schooling using quarter-of-birth instruments

	(1) Born in 1st Quarter of Year	(2) Born in 4th Quarter of Year	(3) Difference (Std. Error) (1) − (2)
ln (weekly wage)	5.892	5.905	−.0135 (.0034)
Years of education	12.688	12.839	−.151 (.016)
Wald estimate of return to education			.089 (.021)
OLS estimate of return to education			.070 (.0005)

Notes: From Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930–39 birth cohorts in the 1980 census 5 percent file. The sample size is 162,515.

Figure 19: Wald estimates of IV - weak/exclusion violated?

- Denote $z = 1$ if born in quarter 1 and $z = 0$ if born in quarter 4

- Wald estimate / IV :

$$\widehat{\beta}_1^{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} = \frac{-0.0135}{-0.151} = 0.089$$

- OLS regression of $y \sim x$ yields $\widehat{\beta}_1^{OLS} = 0.070(0.0005)$. Since it is smaller than IV, ability bias does not seem to drive the result
- Both estimates are small but significant

Can't test by how much IV exclusion is violated, it might be best to use OLS, but in the same sense it may be incorrect - can we search for better instrument? Or,

TABLE 4.1.1 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.071	0.067	0.102	0.13	0.104	0.108	0.087	0.057
	(.0004)	(.0004)	(0.024)	(0.020)	(0.026)	(0.020)	(0.016)	(0.029)
Exogenous Covariates								
Age (in qtrs)								√
Age (in qtrs) squared								√
9 yr-birth dummies		√			√	√	√	√
50 state-of-birth dummies		√			√	√	√	√
Extra Instruments								
Dummy for QoB=1			√	√	√	√	√	√
Dummy for QoB=2				√		√	√	√
Dummy for QoB=3				√		√	√	√
QoB dummies interacted with year-of-birth dummies							√	√
No covariates *k*	2	61	2	2	61	61	61	63
No instruments *m*	2	61	2	4	61	63	93	95

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QoB denotes quarter of birth.

Figure 20: 2SLS estimates of economic returns to schooling

Inflated standard errors : $0.\frac{0.021}{0.0005} = 42$, even though the estimate is significant due to a large sample size, the 9% CI is large.

Problem also of a small $R^2_{x,z}$: the *instruments are weak*, it might be better to use OLS instead of IV

Including more instruments and covariates

- Reduces SE but comes to the cost of potentially having a weak instrument
- Col 3 : just identified 1 instrument
- Col 4 : over identified (3 QoB instruments)
- Col 5/6 : + 59 covariates to 3/4 : m - k = 0 (or 2 resp)
- Col 7 : + 30 Ifs (m-k =32)
- Col 8 : + age and age^2 to x and z ($m - k = 32$)

But there is potential to test for over identifying restrictions, using the Sargon test

Why β larger? Asymptotic variance depends on error variance, depends on the $R^2_{x,z}$ from first stage regression (x on instruments, will be higher if instruments highly relevant and vice versa).

Exercise 6. *Consequences of weak instruments*

1. *High SE*
2. *Slight violation of exclusion restrictions leads to large bias*

Returns to Education

Wooldridge (2010) using data on married working women Looking at the returns to education or specifically the effect of years of education (educ) on the log-wage (lwage) However, there is *omitted variable bias* arising from innate ability

The idea is to use the education of the father as an IV (fatheduc) (one generation before), and the dataset also include information on the mother's education and number of siblings

Model : $lwage = \beta_0 + \beta_1 educ + u$ And it is likely that $E[u|educ] \neq 0$

However, whether this is a valid instrument, we would want to loo at $\text{Cov}(u, fatheduc) = 0$?, but we cannot test this

Though we can say, it is probably not an exogenous instrument, father's education is likely to be correlated with father's ability and ability might be correlated through generations

Relevance However, whether $\text{Cov}(educ, fatheduc) \neq 0$, we can test this, and find there is a significant effect of fatheduc on educ

```
> OLS = lm(lwage ~educ,data = dat)
> summary(OLS)

Call:
lm(formula = lwage ~ educ, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-3.10256 -0.31473  0.06434  0.40081  2.10029

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1852     0.1852  -1.000    0.318
educ          0.1086     0.0144   7.545 2.76e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.68 on 426 degrees of freedom
Multiple R-squared:  0.1179,    Adjusted R-squared:  0.1158
F-statistic: 56.93 on 1 and 426 DF,  p-value: 2.761e-13
```

Thus, 1 more year in educ, increases lwage by 10.9%

Running the first stage of this regression, so $educ \sim fatheduc$ results in an increase of educ by 27% ,and we find the covariance between educ and fatheduc as $\rho = 0.41$ - thus a significant positive correlation

Running the second stage of this regression so $lwage \sim educhat$ results in an increase in lwage of 5.9%, and thus this is the returns to education.

We use the Hausman test and find a marginal non-rejection of the null of exogeneity

```
> IV = ivreg(lwage ~ educ |fatheduc , data = dat)
> summary(IV, diagnostics = TRUE)

Call:
ivreg(formula = lwage ~ educ | fatheduc, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44110    0.44610   0.989   0.3233
educ         0.05917    0.03514   1.684   0.0929 .

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1 426     88.84  <2e-16 ***
Wu-Hausman         1 425      2.47   0.117
```

Figure 21

Including the first stage regression in our $lwage \sim educ$ regression we find the same estimates on educ as with IV

```
> HM = lm(lwage ~educ + FirstStageRes,data = dat)
> summary(HM)

Call:
lm(formula = lwage ~ educ + FirstStageRes, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-3.06788 -0.31334  0.06076  0.40920  2.12339

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.44110    0.43929   1.004    0.316
educ          0.05917    0.03461   1.710    0.088 .
FirstStageRes 0.05979    0.03804   1.572    0.117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6789 on 425 degrees of freedom
Multiple R-squared:  0.123,     Adjusted R-squared:  0.1189
F-statistic:  29.8 on 2 and 425 DF,  p-value: 7.753e-13
```

Summary

- OLS estimates too large (positive omitted variable bias)
- Relatively weak instrument : IV standard errors is 2.5-times bigger than OLS
- Leads to large IV confidence intervals
- The OLS estimates is inside the 95% CI for the IV estimate :we cannot tell whether the estimates are significantly different from each other
- Wu-Hausman test(s) do (marginally) not reject the null of exogeneity with a p-value of 0.117

Exercise 7. *Discuss intuition behind Hausman Test Exogenous regressor then both (testing for whether x endogenous) OLS and IV consistent (if we have valid instrument) If we have an endogenous regressor and OLS is not consistent, difference does not converge to 0 any more, test start follows chi-squared distribution with k degrees of freedom, to test for endogeneity 2 main assumptions, 1 test for, 1 assume*

- *Relies on valid instrument, otherwise $\hat{\beta}$ would not converge at all, this is almost critical assumption in Hausman test*

Why is the exclusion restriction challenging?

Example. *Vietnam war lottery numbers as an IV for military service, styling the impact on mortality*

Where Y is death, D is Vietnam vet, Z is lottery number. Lottery number was randomly assigned as a function of birth-date and thus makes a well-defined design-based view of Z allocation

However, being drafted induces you to change your behaviour to avoid the draft, you may stay in school, or flee to Canada and this would clearly violate the exclusion restriction

Example. *Rainfall as an instrument for income in agriculture environments (since many crops are heavily dependent on it)*

Where Y is conflict, D is income, Z is rainfall.

The exclusion restriction is that rainfall has no effect on conflict beyond income. Whilst the logic seems reasonable, it has been shown that places with dams (which protect against the income shocks due to rain) have similar conflict to those without dams. And thus it is plausible that whilst rain is "random", it might have many channels

Tutorial

Tutorial 1. *IV and simultaneity bias We have expression for y_1 and y_2, we are going to replace this equation, since our asymptotic bias will sum to .? Asymptotic bias = $\frac{Cov(u_1,y_2)}{V(y_2)}$ Then we plug long covariance into $Cov(u_1, y_2)$ We know $cov(u_1, u_2) = 0$ and $cov(u_1, z_2) = 0$, but the problem is the variance is typically positive But depending on assumptions we can determine direction of bias based upon α_2 We show that in the formula we replace by y_1*

Tutorial 2. *IV is asymptotically unbiased That is plim $\tilde{\alpha_1} = \alpha_1$ (the IV estimator)*

Tutorial 3. *why can z_2 not be used as an instrument for y_1 to estimate α_2 the slope of the supply curve It is under-identified, we don't have an endogenous variable at our disposal, we don't have a shock for y_1 → we don't have an instrument, two endogenous variables require 2 instruments*

Tutorial 4. *Exercise 1 (tut3) Regression of log wage on education, estimate using OLS, internet, do you expect OLS to be trustworthy? Education on wage includes ability and motivation etc explaining the wages, that are correlate with education → OVB. We expect a positive omitted variable bias, since ability is likely correlated with log wages Testing relevance condition by FS regression : running education on number of siblings, this is significantly different from 0 and f-stat >10 Then running IV regression, we find the instrument has strong enough F-stat, but it could be that the exclusion restriction is violated, before we found coefficient of 0.059, with IV we find 0.122 (12%), which is higher than OLS, revealing inconsistency already, perhaps our assumption about exogeneity is not fulfilled.*

Tutorial 5. *Exercise 2 Using sibs as iv is not same as plugging sibs into education (as in proxy), we find very different result from our IV estimator, that is big diff from 0 controlling for. Education and birth quarter negatively correlated? B) C) again, we get an increase than the OLS estimator, and larger than when we used siblings as IV. But do we have similar concerns now using birth order Is birth order endogenous? Like the number of siblings? The decision to have children might be related to budget constraints etc. D) identification assumption $\log(wage) = \beta_0 + \beta_1$ Test whether π_2 is significantly different from 0, if we estimate our IV, we need to include all exogenous variables as instruments, we estimate a different coefficient.*

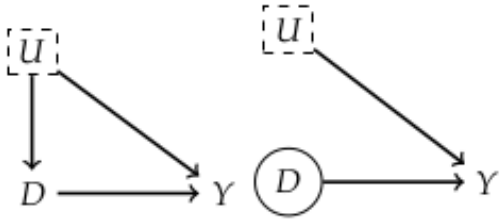
Lecture 5: RCT

Tue 27 Feb 16:02

[L5-RCT]

4 Randomised Experiments

Being able to truly randomise an intervention allows the research to assume by definition that the potential outcomes for units are independent, satisfying the first assumption in strong ignorability.



The effect of D on Y is confounded by U, but a randomised intervention of D breaks any back-door connection since randomisation was the only cause of the intervention.

Issues

1. People may not want to be randomised into different treatments. It may be impractical to randomise their decisions even if their is a clear benefit to doing so, e.g. A firm may not want to randomise their policies
2. It may be unethical to randomise. If there is a clear benefit to treatment, it may be unethical to withhold that treatment from individuals by placing them in the control.
3. It may be impossible to randomise. For example if we are interested in the effect of a policy change, it may be impossible to randomise the policy change across different regions or states.

Randomisation And Design-based Inference Knowledge of the treatment assignment mechanism gives a very powerful tool for thinking about the counterfactual. Randomised intervention with knowledge of treatment assignment mechanism is the "gold standard" and in methods other than this we need to make assumptions about the treatment assignment mechanism and defend them.

There are many things we could want to know about the relationship between D_i and $\tau_i = Y_i(1) - Y_i(0)$ but here we focus on $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau_i$

Design based inference considers the set of potential ways that D could be randomised to the population. We assume that $Y_{1/0}$ are fixed - it is only the random variation in D that creates uncertainty.

We need an estimator for $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau_i$ We already know under random assignment that $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ identifies $E[\tau_i]$ Then the empirical analog is quite easy

$$\begin{aligned}\hat{\tau}(D, Y) &= \frac{D'Y}{\sum_i D_i} - \frac{(1-D)'Y}{\sum_i (1-D_i)} \\ &= n_1^{-1} \sum_i Y_i D_i - n_0^{-1} \sum_i (1-D_i)\end{aligned}$$

Since D is random, we know its marginal distribution over the sample and we can show that this estimator is unbiased (this estimator requires complete random assignment of units across treatment) Under this design, the probability of a unit receiving treatment given a draw D is $\pi = \frac{n_1}{n}$

4.1 Introduction

Motivation : Program Evaluation

Binary treatment on a set of outcomes, lets say the effect of having internet at home on school grades, however this would of course run into selection bias since the decision to have internet at home might depend on other unobserved factors (income etc).

Thus, program evaluation is often about how to overcome the problem of selection bias, using the potential outcomes framework allows us to illustrate this.

