

Revision Notes by Sally Yang

# PROBLEMS OF *Applied* ECONOMETRICS

MICHAELMAS TERM

PROFESSOR MARK SCHANKERMAN  
LONDON SCHOOL OF ECONOMICS 2021/22

# SETUP · MATCHING

Methodological issues in identification/estimation of policy impact on outcomes

## CAUSATION

An exogenous change in  $X$  changes  $Y$ , holding all other (un)observable determinants of  $Y$  constant.

## ASSUMPTIONS

- Statistical
- Behavioural (Rationality, non-maximising)
- Assumption wrt how people respond to intervention heterogeneously

## TREATMENT STATUS

$$D_i = \begin{cases} 0 & \text{if } i \text{ did not receive treatment} \\ 1 & \text{if } i \text{ received treatment} \end{cases}$$

## OUTCOMES

Outcome of  $i$  without treatment =  $Y_{0i}$

Outcome of  $i$  with treatment =  $Y_{1i}$

$$\text{Observed outcome } Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

# RUBIN-HOLLAND CAUSAL MODEL

## TREATMENT EFFECT

Effect of treatment on individual  $i$ :  $\Delta_i = Y_{1i} - Y_{0i}$

But you never directly observe both  $Y_{1i}$  and  $Y_{0i}$  for individual  $i$   
i.e. evaluation is a problem of missing data

We need to construct an appropriate counterfactual

## AVERAGE TREATMENT EFFECT

$$ATE = E(\Delta_i) = E(Y_{1i} - Y_{0i}) = Pr(D_i=1) ATET + Pr(D_i=0) ATEU$$

$$ATET = E(Y_{1i} - Y_{0i} | D_i=1)$$

$$ATEU = E(Y_{1i} - Y_{0i} | D_i=0)$$

↳ e.g. estimate the impact of rolling out a programme to a different age group

There are conditional versions, e.g.  $ATE(x) = E(\Delta_i | X_i=x)$

## OTHER EFFECTS

- % of people benefitting from policy
- Distribution of treatment effects
- Selected quantile

$$Pr(\Delta_i > 0 | D_i=1)$$

$$F(\Delta_i | X_i)$$

$$\inf \{\Delta_i : F(\Delta_i | D_i=1) = q\}$$

## MAIN IDENTIFICATION ISSUES

- Selection into treatment may depend on both  $Y_{0i}$  and  $Y_{1i}$   
(i.e. size of  $\Delta_i$ )  $D_i$  endogenous
- $\Delta_i$  heterogeneous

# ROY'S MODEL

$$Y_{id} = \mu_d + U_{id}, \quad d=0, 1$$

↑  
mean  $E(Y_{id})/E(Y_{id}|X_i=x)$

$Y_{id} - \mu_d$  is absolute advantage in picking d

## SELECTION INTO TREATMENT

i selects into treatment ( $D_i = 1$ ) if  $Y_{i1} > Y_{i0} + C_i$

comparative advantage

Direct cost of choosing 1 instead of 0.

May be heterogeneous, but assume  $C_i = 0$  for now

$$\begin{aligned} ATE &= E(Y_{i1} - Y_{i0}) = E(\mu_1 + U_{i1} - \mu_0 - U_{i0}) = \mu_1 - \mu_0 \\ ATET &= E(Y_{i1} - Y_{i0} | D_i = 1) = E(\Delta_i | \Delta_i \geq C_i) = \mu_1 - \mu_0 + [E(U_{i1} - U_{i0} | U_{i1} - U_{i0} \geq \mu_1 - \mu_0 + C_i)] > 0 \\ ATEU &= E(Y_{i1} - Y_{i0} | D_i = 0) = E(\Delta_i | \Delta_i \leq C_i) = \mu_1 - \mu_0 + [E(U_{i1} - U_{i0} | U_{i1} - U_{i0} \leq \mu_1 - \mu_0 + C_i)] < 0 \\ \text{Var}(\Delta_i) &= \text{Var}(U_{i1}) + \text{Var}(U_{i0}) - [2\text{Cov}(U_{i0}, U_{i1})] \\ &\uparrow \text{Cov}(U_{i0}, U_{i1}) \Rightarrow ATE/ATET/ATEU diverge less \end{aligned}$$