Example. *Do hospitals make people healthier? If we have data on the following questions :*

1. *In the last 12 months have you spent a night in hospital?*
2. *What would you rate you health 1-5 (being excellent)*

Group	Sample Size	Mean Health	SE
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

Figure 22: Naive Hospital Comparison

This naive comparison of individuals hospitalised and not, a difference of 0.72 suggest that non hospitalised people are healthier Thus, can we ask does going to the hospital make people sick? Maybe in some cases, but the main problem is self selection

- *People who decide to go to the hospital are less healthy to begin with*
- *Even if the treatment works, such individuals won't be healthier than those who do not go to the hospital*

We can formalise this with the Potential Outcomes framework

4.2 Potential Outcomes Framework

Treatment Allocation And Outcomes

- Start with single unit I
- Denote the outcome of interest by Y and treatment variable D
  - D = 1 the individual is *treated*
  - D= 0 the individual is *not treated* (control)
- Typical assumption is that one individual can have 2 states
  1. Y(1) - the potential outcome if I receives treatment
  2. Y(0) - the potential outcome if I 'would not' recieve the treatment (control)

- Individual Causal Effect of the treatment for observation I:

$$Y(1) - Y(0)$$

- The *problem of causal inference* - is that it is impossible to observe **both** potential outcomes at the same time, only one is realised → thus is is impossible to observe the causal effect

Stable Unit Treatment Value Assumption

- Generalisation to n units $i = 1, 2, \ldots, n$
- Let D_i be the treatment for unit i
- Each unit can be exposed to the two treatments : *the problem is that in principle the potential outcomes can depend on the treatment of all units*
- *thus we make the assumption* that the potential outcome for unit i depends only on the treatment received by unit i and not on the allocation of other individuals

Assumption 1. *SUTVA Denote $D_{-i} = (D_j) : j \neq i$ treatment status of all other individuals in the population. Then SUTVA states*

$$Y_i(1), Y_i(0) \perp D_{-i}$$

- Aka the 'no interference assumption'
- However, this might be violated if individuals intact
- There cannot be contagion between individuals

Realised Outcomes

Observed outcomes

- Let D_i be the observed treatment
- We only observe the observed outcome Y_i
- Then the observed outcomes is a function of the potential outcomes and the treatment:

$$\begin{aligned} Y_i &= D_i Y_i(0) + (1 - D_i) Y_i(0)_i \\ &= Y_i(0) + [Y_i(1) - Y_i(0)] D_i \end{aligned}$$

  which is sometimes called the "switching equation"

Naive Comparison The natural starting point is a naive comparison of treated and untreated

$$\begin{aligned} \tau_{WW} &= \underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{population raw differential}} \\ &= E[1Y_i(1) + (1-1)Y_i(0)|D_i = 1] - E[0Y_i(1) + (1-0)Y_i(0)|D_i = 0] \\ &\quad \pm E[Y_i(0)|D = 1] \\ &= \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{\text{average treatment effect on treated (ATT)}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}} \\ &\equiv \tau_t + \tau_{SB} \end{aligned}$$

Selection bias

- Differ in average non treated health between those who did and those who did not go to hospital
- Because the sick are more likely to seek treatment, we expect a negative selection bias $E[Y_i(0)|D_i = 1] < E[Y_i(0)|D_i = 0]$
- If the ATT is only slightly positive, we could observe a difference that is negative as with the hospital example

4.3 Treatment Effects

- Average treatment effect (ATE):
$$\tau = E[Y_i(1) - Y_i(0)]$$
  - The average of the individual causal effects in the overall population
  - Average effect of hospitalisation on all people
- Average treatment effect on the treated (ATT):
$$\tau_t = E[Y_i(1) - Y_i(0)|D_i = 1]$$
  - Average of the individual causal effects on those who receive the treatment
  - Average effect of hospitalisation on those who went to the hospital
  - In the hospital example, we would hope that ATT and ATE are positive
- Similarly, average treatment effect on the untreated (ATU)

Observability

- $E[Y_i(0)|D_i = 1]$ is not observable
  - We cannot assess the magnitude or sign of the selection bias
$$\begin{aligned}\tau_{SB} &= E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= \underbrace{E[Y_i(0)|D_i = 1]}_{\text{unobservable}} - \underbrace{E[Y_i|D_i = 0]}_{\text{observable / estimable}}\end{aligned}$$
  - We cannot observe the ATT:
$$\begin{aligned}\tau_t &= E[Y_i(1) - Y_i(0)|D_i = 1] \\ &= \underbrace{E[Y_i|D_i = 1]}_{\text{observable / estimable}} - \underbrace{E[Y_i(0)|Di = 1]}_{\text{counterfactual}}\end{aligned}$$

4.4 Perfect compliance

- Imagine you could manipulate the treatment variable in a way that the treatment assignment Z_i is random
- If $Z_i = 1$, individual i is treated
- If $Z_i = 0$, individual i is not treated

- Since Z_i is randomly assigned, it must hold true that the variable is independent from the potential outcomes:
$$Y_i(1), Y_i(0) \perp Z_i$$
- This means also: $E[Y_i(0)|Z_i = 1] = E[Y_i(0)|Z_i = 0]$ which implies that $\tau_{SB} = 0$
- Which is sometimes called the "brute force design"

Identification under independence The naive comparison recovers the ATE and ATT

$$\begin{aligned} E[Y_i|Z=1] - E[Y_i|Z=0] = E[Y_i(1)|Z_i=1] - E[Y_i(0)|Z_i=0] & \\ = E[Y_i(1)|Z_i=1] - E[Y_i(0)|Z_i=1] & \\ = \tau_t & \\ = E[Y_i(1) - Y_i(0)|Z_i=1] & \\ \text{similarly} & \\ = E[Y_i(1) - Y_i(0)] & \\ = \tau & \end{aligned}$$

Under perfect compliance, $ATE = ATT$ since random assignment eliminates the selection problem

We can estimate the ATT and ATE using the difference in groups means:

$$\hat{tau} = \frac{1}{n_1} \sum_{i:Z_i=1} Y_i - \frac{1}{n_0} \sum_{i:Z_i=0} Y_i$$

Estimation by OLS - constant TE Suppose that the treatment effect is the same for everyone (constant)

$$Y_i(1) - Y_i(0) = \tau$$

We can rewrite the regression in the form:

$$Y_i = \underbrace{\beta_0}_{E[Y_i(0)]} + \underbrace{\tau}_{Y_i(1)-Y_i(0)} Z_i + \underbrace{u_i}_{Y_i(0)-E[Y_i(0)] \equiv u_i(0)}$$

Where u_i is the random part of $Y_i(0)$

We know that

$$\begin{aligned} E[Y_i|Z_i=1] &= \beta_0 + \tau + E[u_i|Z_i=1] \\ E[Y_i|Z_i=0] &= \beta_0 + E[u_i|Z_i=0] \end{aligned}$$

Therefore

$$\begin{aligned} & E[Y_i|Z_i=1] - E[Y_i|Z_i=0] \\ &= \tau + E[u_i|Z_i=1] - E[u_i|Z_i=0] \\ &= \tau + \underbrace{E[Y(0)|Z_i=1] - E[Y(0)|Z_i=0]}_{\tau_{SB}} \end{aligned}$$

Thus, selection bias amount to correlation between error term u_i and the regressor Z_i Inclusion of (exogenous) covariates can increase precision without changing the estimate

Relaxing constant TE

We can relax the constant treatment effect assumption

Define $Y_i(1) = E[Y_i(1) + u_i(1)]$ and $Y_i(0) = E[Y_i(0)] + u_i(0)$ Then recalling that $\tau = E[Y_i(1) - Y_i(0)]$ Hence

$$Y_i(1) - Y_i(0) = \tau + u_i(1) - u_i(0)$$

Then, using the switching equation:

$$Y_i = \underbrace{\beta_0}_{E[Y(0)]} + \tau Z_i + \underbrace{\varepsilon_i}_{u_i(0)+[u_i(1)-u_i(0)]Z_i}$$

Note that the error structure depends on the random assignment, both groups have different variances if the variances of the potential outcomes differ Note again that the zero conditional mean still holds since the potential outcomes are independent of Z_i :

$$E[\varepsilon_i|Z_i] = E[u_i(0)|Z_i] + E[u_i(1) - u_i(0)|Z_i] Z_i = 0$$

Properties

From linear regression we know th following properties

1. Unbiased $E[\hat{\tau}] = \tau$
2. Consistent $\hat{\tau} \xrightarrow{p} \tau$
3. asymptotically normal
4. Variance $\hat{V}(\hat{\tau}) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}$

Since ATE = ATT, this holds for both estimands.

Balancing checks

- Due to randomisation of Z_i, treatment and control groups should have on average similar characteristics
- We can check testing e.g. For $H_0 : \beta_1 = 0$:

$$X_i = \beta_0 + \beta_1 Z_i + u_i$$

- Test only covariates before treatment or covariates that cannot be altered by treatment
- Test for individual significance of the covariates and for joint significance

Including Covariates Choose wisely:

- In a simple randomised experiment, controlling for covariates likely to influence the outcome does not affect the estimated effect of the treatment
- It may reduce variance
- Not include covariates affected by the treatment (ideally collected at baseline)
- Can increase variance if they explain little of the outcome variation

4.5 Imperfect Compliance

- In some cases, its impossible to enforce compliance to the randomisation
- If we were to randomise the offer to have internet at home, students who are randomised to receive a voucher for internet
- Students now can decide on both randomisation arms whether to get internet using a voucher or not (if not, can use other sources)
- Imperfect compliance can typically be subsumed under two forms
  - Encouragement design : individuals in both the treatment and control groups can decide to take up the treatment
  - Eligibility design : the control group can be prevented from taking up the treatment

Setup

- Individuals $i = 1, \ldots, n$ receive a randomised offer Z_i to take up a program
- $Z_i = 1$ if the individual is randomised into the treatment group or is offered the treatment and Z_i otherwise
- Denote the actual program participation or receipt of the treatment by D_i
- $D_i = 1$ if the individual chooses to participate and $D_i = 0$ otherwise
- If $D_i = Z_i$, we have perfect compliance. Otherwise we call the setup imperfect compliance

Intention to Treat

- Comparing individuals who are randomised in with those randomised *out* identifies the **intention to treat** effect (ITT):
$$\tau_{ITT} = E\left[Y_i | Z_i = 1\right] - E\left[Y_i | Z_i = 0\right]$$
- Since Z_i is randomised, we obtain a causal interpretation
- The effect of the randomised offer of treatment (not of the treatment itself)
- Can be a parameter of interest
- However, we are often interested in the effect of the treatment

Encouragement Design

- Individuals can choose to participate in both randomisation arms
- The participation probability conditional on the randomisation arm is in both cases non zero:
$$P(D_i = 1 | Z_i = z) > 0 \quad , \quad z \in \{0, 1\}$$
- Denote the potential participation given the randomisation status by $D_i(z)$
- Observed treatment is therefore
$$D_i = D_i(0) + (D_i(1) - D_i(0))Z_i$$

Individuals receive a voucher for internet at home and can decide to get internet or not. They can also decide to get internet if they do not receive the voucher. Take-up can happen in both groups.

Subpopulations We can split the individuals in 4 different groups

1. Always Takers (AT) : individuals who will always take up the treatment regardless of their randomisation status ($D_i(1) = D_i(0) = 1$)
2. Never takers (NT) : individuals who will never take up the treatment regardless of their randomisation status ($D_i(1) = D_i(0) = 0$)
3. Compliers (C) : individuals who will do as the experiment induces them to do. They take up the treatment when they are randomised and do not when they are randomised out. $D_i(1) = 1, D_i(0) = 0$ which is equivalent to $D_i(1) - D_i(0) = 1$
4. Defiers (D) : individuals who do the opposite of what the experiment induces them to do: they do not take up the treatment when they are randomised in and they do take up the treatment when they are randomised out ($D_i(1) = 0, D_i(0) = 1$)

LATE - The Local Average Treatment Effect

Let $Y_i(z, d)$ be the potential outcome for individual i with treatment status $D_i = d$ and the assignment $Z_i - z, z, d, \in \{0, 1\}$

1. Independence $[\{Y_i(z, d) \forall d, z\}, D_i(1), D_i(0)] \perp Z_i$
2. Exclusion Restriction $Y_i(d, 0) = Y_i(d, 1) = Y_i(d)$
3. First Stage $\mathrm{P}(D_i = 1|Z_i = 1) - \mathrm{P}(D_i = 1|Z_i = 0) > 0$
4. Monotonicity $D_i(1) - D_i(0) \geqslant 0$ for all I

- Independence should be satisfied by good random assignment
- Exclusion need to be discussed, not justified by random assignment since it can be violated
- The first stage ensures that compliers exists, which is equivalent to the Wald estimator in IV, giving the share of compliers under monotonicity. We assume everybody reacts to the treatment in the same way (thus ruling out the existence of defiers)
- *Monotonicity* implies that we cannot have *any defiers*. It needs to be examined but often plausible in this setting when we assume that assigning someone to the active treatment increases the incentive to take the active treatment

LATE Estimand

Under the LATE assumptions, the ITT divided by the share of compliers recovers the LATE

$$E\left[Y_i(1) - Y_i(0)|D_i(1) - D_i(0)\right] = \frac{E\left[Y_i|Z_i = 1\right] - E\left[Y_i|Z_i = 0\right]}{P(D_i = 1|Z_i = 1) - P(D_i = 1|Z_i = 0)}$$

This is the average treatment effect on the subgroup of individuals whose treatment status has been affected by the assignment. In this setting, we cannot identify the ATE or AT or NT without additional assumptions

Late And Bloom Result

Exercise 8. [**Late-bloom**] *Starting from monotonicity, we can write total variation of binary variable, condition this equal to 1 plus same thing conditional on 0, then this term doesn't exist any more, we get rid and our numerator is equal to The switching function exists only due to the exclusion restriction, otherwise we would need to write y as a function observed by both variables* ***The denominator*** *- same thing but for FS, replace now here*

IV estimator provides more meaningful interpretation, provides average treatment effect for average treated people We take the numerator again Replace the observed outcome with switching equation via the switching equation, then replace switching equation since the exclusion restriction holds We also know only the non-treated PO is realised, By independence Since D_i is binary

This is a general framework for a binary IV estimator too. The LATE/ATT is obtained.

Eligibility Design

- In this setting, the individuals who are randomised in can chose to participate but those who are randomised out cannot participate
- Since the take-up on the control group arm is zero, the Wald estimand recovers the ATT
- Hence, the ITT divided by the share of participants recovers the ATT

Bloom Result Suppose the assumptions of the LATE theorem hold and $E\left[D_i|Z_i=0\right]=P\left[D_i=1|Z_i=0\right]=0$ The bloom result is

$$E\left[Y_i(1)-Y_i(0)|D_i=1\right]=\frac{E\left[Y_i|Z_i=1\right]-E\left[Y_i|Z_i=0\right]}{P\left[D_i=1|Z_i=1\right]} \tag{7}$$

The randomisation design recovers the ATT Which, we can notice, the LATE looks just like the Walt estimand from the IV for a binary instrument, which we estimate both designs via 2SLS, instrumenting D_i with Z_i

4.6 Going Further

Stratification

- Use baseline information to stratify / block the sample in order to improve precision
- Divide samples into groups to obtain an equal proportion of treated and untreated within each block
- Can decrease the variance
- Can be used to perform subgroup analysis
- Include the strata in the regression as dummies

Clustering

- If individuals can interact with each outer, the treatment status of one individual can influence the potential outcome of another
- This is a violation of SUTVA

- Randomisation at the group / cluster level where individual cannot interact
- Compute cluster-robust SE
- Often less costly to implement but increases the variance of estimator which means less power

Discussion

- Internal validity : ability to estimate causal effects with the study population
- External validity : ability to general's the results from a specific setting, to other settings (population / outcomes / contexts)
- RCTs can have a strong internal validity but are often criticised for a lack of external validity

4.7 Application

Gerber et al (2009) "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions"

- Randomised experiment to measure the effect of political news content on political behaviour and opinions
- The Washington DC are is served by the Washington times (more right leaning) and the Washington post (more left leaning)
- One month before Virginia election in November 2005, the authors run a short survey to a random selection of households
- 3347 households reported not to receive the post or the times and responded to all questions. Authors then randomly assigned these to get subscriptions to either paper for 10 weeks
- Control group DiD not get either paper in the framework of this experiment
- Households have been drawn from a voters register and consumers list
- Stratified randomisation
- One week after the election, the authors conducted a follow-up survey about political behaviour and opinions
- Authors had further access to state administrative dataset including voter turnout data for the November 2005 and 2006 elections

TABLE 1A—SUMMARY STATISTICS FROM BASELINE SURVEY
(Sample frame: all baseline survey respondents, mean, standard errors, and standard deviations)

	All (1)	Control (2)	*Post* (3)	*Times* (4)	*p*-value (5)
% female	34.76 (0.84) [47.63]	34.44 (1.28) [47.54]	33.01 (1.53) [47.05]	37.02 (1.59) [48.31]	0.18
% voted in 2004 (self-report)	88.62 (0.78) [31.77]	88.51 (1.22) [31.91]	88.82 (1.44) [31.54]	88.57 (1.45) [31.86]	0.99
% voted in 2002 (self-report)	48.08 (1.23) [49.98]	49.04 (1.92) [50.03]	45.76 (2.27) [49.87]	49.06 (2.28) [50.04]	0.48
% voted in 2001 (self-report)	7.30 (0.64) [26.03]	7.07 (0.98) [25.65]	7.66 (1.21) [26.62]	7.28 (1.19) [26.00]	0.93
% from consumer list	50.91 (0.86) [50.00]	52.58 (1.32) [49.95]	49.95 (1.61) [50.03]	49.37 (1.62) [50.02]	0.24
% get news or political magazine	9.20 (0.50) [28.91]	9.36 (0.77) [29.13]	8.81 (0.91) [28.36]	9.37 (0.95) [29.15]	0.88
% prefers Democratic candidate for governor in VA	14.43 (0.61) [35.15]	14.53 (0.93) [35.25]	14.61 (1.14) [35.34]	14.11 (1.13) [34.83]	0.94
% no preference in VA governor race	14.82 (0.61) [35.53]	14.18 (0.92) [34.89]	15.54 (1.17) [36.25]	15.05 (1.16) [35.78]	0.63
% in wave 2 of random assignment	37.14 (0.84) [48.32]	36.87 (1.28) [48.26]	37.31 (1.56) [48.39]	37.37 (1.57) [48.40]	0.96
% participating in follow-up survey	32.30 (0.81) [46.77]	31.70 (1.23) [46.55]	32.02 (1.50) [46.68]	33.47 (1.53) [47.21]	0.65
Number surveyed—baseline	3,347	1,432	965	950	

Notes: Standard errors reported in parentheses; standard deviations in brackets. Column 5 reports the *p*-values for chi-squared tests of independence between treatments for each variable. The second through fourth rows (percent voted) apply only to the voter registration (i.e., nonconsumer) sample frame. All regressions in Tables 2–4 include controls for which sample frame provided the observation. A multinomial logit model predicting assignment to treatment using all of the above baseline variables yields a chi-squared test value of 9.21 (d.f. 18, *p*-value of 0.95).

Figure 23: Comparison of baseline characteristics lets the authors conclude that there are not significant differences

Attrition

- Failure to collect outcome data from some individuals who were part of the initial sample used for the randomisation
- If attrition is random : reduces power
- If attrition is correlated with the treatment, may bias estimates
- Authors argue that covariates appear to be orthogonal

TABLE 4—EFFECT OF TREATMENT ON VOTING BEHAVIOR IN VIRGINIA GOVERNORS RACE
(OLS)

	Voted in 2005 election[a] (1)	Voted in 2005 election[b] (2)	Voted in 2006 election[b] (3)	Voted for Democrat (set to missing if did not vote)[a] (4)	Voted for Democrat (set to zero if did not vote)[a] (5)
Panel A: Separate treatment effects estimated for Washington Post *and* Washington Times					
Washington Post treatment	−0.001 (0.033)	0.011 (0.019)	0.025 (0.019)	0.112 (0.045)	0.072 (0.035)
Washington Times treatment	0.005 (0.033)	−0.006 (0.019)	0.031 (0.020)	0.074 (0.045)	0.060 (0.035)
Adjusted R^2	0.21	0.39	0.31	0.31	0.26
F-test: *Post* = *Times*	0.03	0.65	0.10	0.58	0.09
p-value	0.86	0.42	0.75	0.44	0.76
Panel B: Pooled treatment effect estimated for receiving either newspaper					
Received either *Post* or *Times* treatment	0.002 (0.028)	0.003 (0.016)	0.028 (0.016)	0.093 (0.038)	0.066 (0.029)
Adjusted R^2	0.21	0.39	0.31	0.31	0.26
Observation counts for both panels					
Observations	1,079	2,571	2,571	718	1,003
Refused/not asked	2			363	78
Total not merged (columns 2 and 3)		776	776		
Total surveyed in follow-up	1,081			1,081	1,081

Notes: Standard errors in parentheses. The following covariate variables are included in all specifications: gender; reported age; three separate indicators for voting in the 2001, 2002, and 2004 general elections; an indicator for whether the respondent was drawn from a consumer list; self report of receiving any news or political magazines; baseline survey self reports of preferring the Democratic candidate in the gubernatorial election and having no preference in the gubernatorial election; and an indicator for wave of the study. If a covariate value was missing, an indicator variable was included and the covariate was coded as zero. We include strata indicators, which are variables for each strata formed prior to the randomization, which included unique combinations of the following: intention to vote, receive a paper (non-*Post*/non-*Times*), mentions ever reading a paper, gets a magazine, and asked whether they wish they read the paper more. All results remain qualitatively similar, and statistical significance remains as-is, using probit specifications instead of OLS.
Data source:
[a] Survey.
[b] Administrative voting records.

Figure 24: Outcomes

Admin data : 2.8pp higher voter turnout in 2006 if received either paper

Lecture 6: Panel Data

Wed 06 Mar 12:07

[**L6**]

5 Panel Data Methods

Panel Data methods are another way of dealing with the problem of endogeneity

Panel Data

So far we have only seen a cross section of individuals (i)

Now we introduce panel which combines individual and time dimension (t).

We assume the cross section model (t=1) as :

$$Y_{i,1} = \beta_0 + x_{i,1}\beta + \alpha_i + u_{i,1} \tag{8}$$

Where we have previously estimated the impact of education on wages where *ability* is typically an unobserved omitted variable, our solutions so far have been to (a) find an IV and (b) randomise

The additional time dimension gives us new tools.

There are essentially 3 ways of getting rid of a_i, they are

1. First differences
2. Fixed effects estimator
3. Dummy variable regression

Then, assuming the FE is uncorrelated with the regressor we can use random effects approach or pooled OLS.

Intuition FD Assuming that ability is constant over time and does not change, and that we observe the same individual (i) from eq. (8) (t=2)

$$y_{i,2} = \beta_0 + x_{i,2}\beta + a_i + u_{i,2} \tag{9}$$

The idea is to take the difference between both periods of time, to get rid of the *unobserved* constant effect

Taking the difference between both periods

$$y_{i,2} - y_{i,1} = (x_{i,2} - x_{i,1})\beta + u_{i,2} - u_{i,1} \tag{10}$$

Then, if unobserved ability is constant over time, we can get rid of it.

5.1 First Differences

Typical panel data structure for T=2 (two time periods)

Assuming random sample of individuals that we observe twice at $t = 1$ and $t = 2$

i	t	$y_{i,t}$	1	d_t^1	d_t^2	$x_{i,t}$	a_i
1	1	$y_{1,1}$	1	1	0	$x_{1,1}$	a_1
1	2	$y_{1,2}$	1	0	1	$x_{1,2}$	a_1
2	1	$y_{2,1}$	1	1	0	$x_{2,1}$	a_2
2	2	$y_{2,2}$	1	0	1	$x_{2,2}$	a_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	$y_{n,1}$	1	1	0	$x_{n,1}$	a_n
n	2	$y_{n,2}$	1	0	1	$x_{n,2}$	a_n

We also observe an outcome at both time periods, the outcome of individual 1 in both periods and so on. We also have an intercept, and also 2 binary variables that indicate the relevant time periods. Switching "on" for each period, exactly the opposite of each other.

We also have fixed effects a_i that only vary with i, so $a_1, a_2, \ldots, a_n$.

It is important to note we do not observe all of the *fixed effects*.