$\Delta_i = \mu_1 - \mu_0 + U_{i1} - U_{i0}$  : individual treatment effects (i-specific)

$\bar{\Delta} = \mu_1 - \mu_0 = ATE$  : mean treatment effects

$v_i = \Delta_i - \bar{\Delta} = U_{i1} - U_{i0}$  : deviation of treatment effect around the mean  
 $v_i = 0$  implies no heterogeneity

$\downarrow$   
 $U_{i1} - U_{i0}$  : i's comparative advantage. observed by i, unobservable to the analyst

## THREE CASES

### NO HETEROGENEITY

$\Delta_i$  (Coefficient on  $D_i$ ) is fixed / constant across i given  $X_i$   
 $U_{i1} = U_{i0} \forall i \Rightarrow v_i = 0, \Delta_i = \bar{\Delta} \forall i \quad ATE = ATET$

### RCT

$\Delta_i$  (Coefficient on  $D_i$ ) is random given  $X_i$   
Heterogeneity exists ( $v_i \neq 0$ ) but not acted upon ex ante  
Random assignment / i cannot observe  $U_{i1} - U_{i0}$ . No self selection  
 $D_i | X_i \perp\!\!\!\perp U_{i1}, U_{i0} \Rightarrow E(U_{i1} - U_{i0} | X_i, D_i = 1) = 0$

### SELF SELECTION

$\Delta_i$  (Coefficient on  $D_i$ ) is random given  $X_i$

# SELECTION EFFECT

causes OLS to no longer consistently estimate ATET

We can write outcome (using the Roy model) as

$$Y_i = D_i Y_{ii} + (1 - D_i) Y_{io} = \mu_0 + (\mu_i - \mu_0 + U_{ii} - U_{oi}) D_i + U_{oi} = \alpha + \Delta_i D_i + \varepsilon_i$$

We know  $\hat{\beta}_{OLS} = \frac{\bar{Y}_i D_i - \bar{Y}_i \bar{D}_i}{\bar{D}_i^2 - (\bar{D}_i)^2}$  and  $\text{plim}(\hat{\beta}_{OLS}) = \frac{E(Y_i D_i) - E(Y_i) E(D_i)}{E(D_i^2) - [E(D_i)]^2}$

We can show  $\text{plim}(\hat{\beta}_{OLS}) = \text{ATET} + E(Y_{io}|D_i=1) - E(Y_{io}|D_i=0)$  E

$= \text{ATE} = \text{ATEU}$   
with randomisation  
or with CIA (conditioning on X)

## Selection effect

- Exists even w/o impact heterogeneity ( $v_i = U_{ii} - U_{oi} = 0$ ) if  $\text{Cov}(\varepsilon_i, D_i) = \text{Cov}(U_{io}, D_i) \neq 0$  some underlying characteristic  $U_{io}$  is related to  $D_i$
- Can occur due to self selection
- Need to balance the observables & unobservables

# MATCHING

Pair treated individuals ( $D_i=1$ ) with untreated individuals ( $D_i=0$ ) who are observably similar ( $X_i$ ). With more controls  $X_i$ , it makes us more confident about the CIA.

## CONDITIONAL INDEPENDENCE ASSUMPTION

$$(Y_{ii}, Y_{io}) \perp\!\!\!\perp D_i | X_i, \text{ i.e. } \Pr(D_i | X_i, Y_{io}, Y_{ii}) = \Pr(D_i | X_i)$$

Treatment status independent of the unobservables (that affect outcomes)

Individuals cannot self-select into treatment based on anticipated treatment impact.

CIA will make matching as good as random... but have I included enough observables?

CIA makes  $\text{ATE} = \text{ATET} = \text{ATEU}$  at a given  $X_i$  value, but the distribution of  $X_i$  may differ for the treated and untreated ( $D_i$ ) and  $X_i$  may be correlated! Some ppl may be outside common support) so ATET and ATEU may differ wrt the entire treated and untreated populations

## COMMON SUPPORT ASSUMPTION

$$\begin{aligned} 0 < \Pr(D_i=1 | X_i) < 1 \quad \forall X_i \\ \exists x \in X_i \text{ s.t. } D_i=0 \text{ or } D_i=1 \end{aligned}$$

e.g. if every girl is treated, there's no untreated girl out there who can be the control group

In practice, if  $P(x)=0$  or  $1$  for some  $x$ , treated units with this  $x$ -value should be dropped.

Estimation only takes place over (and the  $\widehat{\text{ATET}}$  only relevant for people in) the common support of  $X$ .

## SUTVA/NO SPILLOVERS

$D_i$  has no effect on the outcome  $Y_j$  of individual  $j \neq i$ .

- No general equilibrium effects / social interactions
- Particularly problematic if treated  $i$  spillover to untreated  $j$ 
  - We can't even tell who's treated!
  - Then, the estimate = direct effect (ATET) - indirect effect
  - Can underestimate direct effect (if both effects same direction)
  - Can massively underestimate social benefit (= direct + indirect effect)

Taking expectations over  $X$

$$\text{Vanilla LIE: } E_x[E(Y|X)] = E(Y)$$

$$\therefore \text{ADVANCED LIE: } E_{x|z}[E(Y|X,z)|z] = E(Y|z)$$

Collapses along  $X$ -dimension,  
but not  $Z$  (which we keep constant)

## PROPENSITY SCORE MATCHING

When there are many controls  $X_i$ , difficult to match exactly (curse of dimensionality)  
 Reduce the multidimensional  $X_i$  into a unidimensional propensity score:

$$P(x) = P(D_i=1 | X_i=x) \quad \forall x \in X_i$$

Theorem: CIA  $\Rightarrow Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | P(X_i)$

- If we assume CIA, we can just condition (match) on  $P(X_i)$  instead of  $X_i$

$$E(Y_{0i} | P(X_j)=P(X_i), D_i=0) \stackrel{\text{CIA}}{=} E(Y_{0i} | P(X_i), D_i=0)$$

outcome of untreated  $j$ s that share  
propensity scores with the treated

$$E(Y_{1i} | P(X_j)=P(X_i), D_i=1) \stackrel{\text{CIA}}{=} E(Y_{1i} | P(X_i), D_i=1)$$

counterfactual untreated outcome  
of the treatment group

$$\begin{aligned} \text{ATET} &= E(Y_{1i} | D_i=1) - E(Y_{0i} | D_i=1) \\ &\stackrel{\text{LIE}}{=} E(Y_{1i} | D_i=1) - E[E(Y_{0i} | X_i, D_i=1) | D_i=1] \\ &\stackrel{\text{CIA}}{=} E(Y_{1i} | D_i=1) - E[E(Y_{0j} | X_j=X_i, D_j=1) | D_i=1] \\ &\stackrel{\substack{\text{Rosenbaum} \\ \text{Rubin}}}{=} E(Y_{1i} | D_i=1) - E[E(Y_{0j} | P(X_j)=P(X_i), D_i=0) | D_i=1] \\ &= E[Y_{1i} - E(Y_j | P(X_j)=P(X_i), D_i=0) | D_i=1] \end{aligned}$$

### ESTIMATING ATET

① Estimate propensity score  $\hat{P}(x_i) := \hat{P}(D_i=1 | X_i=x_i)$  for each person using some model of participation (e.g. latent variable model + probit/logit)

② The estimator for  $E(Y_j | P(X_j)=P(X_i), D_i=0)$  is  $\hat{m}_0(x_i) = \frac{\sum_{j: d_j=0} y_j K(\hat{p}(x_j) - \hat{p}(x_i))}{\sum_{j: d_j=0} K(\hat{p}(x_j) - \hat{p}(x_i))}$

③  $\widehat{\text{ATET}} = \frac{1}{\sum d_i} \sum_{i: d_i=1} [y_i - \hat{m}_0(x_i)]$

$$\widehat{\text{ATEU}} = \frac{1}{\sum_{i: d_i=1}} \sum_{i: d_i=1} [\hat{m}_1(x_i) - y_i] \quad \hat{m}_1(x_i) = \frac{\sum_{j: d_j=1} y_j K(\hat{p}(x_i) - \hat{p}(x_j))}{\sum_{j: d_j=1} K(\hat{p}(x_i) - \hat{p}(x_j))}$$