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \delta_0 d_t^2 + a_i + u_{i,t}$$

Where

1. y_{it} is the outcome of interest, varies over i and t
2. x_{it} is the observed regressor, varying over i and t
3. M d_t^1 (resp d_t^2) is a period 1 (2) dummy varying over t (but only one enters regression - dummy var trap)
4. The unobserved fixed effect a_i only varies over i
5. $u_{i,t}$ is an unobserved idiosyncratic error
6. The time dummy ensures a time varying intercept

Pooled OLS is to estimate a composite error term since we don't observe a_i

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} \delta_0 d_t^2 + \underbrace{v_{i,t}}_{a_i + u_{i,t}}, \; t = 1, 2$$

Then, *taking first differences* to difference out a_i

$$\begin{aligned} y_{i,2} - y_{i.1} &= \delta_0 + \beta_1 (x_{i,2} - x_{i,1}) + u_{i,2} - u_{i,1} \\ \Delta y_i &= \delta_0 + \beta_1 \Delta x_i + \Delta u_i \end{aligned}$$

The change in $y_{i,t}$ between period 1 and 2 is regressed on the change in the regressor(s) and a constant using n observations

i	t	$y_{i,t}$	1	d_t^1	d_t^2	$x_{i,t}$	a_i	$\Delta y_{i,t}$	$\Delta 1$	Δd_t^1	Δd_t^2	$\Delta x_{i,t}$
1	1	$y_{1,1}$	1	1	0	$x_{1,1}$	a_1	.	.	.	.	.
1	2	$y_{1,2}$	1	0	1	$x_{1,2}$	a_1	Δy_1	0	-1	1	Δx_1
2	1	$y_{2,1}$	1	1	0	$x_{2,1}$	a_2	.	.	.	.	.
2	2	$y_{2,2}$	1	0	1	$x_{2,2}$	a_2	Δy_2	0	-1	1	Δx_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	$y_{n,1}$	1	1	0	$x_{n,1}$	a_n	.	.	.	.	.
n	2	$y_{n,2}$	1	0	1	$x_{n,2}$	a_n	Δy_n	0	-1	1	Δx_n

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- New intercept δ_0 : change in the intercept from $t = 1$ to $t = 2$
- We can use the normal framework provided the assumptions are fulfilled. Most importantly that Δx_i and Δu_i are uncorrelated
- This means strict exogeneity : the idiosyncratic error at each time t, u_{it} is uncorrelated with the explanatory variable in both periods
- Allows however x_{it} to be correlated with unobservable s that are constant over time
- If strict exogeneity does not hold, FD is biased and inconsistent
- Need to assume homoskedasticity of ΔU_{it} which means that we can use the normal OLS inference procedures
- Can be extended to more regressors

Variation in Covariates Over t

- Need variation in Δx_i across i
- Even if Δx_i varies only a little : leads to large SE
- Assume we have several regressors : splits covariates into 2 blocks : some change with time and some not

$$Y_{i,t} = x_{i,t}\beta + \delta_0 d_t^2 + a_i + u_{i,t}$$
$$= \underbrace{x_{i,t}^{(1)}}_{\text{changes overt}} + \beta_1 + \underbrace{x_i^{(2)}}_{\text{does not change over t}} \beta_2 + \delta_0 d_t^2 + a_i + u_{i,t}$$

- Taking the first difference:

$$y_{i,2} - y_{i,1} = \delta_0 + (x_{i,2}^{(1)} - x_{i,2}^{(1)})\beta_1 + u_{i,2} - u_{i,1}$$

- The not changing variables disappear
- Only interpretation for variables changing over time

More Time Periods : T=3

$$y_{it} = \delta_0 + \delta_1 d_t^2 + \delta_3 d_t^3 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}\, t = 1, 2, 3$$

Estimate by pooled OLS

$$\Delta y_{it} = \delta_1 \Delta d_t^2 + \delta_3 \Delta d_t^3 + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}\, t = 2, 3$$

- Key assumption Δu_{it} is uncorrelated with Δx_{itj} for all $j = 1, \ldots, k$ and $t = 2, 3$
- We only have $T - 1$ time periods and $T - 1$ differenced time dummies
- Constant is dropped, to include an intercept, include the time dummies starting with d^3 instead of the differenced time dummies

Serial Correlation

- If $T > 2$ we must assume Δu_{it} to be uncorrelated over time for inference to be valid
- Estimated by pooled OLS :
$$\hat{\Delta} U_{i,t} = \rho \hat{\delta} u_{i,t-1} + \varepsilon_{i,t}$$
- Test $H_0 : \rho = 0$
- Potentially make the test robust to heteroskedacity
- If serial correlation suspected : cluster at the cross sectional identifier ("id") level

Drawbacks of FD

- Can be hard to e=collect panel data
- Regression only on n observations : variables are not defined for period 1
- First differencing can reduce variation in covariates
- Only solves *unobserved variable problem* if those are constant over time
- Measurement error can get worse with FD which leads to endogeneity

Lagged Dependent Variables

[L6-LaggedDV]

Consider the dynamic model of wage determination with unobserved heterogeneity:

$$Log(wage_{it}) = \beta_1 \log(wage_{i,t-1}) + c_i + u_{it}\ ,\ t = 1, \ldots, T$$

- β_1 can be considered a measure of how persistent wages are after controlling for unobserved heterogeneity, c_i (e.g. Individual productivity)
- Set $y_{it} = \log(wage_{it})$, hence $y_{it} = \beta_1 y_{i,t-1} + c_i + u_{it}$
- Standard assumption:
$$E\left[u_{it} | y_{i,t-1}, \ldots, y_{i0}, c_i\right] = 0 \tag{11}$$
- Which means all dynamics are captured by the first lag

If we set $x_{it} = u_{i,t-1}$, by **??**, u_{it} is uncorrelated with $(x_{it}, x_{i,t-1}, \ldots, x_{i_1})$ However, u_{it} cannot be uncorrelated with $(x_{it+1}, x_{it+2}, \ldots, x_{iT})$ since $x_{i,t+1} = y_{it}$

$$\begin{aligned} E\left[x_{i,t+1}u_{it}\right] &= E\left[y_{it}u_{it}\right] \\ &= \beta_1 E\left[y_{i,t-1}u_{it}\right] + E\left[c_i u_{it}\right] + E\left[u_{it}^2\right] \\ &= E\left[u_{it}^2\right] > 0 \end{aligned}$$

Because $E\left[y_{i,t-1}u_{it}\right] = E\left[c_i u_{it}\right] = 0$ by eq. (11)

Implications **Strict exogeneity** - this assumption never holds in unobserved effects models with lagged dependent variables

Pooled OLS

- $y_{i,t-1}$ and c_i are necessarily correlated because $y_{i,t-1} = y_{i,t-2} + c_i + u_{i,t-1}$
- $\text{Cov}(y_{i,t-1}, c_i) = \text{Cov}(y_{i,t-2} + c_i + u_{i,t-1}, c_i) \neq 0$
- The standard exogeneity assumption required for pooled OLS is also violated

5.2 Fixed Effects / Within Estimator

Consider the unobserved effects model:

$$y_{i,t} = \beta_1 x_{i,t,1} + \ldots + \beta_k x_{i,t,k} + a_i + u_{i,t} \quad t = 1, \ldots, T$$

Averaging for each i over time, $\hat{s}_i = \frac{1}{T}\sum_{t=1}^{T} s_{i,t}$:

$$\bar{y}_i = \beta_1 \bar{x_{i,1}} + \ldots \beta_k \bar{x_{i,k}} + a_i + \bar{u}_i \tag{12}$$

Then eliminating a_i by demeaning the variables

$$\underbrace{\ddot{y}_{i,t}}_{y_{i,t}-\bar{y}_i} = \beta_1 \underbrace{\ddot{x}_{i,t,1}}_{x_{i,t,1}-\bar{x}_{i,1}} + \ldots + \beta_k \underbrace{\ddot{x}_{i,t,k}}_{x_{i,t,k}-\bar{x}_{i,k}} + \underbrace{\ddot{u}_{i,t}}_{u_{i,t}-\bar{u}_i} \quad t = 1, \ldots, T \tag{13}$$

Within transformation based on time demeaned data: uses the time variation within each cross-sectional observation Estimate by pooled OLS : *Fixed Effects estimator*

FE - Assumptions

Assumption 1. *FE.1*

for each i in the model is

$$y_{i,t} = \beta_1 x_{i,t,1} + \ldots + \beta_k x_{i,t,k} + a_i + u_{i,t} \quad t = 1, \ldots, T \tag{14}$$

Where the β_j are the parameters to estimated and a_i is the unobserved effect

Assumption 2. *FE.2 We have a random sample from the cross section*

Assumption 3. *FE.3 Each explanatory variable changes over time (for at least some i) and no perfect linear relationship exist among the explanatory variables*

Assumption 4. *FE.4 For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero :* $E\left[u_{i,t}|X_i\,, a_i\right] = 0$

- Where X_i contains all X_{itj} for $j = 1, \ldots, k$ and $t = 1, \ldots, T$
- Under **FE.1-4** FE estimator is unbiased and consistent with fixed T and $n \longrightarrow \infty$
- Key is strict **endogeneity** (FE.4)

Assumption 5. *FE.5* $V\left[u_{i,t}|X_i, a_i\right] = V\left[u_{i,t}\right] = \sigma_u^2$ *for all* $t = 1, \ldots, T$

Assumption 6. *FE.6 For all* $t \neq s$*, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and* a_i*) :* $\text{Cov}(u_{i,t}, u_{i,s}|X_i, a_i) = 0$

- Under FE.1-6 : FE estimator of β_j is the best linear unbiased estimator (BLUE)
- Implication : since FD is linear and unbiased, FE is more efficient
- FE.6 implies that the errors are serially uncorrelated

Assumption 7. *FE.7 Conditional on* X_i *and* a_i*, the* $u_{i,t}$ *are iid as* $\mathcal{N}(0,\, \sigma_u^2)$

- This assumption implies FE.4-6 but is stronger as it assumes normality
- This assumption implies that the FE estimator is normally distributed and t and F statistics have exact t and F distributions
- Without FE.7 : rely on asymptotics : large n, small T

In **summary**, to be unbiased under strict exogeneity, the idiosyncratic error should be uncorrelated with each explanatory variable across all time periods.

If the fixed effects are constant over time, they can be correlated with the regressors in any period, but we must be careful since constant regressors are swept away.

For inference, we need homoskedatic and serially uncorrelated error terms.

And, there is a DOF adjustment : for each i, we lose one df because few demeaning and have no intercept. Therefore we have $df = nT - n - k = n(T-1) - k$

Least Squares Dummy Variables Estimator FE estimator by dummy variable regression:

- Traditionally, consider a_i to be estimated for each i
- This represents the intercept for person i
- Approach : add a dummy variable for each observation I
- Without intercept, since $\Sigma_i$individual dummies = 1

i	t	$y_{i,t}$	1	d_t^1	d_t^2	$x_{i,t}$	Individual Dummies			
1	1	$y_{1,1}$	1	1	0	$x_{1,1}$	1	0	...	0
1	2	$y_{1,2}$	1	0	1	$x_{1,2}$	1	0	...	0
2	1	$y_{2,1}$	1	1	0	$x_{2,1}$	0	1	...	0
2	2	$y_{2,2}$	1	0	1	$x_{2,2}$	0	1	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
n	1	$y_{n,1}$	1	1	0	$x_{n,1}$	0	0	...	1
n	2	$y_{n,2}$	1	0	1	$x_{n,2}$	0	0	...	1

FE estimator by dummy variable regression

- Gives exactly he same estimates of the β_j as regression on time demeaned data
- SEs and most stats identical
- Properly computes dof
- One can compute the estimated "intercepts" $\hat{a}_i$

Estimate the FE:

$$\hat{a}_i = \bar{y}_i - \widehat{\beta}_1 \hat{x}_{i,1} - \ldots - \widehat{\beta}_k \hat{x}_{i,k}$$

- Where the $\widehat{\beta}_j$ $j = 1, \ldots, k$ are the FE estimates
- Directly available from the dummy variable regression
- Inconsistent with fixed T but unbiased

FE vs FD

- T = 2
  - FE and FD are identical if the same model
  - For FD easy to compute heteroskedacity robust statistics
- T> 2
  - Both unbiased and consistent for T fixed and $n \longrightarrow \infty$. Therefore compare efficiency.
  - No serial correlation of $u_{i,t}$: FE is more efficient
  - If the FD error is serially uncorrelated (but not $u_{i,t}$), then FD is more efficient
  - In-between : not always easily comparable efficiency
  - Good idea to try both and check

Between Estimator Pooled estimator based on the averaged equation 12 including a constant

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

However, this is rarely used since it is biased if a_i is correlated with $\bar{x}_i$ and if they are not correlated one would use random effects

5.3 Random Effects

Model $y_{i,t} = \beta_0 + \beta_1 x_{i,t,1} + \ldots + \beta_k x_{i,t,k} + a_i + u_{i,t}$ Where they key assumption is

$$\text{Cov}(x_{i,t,j}, a_i) = 0 \; t = 1, \ldots, T \; , j = 1, \ldots, k$$

- Where a_i is uncorrelated with each regressor in all t
- The intercept is included to ensure zero mean of a_i
- Allows for time dummies
- OLS on a single cross section is consistent but discards additional information from other t's
- Pooled OLS consistent but ignores knowledge about functional form

A more efficient estimator uses additional information about the structure of unobserved variables: GLS

Composite error term Put a_i in the error term

$$\nu_{i,t} = a_i + u_{i,t}$$

Resulting in the following model

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t,1} + \ldots + \beta_k x_{i,t,k} + \nu_{i,t} \tag{15}$$

Where $\nu_{i,t}$ are serially correlated across time

- Pooled OLS standard errors ignore this correlation and are incorrect
- Accounting for the error structure: apply GLS
- Requires n large and T fixed : can be extended to unbalanced panels

In addition to Assumption 1 (model) Assumption 2 (random sample) Assumption 4 (exogeneity) Assumption 5 (time constant variance) and Assumption 6 (no serial correlation).