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^N \frac{y_i(d_i - \hat{p}(x_i))}{\hat{p}(x_i)[1-\hat{p}(x_i)]}$$

weight↑ when predicted participation diverges from actual participation  
weight↑ when  $\hat{p}(x_i)$  gets away from  $\frac{1}{2}$

# RCT

With random assignment,  $R_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$

$$\boxed{\text{FULL COMPLIANCE}} \quad D_i = 1 \Leftrightarrow R_i = 1 \quad D_i = 0 \Leftrightarrow R_i = 0 \quad + \text{random assignment} \Rightarrow D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$$

$$\begin{aligned} \text{ITT} &= E(Y_i | R_i = 1) - E(Y_i | R_i = 0) \\ &= E(Y_{1i} | R_i = 1) - E(Y_{0i} | R_i = 1) \text{ by random assignment} \\ &= E(Y_{1i} | D_i = 1) - E(Y_{0i} | D_i = 1) \text{ by full compliance} = \text{ATET} \\ &= E(Y_{1i}) - E(Y_{0i}) = \text{ATE} = \text{ATEU} \dots \end{aligned}$$

$$\boxed{\text{ONE-SIDED NONCOMPLIANCE}} \quad R_i = 0 \Rightarrow D_i = 0 \quad \stackrel{\text{contrapositive}}{D_i = 1 \Rightarrow R_i = 1}$$

selecting out of treatment

$$\begin{aligned} \text{ITT} &= E(Y_i | R_i = 1) - E(Y_i | R_i = 0) \\ &= E(Y_{1i} | R_i = 1) - E(Y_{0i} | R_i = 1) \\ &= E(Y_{1i} | R_i = 1, D_i = 1) \Pr(D_i = 1 | R_i = 1) + E(Y_{1i} | R_i = 1, D_i = 0) \Pr(D_i = 0 | R_i = 1) \\ &\quad - E(Y_{0i} | R_i = 1, D_i = 1) \Pr(D_i = 1 | R_i = 1) - E(Y_{0i} | R_i = 1, D_i = 0) \Pr(D_i = 0 | R_i = 1) \\ &= E(Y_{1i} - Y_{0i} | R_i = 1, D_i = 1) \Pr(D_i = 1 | R_i = 1) \\ &= E(Y_{1i} - Y_{0i} | D_i = 1) \Pr(D_i = 1 | R_i = 1) \quad \text{since } D_i = 1 \Rightarrow R_i = 1 \\ &= \text{ATET} \times \text{Compliance rate} \\ &= \frac{\text{first stage / denominator of Wald estimator}}{\Pr(D_i = 1 | R_i = 1) - \Pr(D_i = 0 | R_i = 1)} \end{aligned}$$

Conclusion: Wald  $\hat{\beta}_{IV}$  consistently estimates ATET under one-sided noncompliance and random assignment... though under one-sided noncompliance you typically just estimate ITT Wald estimator is not very necessary.

$$\boxed{\text{GENERAL NONCOMPLIANCE}} \quad \text{must assume monotonicity: } D_i(1) \geq D_i(0) \quad \forall i$$

selecting in and out of treatment

$$\text{and exclusion restriction: } (Y_{0i}, Y_{1i}, D_i(Z_i)) \perp\!\!\!\perp Z_i$$

$$\begin{aligned} \text{ITT} &= E(Y_i | R_i = 1) - E(Y_i | R_i = 0) \\ &= E[D_i(R_i) Y_{1i} + (1 - D_i(R_i)) Y_{0i} | R_i = 1] - E[D_i(R_i) Y_{1i} + (1 - D_i(R_i)) Y_{0i} | R_i = 0] \\ &= E[D_i(1) Y_{1i} + (1 - D_i(1)) Y_{0i}] - E[D_i(0) Y_{1i} + (1 - D_i(0)) Y_{0i}] \quad \text{by exclusion restriction} \\ &= E\{(Y_{1i} - Y_{0i}) [D_i(1) - D_i(0)]\} \\ &= E[(Y_{1i} - Y_{0i}) \times 1 | D_i(1) - D_i(0) = 1] \Pr[D_i(1) - D_i(0) = 1] \\ &\quad + E[(Y_{1i} - Y_{0i}) \times 0 | D_i(1) - D_i(0) = 0] \Pr[D_i(1) - D_i(0) = 0] \\ &\quad + E[(Y_{1i} - Y_{0i}) \times -1 | D_i(1) - D_i(0) = -1] \Pr[D_i(1) - D_i(0) = -1] \\ &= E[Y_{1i} - Y_{0i} | D_i(1) - D_i(0) = 1] \Pr[D_i(1) - D_i(0) = 1] \quad \text{by monotonicity} \\ &= \text{LATE} \times \% \text{ Compliers} \\ &= \frac{\text{first stage / denominator of Wald estimator}}{\Pr[D_i(1) = 1] - \Pr[D_i(0) = 1]} \\ &\quad \text{because } \Pr[D_i(1) - D_i(0) = 1] = \Pr[D_i(1) = 1 \cap D_i(0) = 0] \\ &\quad \text{total probability theorem} = 1 - \Pr[D_i(1) = 0] - \Pr[D_i(0) = 1] \\ &\quad = \Pr[D_i(1) = 1] - \Pr[D_i(0) = 1] \end{aligned}$$

Conclusion: Wald  $\hat{\beta}_{IV}$  consistently estimates LATE under general noncompliance, monotonicity, exclusion restriction and relevance ( $\Pr[D_i(1) = 1] - \Pr[D_i(0) = 1] \neq 0$ )

# IV

When self selection as per Roy model exists,

$$\begin{aligned}
 Y_i = D_i Y_{ii} + (1 - D_i) Y_{oi} &= \mu_0 + (\mu_i - \mu_0 + U_{ii} - U_{oi}) D_i + U_{oi} \\
 &= \mu_0 + [\mu_i - \mu_0 + E(U_{ii} - U_{oi} | D_i = 1)] D_i + [U_{ii} - U_{oi} - E(U_{ii} - U_{oi} | D_i = 1)] D_i + U_{oi} \\
 &= \mu_0 + \text{ATE} D_i + \eta_i \quad \text{Cov}(D_i, \eta_i) \neq 0!
 \end{aligned}$$

**WALD ESTIMATOR (BINARY)**  $\hat{\beta}_{IV} = \frac{\bar{Y}_i Z_i - \bar{Y}_i \bar{Z}_i}{\bar{D}_i Z_i - \bar{D}_i \bar{Z}_i}$

$$\begin{aligned}
 \text{plim } \hat{\beta}_{IV} &= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{E(Y_i Z_i) - E(Y_i) E(Z_i)}{E(D_i Z_i) - E(D_i) E(Z_i)} \\
 &= \frac{E(Y_i | Z_i = 1) \Pr(Z_i = 1) - [E(Y_i | Z_i = 0) \Pr(Z_i = 0) + E(Y_i | Z_i = 1) \Pr(Z_i = 1)] \Pr(Z_i = 1)}{E(D_i | Z_i = 1) \Pr(Z_i = 1) - [E(D_i | Z_i = 0) \Pr(Z_i = 0) + E(D_i | Z_i = 1) \Pr(Z_i = 1)] \Pr(Z_i = 1)} \\
 &= \frac{E(Y_i | Z_i = 1) - \{E(Y_i | Z_i = 0)[1 - \Pr(Z_i = 1)] + E(Y_i | Z_i = 1) \Pr(Z_i = 1)\}}{E(D_i | Z_i = 1) - \{E(D_i | Z_i = 0)[1 - \Pr(Z_i = 1)] + E(D_i | Z_i = 1) \Pr(Z_i = 1)\}} \\
 &= \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)} \times \frac{\Pr(Z_i = 1)[1 - \Pr(Z_i = 1)]}{\Pr(Z_i = 1)[1 - \Pr(Z_i = 1)]} \\
 &= \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)} \text{ reduced form (Rachael's notation: ITT}_Y\text{)} \\
 &\quad \text{first stage (Rachael's notation: ITT}_D\text{)}
 \end{aligned}$$

Condition for  $\hat{\beta}_{IV}$  consistency:  $\text{Cov}(Z_i, \eta_i) = E(Z_i \eta_i) - E(Z_i) E(\eta_i) = 0$

If  $U_{oi} = U_{ii}$ ,  $\eta_i = 0$ , condition automatically true. In fact,  $\text{plim } \hat{\beta}_{IV} = \text{ATE} = \text{ATE}$

If  $U_{oi} \neq U_{ii}$ , we need to make assumptions s.t.  $\text{Cov}(Z_i, \eta_i) = 0$ .