Replace with the following:

Assumption 8. *RE.1 There are no perfect linear relationships among the explanatory variables*

Assumption 9. *RE.2 In addition to Assumption 4, the expected value of a_i, given all explanatory variables, is constant $E[a_i|X_i] = \beta_0$*

The key difference being that Assumption 9 rules out the correlation between the unobserved effect and the regressors

Assumption 10. *RE.3*

In addition to Assumption 5, the variance of a_i given all explantory variables is constant

$$V[a_i|X_i] = \sigma_a^2$$

- Under Assumption 1, Assumption 2, Assumption 8, Assumption 9 (includes Assumption 4), Assumption 10 (includes Assumption 5), Assumption 6, the RE estimator is consistent and asymptotically normally distributed for n large and fixed T
- *First four*: consistency and asymptotic normality
- Last two: inference
- Under all RE assumptions the RE estimator is asymptotically efficient

Variance-Covariance Structure Define, for $s \neq t$:

$$\begin{aligned} \text{Cov}(\nu_{i,t}, \nu_{i,s}) &= \sigma_a^2 \\ V[\nu_{i,t}] &= \sigma_a^2 + \sigma_u^2 \\ \text{Cor}(\nu_{i,t}, \nu_{i,s}) &= \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} \end{aligned}$$

The resulting error structure is serially correlated over time

GLS transformation

Applying matrix algebra

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}} \quad \text{where } 0 < \theta < 1$$

And the transformed equation:

$$y_{i,t} - \theta\bar{y}_i = \beta_0(1-\theta) + \beta_1(x_{i,t,1} - \theta\bar{x}_{i,1}) + \ldots + \beta_1(x_{i,t,k} - \theta\bar{x}_{i,k}) + (\nu_{i,t} - \theta\bar{\nu}_i) \tag{16}$$

So, GLS is pooled OLS on the transformed equation. Intuitively we have quasi-demeaned the data by θ

In order to estimate θ, we need to estimate σ_u^2 and σ_a^2. This can be done based on the FE or pooled OLS or even on the between estimator

Approach

- Estimate $\hat{\nu}_{i,t}$ by pooled OLS in **??** and obtain the resulting $\widehat{\sigma}_\nu^2$
- Estimate

$$\widehat{\sigma}_a^2 = [nT(T-1)/2 - (k+1)]^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{\nu}_{i,t}\hat{\nu}_{i,t}$$

- Estimate $\widehat{\sigma}_u^2 = \widehat{\sigma}_\nu^2 - \widehat{\sigma}_a^2$
- Estimate $\hat{\theta}$ using $\widehat{\sigma}_a^2$ and $\widehat{\sigma}_u^2$ and eq. (16) by pooled OLS

The resulting random effects estimator is feasible GLS

Notes on RE

- Allows for time-varying and constant data
- Under the assumptions, it is consistent for large n but fixed T (but not unbiased)
- $\theta = 1$: GLS = FE. One can see that as $T \longrightarrow \infty$, $\theta \longrightarrow 1$. So GLS gets close to FE
- $\theta = 0$: GLS = Pooled OLS

5.4 FE vs RE

Hausman Test

$$H_0 : a_i \text{ is not correlated with } x_{i,t,j} \quad \text{vs.} \quad H_1 : a_i \text{ is correlated with } x_{i,t,j}$$

$$\text{Under } H_0 : \left.\begin{array}{l} \hat{\beta}^{RE} \xrightarrow{H_0} \beta \\ \hat{\beta}^{FE} \xrightarrow{H_0} \beta \end{array}\right\} \quad \hat{\beta}^{FE} - \hat{\beta}^{RE} \xrightarrow{H_0} 0$$

$$\text{Under } H_1 : \left.\begin{array}{l} \hat{\beta}^{RE} \overset{H_1}{\nrightarrow} \beta \\ \hat{\beta}^{FE} \overset{H_1}{\nrightarrow} \beta \end{array}\right\} \quad \hat{\beta}^{FE} - \hat{\beta}^{RE} \overset{H_0}{\nrightarrow} 0$$

$$\text{Under } H_0 : \quad \left[\hat{\beta}^{FE} - \hat{\beta}^{RE}\right]' \left[\widehat{\text{AV}}(\hat{\beta}^{FE}) - \widehat{\text{AV}}(\hat{\beta}^{RE})\right]^{-1} \left[\hat{\beta}^{FE} - \hat{\beta}^{RE}\right] \xrightarrow{d} \chi^2_m$$

Where m is the number of elements in $\widehat{\beta}^{FE}$ and $\widehat{\beta}^{RE}$

- The Hausman test statistic compares RE and FE estimators
- In some cases, the variance might not be invertible
- If we are interested in a single regressors, then we cane state the test statistic for a single β_j
- We rely on strict exogeneity for $x_{i,t,j}$ with respect to $u_{i,t}$ under H_0 and H_1
- Can only compare coefficients on time-varying variables as FE can only include those

5.5 Application

Verbeek (2012) : *whose wage do unions raise? a dynamic model of unionism and wage rate determination for young men*

Using a sample of 545 full time working males who completed their schooling by 1980, followed over 1980-87. With the outcome log wages and covariates including years of schooling, years of experience, being a union member, public sector and being married

```
Call:
lm(formula = lwage ~ school + exper + expersq + union + married +
    pub, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2500 -0.2509  0.0330  0.2962  2.5788

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0336849  0.0633445  -0.532 0.594910
school       0.0989574  0.0046272  21.386  < 2e-16 ***
exper        0.0860881  0.0101514   8.480  < 2e-16 ***
expersq     -0.0027313  0.0007102  -3.846 0.000122 ***
union        0.1681800  0.0171598   9.801  < 2e-16 ***
married      0.1229150  0.0155811   7.889 3.83e-15 ***
pub          0.0072708  0.0376254   0.193 0.846779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4829 on 4353 degrees of freedom
Multiple R-squared:  0.179,     Adjusted R-squared:  0.1779
F-statistic: 158.2 on 6 and 4353 DF,  p-value: < 2.2e-16
```

Figure 25: Pooled OLS

Where the dependent variable is log wage

```
Call:
plm(formula = lwage ~ school + exper + expersq + union + marrie
    pub, data = data, model = "random", index = c("nr",
    "year"))

Balanced Panel: n = 545, T = 8, N = 4360

Effects:
                 var std.dev share
idiosyncratic 0.1234  0.3513 0.536
individual    0.1070  0.3271 0.464
theta: 0.645

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-4.57041 -0.14525  0.02313  0.18511  1.54085

Coefficients:
              Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -0.11570236  0.10728871 -1.0784    0.2808
school       0.10101175  0.00878462 11.4987 < 2.2e-16 ***
exper        0.11113480  0.00827097 13.4367 < 2.2e-16 ***
expersq     -0.00403377  0.00059214 -6.8121 9.615e-12 ***
union        0.10316121  0.01785117  5.7790 7.516e-09 ***
married      0.06656405  0.01674013  3.9763 6.999e-05 ***
pub          0.03102262  0.03649394  0.8501    0.3953
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    655.79
Residual Sum of Squares: 540.06
R-Squared:      0.17647
Adj. R-Squared: 0.17534
Chisq: 932.799 on 6 DF, p-value: < 2.22e-16
```

Figure 26: Random Effects

```
Call:
plm(formula = lwage ~ school + exper + expersq + union + married +
    pub, data = data, model = "within", index = c("nr",
    "year"))

Balanced Panel: n = 545, T = 8, N = 4360

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-4.1724327 -0.1256540  0.0099207  0.1589191  1.4701775

Coefficients:
           Estimate  Std. Error t-value  Pr(>|t|)
exper    0.11645699  0.00843090 13.8131 < 2.2e-16 ***
expersq -0.00428857  0.00060544 -7.0834 1.668e-12 ***
union    0.08120303  0.01931592  4.2039 2.683e-05 ***
married  0.04510613  0.01831141  2.4633   0.01381 *
pub      0.03492672  0.03860819  0.9046   0.36571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 470.1
R-Squared:      0.17822
Adj. R-Squared: 0.059807
F-statistic: 165.256 on 5 and 3810 DF, p-value: < 2.22e-16
```

Figure 27: Fixed Effects

Where the time invariant variables are dropped, should we drop experience also?

```
> phtest(fixed_effects_plm, random_effects_plm)

        Hausman Test

data:  lwage ~ school + exper + expersq + union + married + pub
chisq = 33.866, df = 5, p-value = 2.532e-06
alternative hypothesis: one model is inconsistent
```

Figure 28: Hausman Test

Explaining Individual Wages Rejects the null of fixed effects so Random effects it is. Since the null hypothesis is that the preffered model (RE) is consistent .

- Marital status likely to be correlated with unobserved FE
- Typically captures other unobservable differences between married and unmarried workers
- It is confirmed by FE regression that the effect of being married reduces to 4.6%

```
> p_m = 100*(exp(fixed_effects_plm$coefficients[4])-1)
> p_m
 married
4.613888
```

- Effect only identified through people who change marital status
- Similar for union status
- Others have concentrated on the impact of endogenous union status on wages for this group of workers and consider alternative, more complicated estimators

Tutorial

[**Tutorial 5**]

1. Interpretation of coefficient, 0.38% log-log
2.
3.
4. Use fixed effects since . . . Demean

Lecture 7: Differences-in-Differences

Thu 07 Mar 15:03

[**L7-DID**]

6 Differences-in-differences

In many applications, we want to estimate the effect of a policy across groups. However, the policy assignment is not necessarily uncorrelated with group characteristics. In this case, we can use a DiD to identify the effect of the policy without being confounded by these level differences.

To compare the difference between two groups before and after a change to establish causality

Example. *John Snow Cholera London Cholera epidemic in London, he wanted to establish that Cholera was transmitted through contaminated drinking water Since districts were served by two water companies, To compare the difference between 2 groups before and after a change to establish causality In 1849, both companies obtained their water supply from the dirty Thames In 1852 Lambeth company moved its water works upriver to an area less contaminated with sewage Death rates fell in districts supplied by Lambeth compared to the change in death rates in districts supplied by Southwark and vauxhall*

Canonical DiD

- Data with a time dimension, at least 2 repeated cross sections
- An exogenous treatment : no self selection into the event and no change of behaviour in anticipation of the event
- Measure the outcomes of interest before and after the event
- Two groups, one impacted by the event and the other not

One can introduce more groups, time periods and covariates to the *canonical* 'simple 2x2' model.

Simple Setup in Basic 2x2 Diff-in-diff Assume we have n units (i) and T = 2 time periods (t). Consider a binary policy D_{it} and we are interested in estimating its effect on outcomes Y_{it}. Consider the potential outcome notation for Y_{it} :

- $Y_{it}(0,0)$ is the potential outcome in period $t \in \{1,2\}$ if untreated in both periods
- $Y_{it}(0,1)$ is the outcome in period $t \in \{1,2\}$ if untreated in first period, treated in second

The inherent problem is that D_{it} is not necessarily randomly assigned, but we still want to estimate the ATT in period 2:

$$\tau_2^{ATT} = E\left[Y_{i,2}(1) - Y_{i,2}(0)|D_i = 1\right]$$

In order to identify the ATT, we have to assume:

1. Parallel trends: "in the absence of the treatment, the average outcomes would have evolved in parallel"
$$E\left[Y_{i,2}(0) - y_{i,1}(0)|D_i = 1\right] = E\left[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0\right]$$
   - Absent the policy, units may have different *levels*, but their changes would be the same
   - A sufficient parametric formulation: $Y_{i,t}(0) = \gamma_l + \alpha_i + \varepsilon_{it}$
2. No-anticipation: policy has no effect prior to treatment
$$Y_{i,1}(0) = Y_{i,1}(1)$$

Where out typical estimand of interest is the ATE or the ATT:

$$\tau_{ATE} = E\left[Y_{it}(1) - Y_{it}(0)\right] = E\left[\tau_i\right]$$
$$\tau_{ATT} = E\left[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1\right] = E\left[\tau_i|D_{it} = 1\right]$$

Since D is not randomly assigned, this model is inherently not identified without the additional assumptions (and two time periods)

Since D_i could be correlated with α_i, recall that our plug-in estimator approaches need estimates for $E\left[Y_{it}(1)\right]$ and $E\left[Y_{i,t}(0)\right]$. But the correlation prevents this without the conditional exogeneity assumptions.

2 x 2 DID estimation

	t =0	t = 1
$D = 0$	$\gamma_0 + \alpha_i$	$\gamma_1 + \alpha_i$
$D = 1$	$\gamma_0 + \alpha_i + \tau_i$	$\gamma_1 + \alpha_i + \tau_i$

The within unit difference :

$$Y_{i1} - Y_{i0} = (\gamma_1 - \gamma_0) + \tau_i (D_{i1} - D_i(0))$$

Hence

$$E[Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 1] - E[Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 0] = E[\tau_i|D_{i1} - D_{i0} = 1]$$

Where the simplifying assumption is that treatment only goes 1 way in period 1, the "absorbing adoption", e.g. $D_{i0} = 0$

$$E[Y_{i1} - Y_{i0}|D_{i1} = 1] - E[Y_{i1} - Y_{i0}|D_{i1} = 0] = \underbrace{E[\tau_i|D_{i1} = 1]}_{\text{ATT}}$$

Or we can rewrite the parallel trends assumption:

$$E(Y_{i,2}(0)|D_i = 1) = E(Y_{i,1}(0)|D_i = 1) + E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0)$$

Where in other words, the counterfactual "untreated" state is the untreated outcome in the pre-period for the treated group, plus the change from the other untreated group.