Three possible assumptions:  $Z_i \perp\!\!\!\perp \eta_i$   $\Rightarrow E(\eta_i | Z_i) = E(\eta_i) \Rightarrow E(Z_i \eta_i) = 0$

independence  
sufficient (strong)    mean independence  
sufficient (usual)    orthogonality  
necessary

And  $\text{plim } \hat{\beta}_{IV} = \text{ATE}$  or LATE or ATE on compliers and defiers (see below).

**MONOTONICITY**  $\forall z_1, z_2 \in Z_i$  s.t.  $z_1 \neq z_2$ ,  $D_i(z_1) \geq D_i(z_2)$  or  $D_i(z_1) \leq D_i(z_2)$   $\forall i$

**EXCLUSION RESTRICTION**  $(Y_{oi}, Y_{ii}, D_i(z_i)) \perp\!\!\!\perp Z_i$

**RELEVANCE**  $E(D_i | Z_i)$  should be a nontrivial function of  $Z_i$   
 Binary case:  $E(D_i | Z_i = 1) - E(D_i | Z_i = 0) \neq 0$

**LATE**  $\text{LATE} = E(Y_{ii} - Y_{oi} | D_i(1) - D_i(0) > 0)$  is ATE on compliers

- $\text{plim } \hat{\beta}_{IV} = \text{LATE}$  under monotonicity and exclusion restriction (previous page)

LATE with continuous  $Z_i$  is

$$\text{LATE}(z, z') = E[Y_{ii} - Y_{oi} | P(z') - P(z) = 1] = \frac{E(Y_i | Z_i = z) - E(Y_i | Z_i = z')}{E(D_i | Z_i = z) - E(D_i | Z_i = z')} \stackrel{\leftarrow \text{Rosenbaum-Rubin theorem}}{=} \frac{E[Y_i | P(Z_i) = P(z)] - E[Y_i | P(Z_i) = P(z')]}{P(z') - P(z)} \stackrel{\leftarrow \text{Binary } D_i \rightarrow}{=} P(z') - P(z)$$

EC333 definition

$$\text{LATE}(P(z), P(z')) = E[Y_{ii} - Y_{oi} | D_i(P(z')) - D_i(P(z)) = 1] = \frac{E[Y_i | P(Z_i) = P(z)] - E[Y_i | P(Z_i) = P(z')]}{P(z') - P(z)}$$

ATE on people who'd participate if  $Z_i = z'$  ( $U_{D_i} \leq P(z')$ )  
 but not if  $Z_i = z$  ( $U_{D_i} \geq P(z)$ )

**LOCAL IV** Given continuous IV  $Z_i$ ,  $\Delta_{IV}(p) = \frac{\partial E(Y_i | P(Z_i) = p)}{\partial p}$

# MARGINAL TREATMENT EFFECT

Roy model:  $D_i = \mathbb{1}[\mu_i(Z_i) - \mu_o(Z_i) + U_{ii} - U_{oi} > C_i] = \mathbb{1}[\mu_D(Z_i) > U_{Di}]$

$\mu_D(Z_i) = \mu_i(Z_i) - \mu_o(Z_i)$  observable characteristics making  $i$  more likely to opt in  
 $U_{Di} = C_i - U_{ii} + U_{oi}$  unobservable characteristics making  $i$  more likely to opt out (resistance)