Then, thanks to no-anticipation, we can replace $E[Y_{i,1}(0)|D_i = 1]$ with $E[Y_{i,1}(1)|D_i = 1]$, which has an empirical analog:

$$E[Y_{i,2}(0)|D_i = 1] = E[Y_{i,1}|D_i = 1] + E[Y_{i,2} - Y_{i,1}|D_i = 0]$$

So parallel trends and the no anticipation generates our counterfactual outcome

Estimation using linear regression A simple linear regression will identify $E[\tau_i|D_i = 1]$ with two time periods:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta + \varepsilon_{it}$$

Which is sometimes referred to as the two-way fixed effects (TWFE) estimator Note, we can also have estimated τ directly:

$$\hat{\tau} = n^{-1} \sum_i \underbrace{D_i (Y_{i1} - Y_{i0})}_{\Delta\bar{Y}_1} - \underbrace{(1 - D_i)(Y_{i1} - Y_{i0})}_{\Delta\bar{Y}_0}$$

Intuitively, we generate a counterfactual for the treatment using the changes in untreated units: $E[Y_{i1} - Y_{i0}|D_i = 0]$

6.1 Differences-in-differences - Simple Case

Following frolick and sperling (2019)

Considering the arrival of a large number of refugees in one city. The idea is to *estimate the impacts of this increase in refugees on local markets*, employment Supposing we have data on an outcome variable Y for a time period t after the influx of refugees for a time period $t-1$, before the influx of refugees. The immigrants arrive some time between $t-1$ and t. Thus the before-after-difference is $Y_t - Y_{t-1}$

Then, if the time periods are far apart, it may be that other changes have an impact during this time. Then, we can subtract the time trend that *would have happened* if no influx of refugees had occurred. With unaffected neighbouring regions helping us to identify this *unobserved trend*.

We have data for city A (arrival) and B (no arrival)

We compare the differences:

$$\begin{aligned}\Delta Y_{t,A} - \Delta Y_{t,B} &= \underbrace{(Y_{t,A} - Y_{t-1,A})}_{\text{diff over time}} - \underbrace{(Y_{t,B} - Y_{t-1,B})}_{\text{diff over time}} \\ &= \underbrace{(Y_{t,A} - Y_{t,B})}_{\text{diff between cities}} - \underbrace{(Y_{t-1,A} - Y_{t-1,B})}_{\text{diff between cities}}\end{aligned} \tag{17}$$

Where taking the differences in the 'differences' over time is the same as taking the differences in the differences between cities

The idea is to use the changes of the outcomes in the control groups to construct the counterfactual outcome for the treated

We assume the *common trends* and *SUTVA* assumptions.

Potential Outcomes

Define $D = 1$, if city A and $D = 0$ if city B. Let $d_t = 1$ if $t = 1$ and $d_t = 0$ if $t = 0$. We denote the potential outcomes $Y_t(1)$ or $Y_t(0)$, where an observation is treated if $D = 1$ and $d_t = 1$. In $t = 0$, both groups do not receive treatment.

The observed outcome is a linear function of t, D and the Potential Outcomes:

$$Y_t = Y_t(1) \cdot d_t + Y_t(0)\,(1 - D \cdot d_t) \tag{18}$$

$$= \begin{cases} y_{t=1} = Y(1)_{t=1}D + Y(0)_{t=1}(1 - D), & \text{if } t = 1 \\ Y_{t=0} = Y(0)_{t=0} & \text{if } t = 0 \end{cases} \tag{19}$$

Assumption 1. *Common Trends Assumption (CT)*

During the period $[t-1, t]$*, the potential non-treatment outcomes* $Y(0)$ *followed the same linear trend in the treatment group as in the control group:*

$$E[Y(0)_{t=1} - Y(0)_{t-0}|D = 1] = E[Y(0)_{t=1} - Y(0)_{t=0}|D = 0]$$

or the Parallel trend or Parallel Path

Proof 1. *DiD identifies ATT*

Recalling that we have only treated individuals in $t = 1$ *therefore in* $t = 1$*,*

$$\tau = E\left[Y(1)_{t=1} - Y(0)_{t=1}|D = 1\right]$$

while we can also identify $E\left[Y(1)_{t=1}|D = 1\right] = E\left[Y_{t=1}|D = 1\right]$*, the CT assumption helps us to identify* $E\left[Y(0)_{t=1}|D = 1\right]$ *Then, rearranging CT:*

$$E\left[Y(0)_{t=1}|D = 1\right] = E\left[Y(0)_{t=0}|D = 1\right] + E\left[Y(0)_{t=1} - Y(0)_{t=0}|D = 0\right]$$

then, by equation 1

$$= E\left[Y_{t=0}|D=1\right] + E\left[Y_{t=1} - Y_{t=0}|D=0\right]$$

hence,

$$\tau = E\left[Y_{t=1} - Y_{t=0}|D=1\right] - E\left[Y_{t=1} - Y_{t=0}|D=0\right]$$

[Slide]

Toy proof?

Then, taking the sample analog of the conditional expectations:

$$\hat{\tau} = \hat{E}\left[Y_{t=1} - Y_{t=0}|D=1\right] - \hat{E}\left[Y_{t=1} - Y_{t=0}|D=0\right]$$
$$= \hat{E}\left[Y|D=1, t=1\right] - \hat{E}\left[Y|D=1, t=0\right] - \left\{\hat{E}\left[Y|D=0, t=1\right] - \hat{E}\left[Y\right.\right.$$

That is, to estimate the Did estimator, we only need 4 data points.

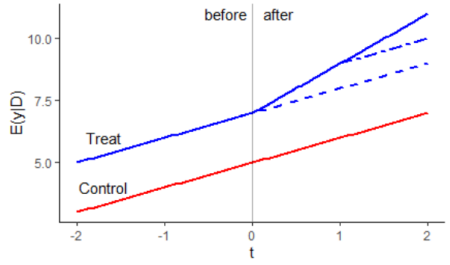


Figure 29: Illustration of Common Trends

Common Trends Extension

- The CT assumption might sometimes be hard to argue for
- In some cases more credible to hold, conditional on confounders
- That is, matching DiD or conditional DiD
- Main assumption:

For confounders not affected by the treatment, ie $X(0) = X(1) = X$, we have

$$E\left[Y(0)_{t=1} - Y(0)_{t=0}|X, D=1\right] = E\left[Y(0)_{t=1} - Y(0)_{t=0}|X, D=0\right]$$

6.2 Regression

We can obtain the AT estimator by linear regression, including the interaction term

$$Y_{i,g,t} = \beta_0 + \gamma D_g + \delta d_t + \tau D_g \cdot d_t + u_{i,g,t}$$

Where i would be a person, family, firm, school. Belonging to a pair (g, t), that could represent a city, state, county

Alternative Representation

There is an alternative way of writing to represent the potential non-treatment outcome $Y(0)$ as

$$Y_{i,g,t}(0) = \beta_0 + \delta d_t + \gamma D_g + u_{i,g,t}(0)$$

Where treatment status is defined as $W)_{g,t} = D_g \cdot d_t$

Then, if we are interested in the ATT, then we do not need a model for $Y(1)$ because

$$E\left[Y(1) - Y(0)|W_i = 1\right] = E\left[Y|D = 1, t = 1\right] - E\left[Y|D = 1, t = 0\right] - \{E\left[Y|D = 0, t = 1\right] - E\left[Y|D = 0, t = 0\right]\}$$

Including covariates With time invariant covariates that treatment does not affect, controlling for covariates is conditional parallel trends assumption (although this isn't enough to satisfy the TWFE regression approach in 2 periods)

$$Y_{i,g,t} = \beta_0 + \gamma D_g + \delta d_t + \tau D_g \times d_t + X_{i,g,t}\theta + u_{i,g,t}$$

Where $X_{i,g,t}$ can include individual level characteristics as well as time varying variables at the group level And, individual level covariates can increase precision

6.3 Multiple Groups And Time Periods

Multiple Time Periods in Basic Setup Consider a policy that occurs all at t_0

More time periods helps in several ways:

1. If we have multiple periods before the policy implementation, we can partially test the underlying assumptions (this is sometimes referred to as "pre-trends")
2. If we have multiple time periods after the policy implementation, we can examine effect timing
   - Is it an immediate effect? Does it die off? Is it persistent?
   - If we pool all time periods together into one "post" variable this estimates the average effect. But if the sample is not balanced, this can have unintended effects

To implement:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1 t \neq t_0}^{T} \delta_t D_{it} + \varepsilon_{it}$$

However, one of the coefficients is fundamentally unidentified because of α_i. So, all coefficients measure the effect *relative to period* t_0

But, implicit here is a strong assumption about trends We assumed that $Y_{it}(d) - Y_{i,t-k}(d) = \gamma_t - \gamma_{t-k}$, which is testable pre treatment (pre-test)

General framework :

- Policy intervention at group level
- I belongs to a pair (g,t)
- There should be a before and after period for at least some of the groups

- Switch treatment definition. Define treatment now as $W_{g,t}$:

$$W_{g,t} = \begin{cases} 1 & \text{if group g in year t is subject to intervention} \\ 0, & \text{otherwise} \end{cases}$$

We estimate this by pooled OLS

$$y_{i,g,t} = \delta_t + \gamma_g + \beta W_{g,t} + X_{i,g,t}\theta + u_{i,g,t}$$
$$g = 1, \ldots, G \; ; \; t = 1, \ldots, T$$

- Outcome and covariates measured at unit level
- δ_t is the aggregate time effect, include time dummies d_t for each t
- γ_g is the group effects. Include dummies for each group d_g
- In practice, intercept is included and one of the time and group dummies are excluded

6.4 Application

Card and Krueger (1994) study the impact of New Jersey increasing the minimum wage from 4.25 to 5.05 dollars an hour on April 1, 1992 The key question is what *impact does this have on employment?*

To answer this, we need a counterfactual for NJ, so use Pennsylvania as a control

- 2 periods : Feb 1992 and Nov 1992
- Data in fast food restaurants in each period in both states
- D_i is NJ vs PA, and $t = 0$ is Feb 1992 and $t = 1$ is Nov 1992
- The outcome : employment at restaurant i in state g in year t
- Analysis : compare the difference November - Feb change of employment in NJ to the difference in Pennsylvania
- Under adequate assumption, this can recover the causal effect of the policy change

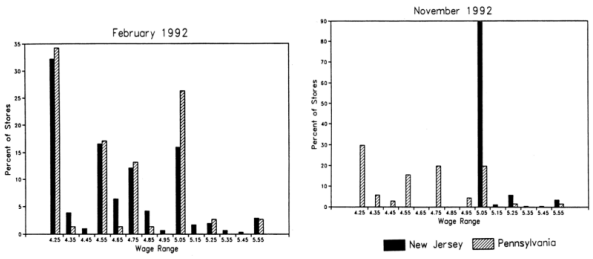


Figure 30: *Stark Effect on Wages in Card And Krueger (1994)*

TABLE 5.2.1
Average employment in fast food restaurants before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (.94)	21.03 (.52)	−.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	.59 (.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), table 3. The table reports average full-time-equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all restaurants with data on employment. Employment at six closed restaurants is set to zero. Employment at four temporarily closed restaurants is treated as missing. Standard errors are reported in parentheses.

Figure 31: Source: table from Angrist and Pischke (2009), chapter 5.

However, despite a large increase in wages, seemingly no negative impact on employment, and in fact, marginally significant positive impact.

Looking at the raw data, the positive impact is driven by a -2.16 decline in PA, which is reasonable if you think that PA is a good counterfactual, since 1992 is in the middle of a recession.

- The contradiction with economic theory seems to have led to further investigation
- In a follow up study, they obtained additional payroll data and included more periods before the treatment
- In 1996, the federal minimum wage increased to $4.75 while the min wage in NJ stayed at 5.05
- A new policy experiment

A second comparison can be run with stores whose starting wage in pre-period was abode the treatment cutoff, where these stores perform similarly to PA

Stores in New Jersey[a]		
Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)
19.56 (0.77)	20.08 (0.84)	22.25 (1.14)
20.88 (1.01)	20.96 (0.76)	20.21 (1.03)
1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)

Figure 32: Above Minimum Wage Increase Stores

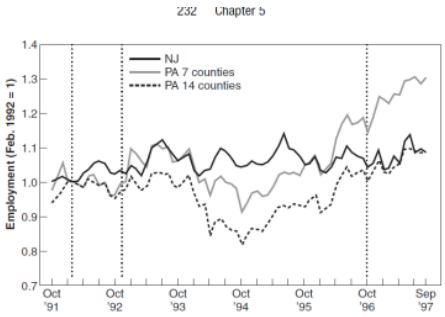


Figure 5.2.2 Employment in New Jersey and Pennsylvania fast food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum wage increase.