Assume  $U_{Di}$  is continuous, and CSA is supported

$$D_i = \mathbb{1}[\mu_D(Z_i) > U_{Di}] = \mathbb{1}[F_{U_{Di}}(\mu_D(Z_i)) > F_{U_{Di}}(U_{Di})] = \mathbb{1}[\Pr(\mu_D(Z_i) \geq U_{Di}) > F_{U_{Di}}(U_{Di})]$$

$$= \mathbb{1}[\Pr(D_i = 1 | Z_i) > u_{Di}], U_{Di} \sim U[0, 1]$$

$$F_{U_{Di}}(U_{Di}) \sim U[0, 1] \text{ as } \forall x \in [0, 1], \Pr[F_{U_{Di}}(U_{Di}) \leq x] = \Pr[U_{Di} \leq F_{U_{Di}}^{-1}(x)] = F_{U_{Di}}[F_{U_{Di}}^{-1}(x)] = x$$

## DEFINITION

$MTE(u) = E(\Delta_i | U_{Di} = u)$  is the ATE on people with unobserved resistance  $U_{Di} = u$ .

$MTE(P(Z_i)) = E(\Delta_i | U_{Di} = P(Z_i))$  is the ATE on people who are indifferent towards participation.  
their unobserved resistance equals propensity to opt in (based on observables)

## CONNECTION TO LATE

$$E[Y_i | P(Z_i) = P(z)] = P(z) E[Y_{ii} | P(Z_i) = P(z), D_i = 1] + [1 - P(z)] E[Y_{io} | P(Z_i) = P(z), D_i = 0]$$

$$= P(z) \underbrace{\int_0^{P(z)} E(Y_{ii} | U_{Di} = u) du}_{\text{Sum of ATE on people whose } U_{Di} \text{ is below } P(z), \text{ averaged over the no. of these people}} + [1 - P(z)] \underbrace{\int_{P(z)}^1 E(Y_{io} | U_{Di} = u) du}_{\text{Sum of ATE on people whose } U_{Di} \text{ is above } P(z), \text{ averaged over the no. of these people}}$$

$$= \int_0^{P(z)} E(Y_{ii} | U_{Di} = u) du + \int_{P(z)}^1 E(Y_{io} | U_{Di} = u) du$$

$$E[Y_i | P(Z_i) = P(z)] - E[Y_i | P(Z_i) = P(z')] = \int_0^{P(z)} E(Y_{ii} | U_{Di} = u) du + \int_{P(z)}^1 E(Y_{io} | U_{Di} = u) du - \int_0^{P(z')} E(Y_{ii} | U_{Di} = u) du - \int_{P(z')}^1 E(Y_{io} | U_{Di} = u) du$$

$$= \int_{P(z')}^{P(z)} E(Y_{ii} | U_{Di} = u) du - \int_{P(z)}^{P(z')} E(Y_{io} | U_{Di} = u) du$$

$$= \int_{P(z)}^{P(z')} E(Y_{ii} - Y_{io} | U_{Di} = u) du$$

$$\text{LATE}(P(z), P(z')) = \frac{E[Y_i | P(Z_i) = P(z)] - E[Y_i | P(Z_i) = P(z')]}{P(z') - P(z)} = \frac{\int_{P(z)}^{P(z')} E(\Delta_i | U_{Di} = s) ds}{P(z') - P(z)} = \frac{\int_{P(z)}^{P(z')} MTE(s) ds}{P(z') - P(z)}$$

$$= E[MTE(U_{Di}) | P(z) \leq U_{Di} \leq P(z')] \text{ ATE on people who'd participate if } Z_i = z' \text{ (} U_{Di} \leq P(z') \text{ but not if } Z_i = z \text{ (} U_{Di} \geq P(z) \text{)}$$

$$= \frac{\int_{P(z)}^{P(z')} MTE(u) du}{P(z') - P(z)}$$

## CONNECTION TO ATE, ATET

Rarely possible as  $P(Z_i)$  usually doesn't have full support

$$\left\{ \begin{array}{l} \text{ATE} = \int_0^1 MTE(u) du \\ \text{ATET} = \int_0^1 MTE(u) h_{\text{ATET}}(u) du \end{array} \right.$$

$\xrightarrow{\int_0^1 \frac{1 - F_{P(Z_i)}(u)}{[1 - F_{P(Z_i)}(t)] dt} dt} \Pr[P(Z_i) \geq u]$  weight by probability that someone with  $U_{Di} = u$  will be treated