Figure 33: Figure: Source: table from Angrist and Pischke (2009), chapter 5.

Key considerations However, the treatment cannot really be thought of as randomly assigned, since:

- Treatment is completely correlated within states
- As a result, any within-state correlation of errors will be correlated with treatment status

Given the limited number of states, time periods, and treatments, it is more valuable to view this as a case study. Under strong parametric assumptions we can infer causality.

Leads And Lags If we have several pre-treatment periods, it is common to use an estimation strategy to include leads and lags

- Autor 2003 - whether increased employment protection affects firm's use of temporary help
- US labour law usually allows workers to be hired and fired at will
- Some states have allowed exceptions leading to lawsuits for unjust dismissal
- Autor wants to understand whether fear of employee law suits makes firms more likely to use temporary workers instead of hiring in workforce
- Identification: uses dummy variables to indicate state court rulings that allow exceptions to the employment-at-will doctrine and then assesses their effect on the use of temporary workers
- Includes leads and lags: 2 year ahead and 4 years behind

$$y_{i,g,t} = \delta_t + \gamma_g + \sum_{\tau=0}^{m} \beta_{-\tau} W_{g,ti\tau} + \sum_{\tau=1}^{q} \beta_{+\tau} W_{g,t+\tau} + X_{i,g,t}\theta + u_{i,g,t}$$

Where sums allow for m-lags, posttreatment effects or q leads anticipatory effects

Example

Example. *Policy: Providing Additional Financial Resources to Poorly Performing Schools*

Difference-in-Differences (DiD) Method:

- *Compares average school outcomes between treated and control schools before and after the intervention.*
- *School outcomes are measured in the same manner before and after the intervention (with different pupils).*
- *Schools are selected for treatment based on average performance of their pupils, specifically those below a certain threshold.*

Analysis of Treatment Effect:

- ***Positive or Negative Treatment Effect?***
  - *Positive.*
  - *In the control group, student performance improves over time (upward trend), with observations taken at $t = 0$ and extrapolated to $t = 1$.*
  - *For the treatment group, performance initially drops just before the intervention. After the treatment, their performance should average out to be similar to the control group. Misidentifying the treatment effect could occur if this trend is not correctly accounted for.*
  - *The parallel trends assumption is violated just before the treatment.*

Is the Common Trend Assumption Likely to Hold?

- *No, the common trend assumption is likely violated due to the drop in performance just before the treatment in the treatment group.*

Questions:

1. *How can the violation of the parallel trends assumption impact the validity of the DiD method?*
2. *What alternative methods can be used if the common trend assumption does not hold in the DiD analysis?*
3. *What additional data could help in more accurately measuring the treatment effect of financial resources on school performance?*

Lecture 8: Regression Discontinuity Designs

Tue 12 Mar 16:19

7 Regression Discontinuity Design

[L8-RDD] [cont]

Motivation

Regression discontinuity exploits precise knowledge of the rules determining treatment

Example such as the minimum legal drinking age, to determine the **cause effect** of legal access to alcohol on death rates (Angrist and Pischke 2014)

Suppose X_i is the individuals drinking age, x_0 the threshold (the MCDA) and D_i is legal drinking

$$D_i = \begin{cases} 1 & , \text{ if } X_i \geqslant 21 \\ 0 & , \text{ if} X_i < 21 \end{cases}$$

3 Requirements

- A score or running variable (e.g. age)
- A cutoff or threshold (e.g. 21)
- A treatment (e.g. Legal drinking)

2 different styles Sharp

- Everyone whose score is above (or below) the threshold receives treatment
- Selection on observable threshold / score

Fuzzy

- Imperfect compliance with the treatment assignment
- Leads to an IV type setup

Essentially, there are two ways of looking at the mechanism,

1. The threshold acts like a random assignment mechanism: an individual is by chance right above or below x_0
2. The threshold creates a local instrumental variable: an instrument only valid at or around the threshold

RD only provides identification around the threshold x_0

7.1 Sharp Design

Treatment status D_i is a deterministic and discontinuous function of a covariate (score) X_i Where x_0 is a known threshold

$$D_i = \begin{cases} 1 & , \quad if X_i > x_0 \\ 0 & , \quad if X_i < x_0 \end{cases}$$

Once we know X_i, we know D_i. With potential outcomes:

$$Y_i = \begin{cases} Y_i(1) & \text{if} D_i = 1 \\ Y_i(0) & , \quad \text{if} D_i = 0 \end{cases}$$

Treatment is a discontinuous function of X because no matter how close $X = x$ gets to x_0, treatment is unchanged until $X = x_0$

$$\lim_{\varepsilon \to 0} E\left[D | X = x_0 + \varepsilon\right] - \lim_{\varepsilon \to 0} E\left[D | X = x_0 + \varepsilon\right] = 1$$

Assumption 1. *RDD-2* $E\left[Y(d)|X = x\right]$ *is continuous in* x *at* x_0 *for* $d \in 0, 1$ *Potential outcomes are essentially the same on both sides of the threshold and can be violated if other things happen at the threshold. Sometimes, one finds a stronger assumption*

$$Y_i(d) \perp X_i \ , \ inear x_0$$

Based on this assumption, we can identify the potential outcomes as follows:

$$E\left[Y(1)|X=x_0\right] = \lim_{\varepsilon\to 0} E\left[Y(1)|X=x_0+\varepsilon\right] = \lim_{\varepsilon\to 0} E\left[Y|X=x_0+\varepsilon\right]$$

similarly

$$E\left[Y(0)|X=x_0\right] = \lim_{\varepsilon\to 0} E\left[Y(0)|X=x_0-\varepsilon\right] = \lim_{\varepsilon\to 0} E\left[Y|X=x_0-\varepsilon\right]$$

Hence, sharp RD identifies the ATE = ATT at (or near) x_0 :

$$\tau(x_0) = E\left[Y(1)-Y(0)|X=x_0\right] = \lim_{\varepsilon\to 0} E\left[Y|X=x_0+\varepsilon\right] - \lim_{\varepsilon\to 0} E\left[Y|X=x_0-\varepsilon\right]$$

Estimation - OLS

Assuming linearity of the regression functions, we can use OLS To do so, we estimate the regression to the left ($X < x_0$) and to the right of the threshold ($X \geqslant x_{00}$):

$$Y = \beta_{0,l} = \beta_{1,l}(X-x_0) + u$$
$$Y = \beta_{0,r} = \beta_{1,r}(X-x_0) + u$$

Centre the running variable around the cut-off. Taking the difference between the intercepts gives the treatment effect $\tau = \beta_{0,l} - \beta_{0,r}$

Pooled OLS The pooled regression yields the same τ:

$$Y = \beta_{0,l} + \tau D + \beta_1(X-x_0) + u$$

Near the cutoff (switching D on and off) :

$$\lim_{\varepsilon\to 0} E\left[Y|X=x_0+\varepsilon\right] = \beta_{0,l} + \tau \quad \text{and} \lim_{\varepsilon\to 0} E\left[Y|X=x_0-\varepsilon\right] = \beta_{0,l}$$

The effect of the policy is the jump at the cutoff:

$$\tau = \lim_{\varepsilon\to 0} E\left[Y|X=x_0+\varepsilon\right] - \lim_{\varepsilon\to 0} E\left[Y|X=c_0-\varepsilon\right]$$

This assumes the same slope above and below the cutoff

Functional Forms

- There is no value of the score at which you observe both treatment and control observations
- RD relies on extrapolation across covariate values
- Consequently, we cannot be agnostic about the functional form of the regression. Possible generalisation
  - Allow for the slopes on the right and left of the threshold to differ
  - Use more flexible specifications including non linear relationships
  - Use non-parametric estimation

Sharp And Linear Include interaction terms

$$Y = \beta_0 + \tau D + \beta_1(X - x_0) + \gamma_1 D \cdot (X - x_0) + u$$

Effect far away from cutoff

$$E[Y|X \geqslant x_0] = \beta_0 + \tau + (\beta_1 + \gamma_1)(X - x_0)$$
$$E[Y|X < x_0] = \beta_0 + \beta_1(X - x_0)$$

Near the cutoff:

$$\lim_{\varepsilon \to 0} E[Y|X = x_0 + \varepsilon] = \beta_0 + \tau \text{ and } \lim_{\varepsilon \to 0} E[Y|X = x_0 + \varepsilon] = \beta_0$$

The effect at the cut-off is still τ:

$$\tau = \lim_{\varepsilon \to 0} E[Y|X = x_0 + \varepsilon] - \lim_{\varepsilon \to 0} E[Y|X = x_0 + \varepsilon]$$

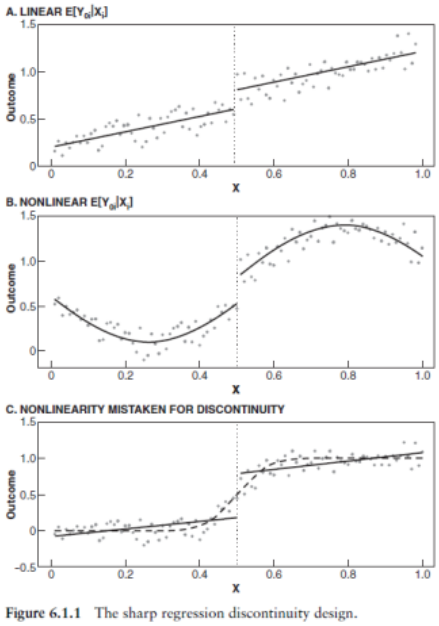


Figure 6.1.1 The sharp regression discontinuity design.

Figure 34: Non Linearities

One must be careful not to confuse a non-linear relationship with a discontinuity

Parametric RD Include higher order polynomials to account for non linearities, eg cubic polynomials

$$Y = \beta_0 + \tau D + \beta_1 g(X - x_0) + u$$

Where $g(\cdot)$ is a higher order (cubic) polynomial:

$$Y = \beta_0 + \tau D + \beta_1(X - x_0) + \beta_2(X - x_0)^2 + \beta_3(X - x_0)^3 + u$$

And we can allow for different slopes, e.g. a quadratic polynomial such that τ the treatment effect near the cutoff is still identified.

$$Y = \beta_0 + \tau D + \beta_1(X - x_0) + \beta_2(X - x_0)^2 + \gamma_1 D \cdot (X - x_0) + \gamma_2 D \cdot (X - x_0)^2 + u$$

Non-parametric RD

- Focus only on observations near the cutoff:

$$Y = \beta_0 + \tau D + \beta_1(X - x_0) + \beta_2 D \cdot (X - x_0) + u$$

  in a sample such that $x_0 - h \leqslant X \leqslant x_0 + h$
- Where h is called the bandwidth of the window
- e.g. Is the cutoff is at 21, then one could decide to only include observations between 20 to 22 which represent a bandwidth of 1

Bandwidth Choice

- However, one has to choose a bandwidth
- There is a tradeoff
  - A small bandwidth will include observations near the cut-off and decrease a potential extrapolation bias
  - But we lose information, less observations increase the standard errors
  - This is a variance bias tradeoff

7.2 Fuzzy And Mixed Designs

Fuzzy RD

- Imperfect compliance at the threshold
- The running variable determines eligibility / encouragement Z of the treatment but individuals still can choose to take up the treatment D or not:

$$Z_i = \begin{cases} 1, & \text{if} X_i \geqslant x_0 \\ 0, & \text{if} X_i < x_0 \end{cases}$$

- The discontinuity lies now in the probability of treatment receipt but does not jump from 0 to 1 as in the sharp design

Assumption 2. *RDD-1*

$$\lim_{\varepsilon \to 0} p\left(D = 1 | X = x_0 + \varepsilon\right) \neq \lim_{\varepsilon \to 0} P\left(D = 1 | X = x_0 - \varepsilon\right)$$

Or, the probability of $D = 1$ has a discontinuity at $X = x_0$

Local Compliers Concept

- Let $D(x)$ be the treatment status of individual i if X was exogenously set to x
- Moving $X = x$ a bit around the threshold, leads to four different types of people
  1. Local always takers $D(x_0 - \varepsilon) = D(x_0 + \varepsilon) = 1$
  2. Local never takers $D(x_0 - \varepsilon) = D(x_0 + \varepsilon) = 0$
  3. Local compliers $D(x_0 - \varepsilon)$ and $D(x_0 + \varepsilon) = 1$
  4. Local defiers $D(xo - \varepsilon) = 1$ and $D(x_0 + \varepsilon) = 0$

Assumption 3. *RDD-3*

$$\{Y_i(1) - Y_i(0), D_i(x)\} \perp X_i near\ x_0$$

and there exists $\varepsilon > 0$ *such that for all* $0 < \varepsilon < e$

$$D_i(x_0 + \varepsilon) \geqslant D_i(x_0 - \varepsilon)$$

The first line : similar to an instrument exclusion restriction The second line : local monotonicity restriction which assumes away the existence of local defiers in a neighbourhood of x_0

Local LATE

Under assumptions RDD - 1,2,3. Fuzzy RD identifies the local LATE:

$$\tau_{LATE}(x_0) = \lim_{\varepsilon \to 0} E\left[Y(1) - Y(0) | D(x_0 + \varepsilon) > D(x_0 - \varepsilon), X = x_0\right]$$

It can be shown that the ATE on the local compliers can be identified as:

$$\tau_{LATE}(x_0) = \frac{\lim\limits_{\varepsilon \to 0} E\left[Y | x_0 + \varepsilon\right] - \lim\limits_{\varepsilon \to 0} E\left[Y | x_0 - \varepsilon\right]}{\lim\limits_{\varepsilon \to 0} E\left[D | x_0 + \varepsilon\right] - \lim\limits_{\varepsilon \to 0} E\left[D | x_0 - \varepsilon\right]}$$

The $\tau_{LATE}(x_0)$ is local twice : for $X = x_0$ and for compliers

Mixed Design - Identification

Mixed RD recovers $ATT(x_0)$ One sided non compliance : when treatment eligibility depends strictly on the threshold Participation is voluntary when the threshold is passed. Non eligible implies no participation

$$\lim_{\varepsilon \to 0}$$

In a mixed design, under RDD - 1 and RDD - 2, we can show that the LATE and the ATT are the same at the threshold

$$\tau_t(x_0) = E\left[Y(1) - Y(0) | D = 1, X = x_0\right]$$

The main assumption needed is the mean of $Y(0)$ is continuous at the threshold (RDD-2)

Estimation

The LATE at the threshold looks just like the IV estimator at the threshold for a binary instrument. We can use standard instrumental variable techniques Recall:

$$Z_i = \begin{cases} 1, & if X_i \geqslant x_0 \\ 0, & \text{if} X_i < x_0 \end{cases}$$

The estimation follows a similar reasoning as for the sharp RD, with this time instrumenting D with Z

Thus, fuzzy RD can be described by the two equations system

$$\begin{aligned} \text{first stage} : D &= \gamma + \delta Z + g(X - x_0) + \nu \\ \text{Second stage} : Y &= \beta_0 + \tau D + f(X - x_0) + u \end{aligned}$$

Where $f(X - x_0)$ and $g(X - x_0)$ are again p-order polynomials 2SLS : Instrument D with Z

Note that substituting the treatment determining equation into the outcome equation yields the reduced form

$$Y = \beta_{0,r} + \tau_r Z + f_r(X - x_0) + u_r$$

With $\tau_r = \tau\delta$. Then, if the same polynomial is used for $f(.)$ and $g(\cdot)$, 2SLS is numerically identical to $\tau = \frac{\tau_r}{\delta}$

- We can also allow for different slopes on both sides of the threshold by using interaction terms where D and its interactions are instrumented by Z and its interactions:
- E.g. for a quadratic polynomial:

$$\begin{aligned} \text{First Stage} : \quad & D = \gamma + \delta Z + \gamma_1(X - x_0) + \gamma_2(X - x_0)^2 + \gamma_3 Z \cdot (X - x_0) + \gamma_4 Z \cdot (X - x_0)^2 + \nu \\ \text{Second Stage} : \quad & Y = \beta_0 + \tau D + \beta_1(X - x_0) + \beta_2(X - x_0)^2 + \beta_3 D \cdot (X - x_0) + \beta_4 D \cdot (X - x_0)^2 + u \end{aligned}$$

- The estimation results in estimating a regression of

$$Y \sim D + (X - x_0) + (X - x_0)^2 + D(x - x_0) + D(X - x_0)^2$$

- Then, instrumenting D, $D \cdot (X - x_0)$, $D \cdot (X - x_0)^2$ with $Z, Z \cdot (X - x_0), Z \cdot (X - x_0)^2$

The non-parametric versions consists of an IV estimation within a small neighbourhood around the cutoff

For example with interaction terms

$$\begin{aligned} \text{First Stage} : \; & D = \gamma + \delta Z + \gamma_1(X - x_0) + \gamma_2(X - x_0)^2 + \gamma_3 Z \cdot (X - x_0) + \gamma_4 Z \cdot (Z - x_0)^2 + \nu \\ \text{Second Stage} : \; & Y = +\beta_0 + \tau D + \beta_1(X - x_0) + \beta_1(X - x_0)^2 + \beta_1 D \cdot (X - x_0) + \beta_1 D \cdot (Z - x_0)^2 + u \\ & \text{in a sample such that: } x_0 - h \leqslant X \leqslant Z \cdot (X - x_0) \end{aligned}$$

Where h is called the bandwidth of the window, and we encounter again a *variance-bias* tradeoff.

Non-parametric Estimation

Instead of assuming the shape of regression function, assume that $E[Y|X = x] - \mu(x)$ is any function of the running variable The idea is to compare people just before and just after the threshold.

We have seen the naive non-parametric estimator: define a window (or bandwidth) around the threshold and estimate e.g. the mean for all observations above / below the cutoff. The difference in mean outcomes gives you the causal effect.

The implicit weighting function of each individual is very basic : each individual gets the same weight. However, intuitively, we would want to give higher weights to observations closer to the cut-off and less weights to those further away from the cut-off. In order to estimate, we must select a kernel and bandwidth

Kernels The weighting function is called a Kernel $K(u)$ and determines which observations are included in the estimation and how. The word kernel refers to any smooth function K such that $K(u) \geqslant 0$ and $\int K(u)du = 1$, $\int uK(u)du = 0$ and $\sigma_K^2 = \int u^2 K(u)du > 0$ A *uniform* or *boxcar* kernel will allocate the same weight to everyone as in the naive approach: $K(u) = \frac{1}{2}1_{|u|\leqslant 1}$ Or other popular kernel choice are the Gaussian and Epanechinkov

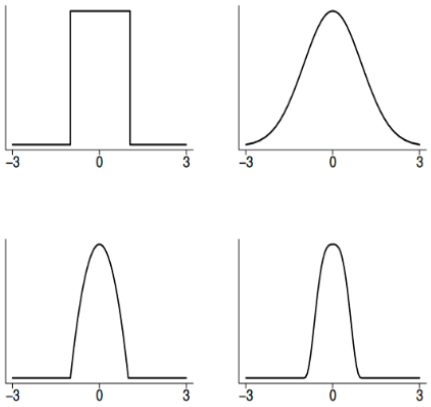


Figure 35: igure: Source: Wasserman (2006), chapter 4

Local constant kernel estimator Estiamte $\tau_r = \lim_{\varepsilon \to 0} E[Y|x_0 + @E]$ and $\tau_i = \lim_{\varepsilon \to 0} E[Y|x_0 - \varepsilon]$:

$$\hat{\tau}_r(x_0) = \frac{\Sigma_{X_i \geqslant x_0} Y_i K(\frac{X_i - x_0}{h}}{\Sigma_{X_i \geqslant x_0} K(\frac{X_i - x_0}{h})}$$

$$\hat{\tau}_l(x_0) = \frac{\Sigma_{X_i \leqslant x_0} Y_i K(\frac{X_i - x_0}{h}}{\Sigma_{X_i \leqslant x_0} K(\frac{X_i - x_0}{h})}$$

The outcomes are weighted by the kernel function. The causal effect is then estimated as follows:

$$ATE(x_0) = \hat{\tau}_r(x_0) - \hat{\tau}(x_0)$$

And it is recommended to use boundary correction kernels for the local constant kernel estimator since the estimator is biased at the boundary (cutoff)

However, since the estimation involves boundaries (cutoffs), local linear regression estimators are often preffered since they are expected to have better boundary properties than many other estimators

Non-parametric Local Linear Regression Intuition is to use a weighted regression where the weights are determined by the kernel function.

The causal effect is then estimated as follows:

$$ATE(x_0) = \hat{a}_r - \hat{a}_l$$

Note that the bandwidths h_r and h_l do not have to be the same on each side of the threshold.

7.3 Application

Angrist and Lavy (1999) aim at estimating the effect of class size on children's test scores (fuzzy RD)

- There are 2 generalisations
    1. Class size, takes on many values, the first stage exploits jumps in average class size instead of probabilities
    2. Multiple discontinuities
- There is a class size rule capped at 40, meaning that classes with 41 students are split into 2 classes
- Enrollment represents the running variable (X)
- Class-size represents take-up (D) and the class size rule, rules the instrument (Z)
- Test scores are the outcomes variable (Y)

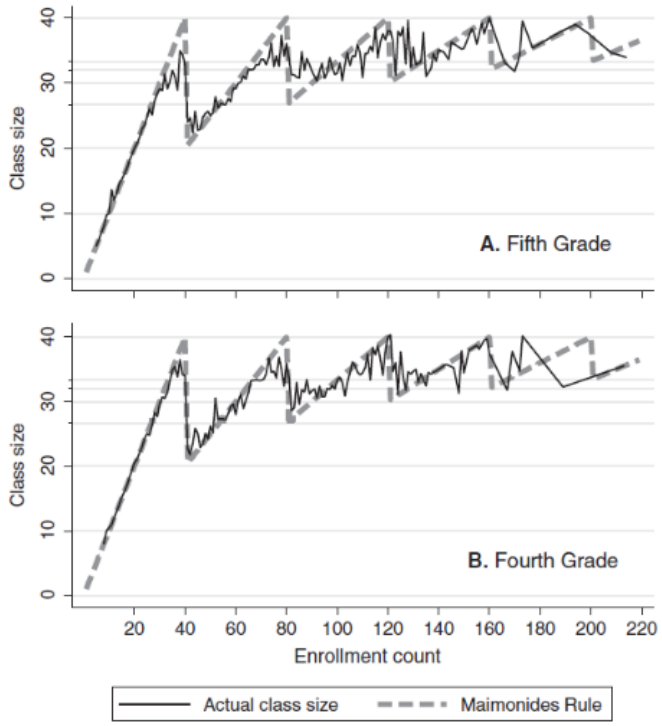


Figure 6.2.1 The fuzzy-RD first-stage for regression-discontinuity estimates of the effect of class size on test scores (from Angrist and Lavy, 1999).

Figure 36: Validity of The Instrument

- There are discontinuities at enrollment of 40,80,120
- This is a fuzzy design since some schools decide to cap earlier than at 40
- Estimation via 2SLS with higher polynomials
- Class-size is instrumented with the rule
- Additional controls are the proportion of students from disadvantaged backgrounds

TABLE 6.2.1
OLS and fuzzy RD estimates of the effect of class size on fifth-grade math scores

	OLS			2SLS				
				Full Sample		Discontinuity Samples		
						±5		±3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean score		67.3		67.3		67.0		67.0
(SD)		(9.6)		(9.6)		(10.2)		(10.6)
Regressors								
Class size	.322	.076	.019	−.230	−.261	−.185	−.443	−.270
	(.039)	(.036)	(.044)	(.092)	(.113)	(.151)	(.236)	(.281)
Percent disadvantaged		−.340	−.332	−.350	−.350	−.459	−.435	
		(.018)	(.018)	(.019)	(.019)	(.049)	(.049)	
Enrollment			.017	.041	.062		.079	
			(.009)	(.012)	(.037)		(.036)	
Enrollment squared/100					−.010			
					(.016)			
Segment 1 (enrollment 38–43)								−12.6
								(3.80)
Segment 2 (enrollment 78–83)								−2.89
								(2.41)
R^2	.048	.249	.252					
Number of classes		2,018		2,018		471		302

Notes: Adapted from Angrist and Lavy (1999). The table reports estimates of equation (6.2.6) in the text using class averages. Standard errors, reported in parentheses, are corrected for within-school correlation.

Figure 37: Estimates

- Column 1: A one unit increase in class size is associated with a 0.332 % increase in math scores, however only 4.8 % of the variation is explained
- Column 2: the effect of class size is reduced when controlling for mean score and percent disadvantaged. But we find higher percentage f disadvantaged students is associated with lower math scores (-0.34)
- Column 3: adding enrollment squared and segments
- Columns 4-5: full sample 2SLS estimates where class size has a negative effect on moth scores using IV
- Columns 6-8: discontinuity samples focusing on smaller bandwidths around the cutoff. Finding larger negative effects but with increases standard errors

Thus, there is an initial positive association between class size and math scores, which diminishes and becomes almost negligible when controlling for additional, important factors. Though we find there is a high percentage of disadvantaged students consistently negatively associated with math scores.