## ESTIMATE MTE

$$\begin{aligned}
 MTE(P(z)) &= \lim_{P(z') \rightarrow P(z)} LATE(P(z), P(z')) = \lim_{P(z') \rightarrow P(z)} \frac{E[Y_i | P(Z_i) = P(z)] - E[Y_i | P(Z_i) = P(z')]}{P(z') - P(z)} \\
 &= \frac{\partial E[Y_i | P(Z_i) = P(z)]}{\partial P(z)} = \Delta_{LIV}(P(z))
 \end{aligned}$$

- ① Estimate  $P(Z_i)$  function using probit, logit, non-parametric
- ② Estimate conditional mean  $E[Y_i | P(Z_i)]$

③ Estimate slope,  $\Delta_{LIV}(P(Z_i)) = \frac{\partial E[Y_i | P(Z_i) = P(z)]}{\partial P(z)} = MTE(P(z))$  at each  $P(z)$

④ Integrate using MTE function for ATE, ATET, LATE

# DIFFERENCE-IN-DIFFERENCES

## MODEL

$$Y_{it} = \alpha + \Delta_i D_{it} + \varepsilon_{it}$$

$$\left\{ \begin{array}{l} \Delta_i = Y_{it_1} - Y_{it_0}, \\ D_{it} = 0 \text{ if } t=t_0, \\ \varepsilon_{it} = m_t + \eta_i + v_{it} \end{array} \right.$$

"macro" time trends      individual fixed effects  
common to everyone      constant across time

## IDENTIFYING ASSUMPTION

$$E(\varepsilon_{it} | D_{it}) = m_t + E(\eta_i | D_{it}) + 0 \quad \text{i.e. } E(v_{it} | D_{it}) = 0$$

allowed to vary with time      allowed to vary by treatment status      No selection on individual-specific  
but must be common to treatment/control      but fixed across time

No selection on trends:  $E(Y_{it_1} - Y_{it_0} | D_{it}) = E(Y_{it_1} - Y_{it_0})$  (parallel trends!!!)

- This allows identification of ATET. For ATE we also need  $\Delta_i \perp\!\!\!\perp D_{it} \Leftrightarrow (Y_{it_1} - Y_{it_0}) \perp\!\!\!\perp D_{it}$
- Requires  $E(v_{it} | D_{it}) = 0$ .
  - Time-varying shocks (macro trends) must have same average effect on treated and control.
  - Entity-specific shocks must not be time-varying
  - **Ashenfelter's dip:** If enrolment in training is more likely when earnings temporarily fall, treatment group will experience faster earnings growth even absent the treatment
  - Can check pretrends, but you can only assume post-treatment underlying trends are parallel

Selection on levels is allowed:  $E(Y_{it_0} | D_{it_1})$  can depend on  $D_{it_1}$ ,

- $E(\eta_i | D_{it})$  can be a function of  $D_{it}$
- e.g. people who enroll in job training usually have lower earnings historically

## ESTIMATOR

$$\begin{aligned} \text{ATET} &= E(Y_{it_1} - Y_{it_0} | D_{it_1} = 1) \\ &= E(Y_{it_1} - Y_{it_0} | D_{it} = 1) - E(Y_{it_1} - Y_{it_0} | D_{it_1} = 0) \\ &= E(Y_{it_1} - Y_{it_0} | D_{it} = 1) - E(Y_{it_1} - Y_{it_0} | D_{it_1} = 0) \quad \text{by parallel trends} \\ &= E(Y_{it_1} - Y_{it_0} | D_{it_1} = 1) - E(Y_{it_1} - Y_{it_0} | D_{it_1} = 0) \end{aligned}$$

$$\widehat{\text{ATET}} = \frac{\sum_i d_{it_1} (y_{it_1} - y_{it_0})}{\sum_i d_{it_1}} - \frac{\sum_i (1-d_{it_1}) (y_{it_1} - y_{it_0})}{\sum_i (1-d_{it_1})}$$

Can use repeated cross section if treatment and control group can be separated before the treatment and composition of the groups (wrt entity FE)s didn't change

1<sup>st</sup> "diff" within group removes entity FE

$$E(Y_{it_1} - Y_{it_0} | D_{it_1} = 1) = \alpha + E(\Delta_i | D_{it_1} = 1) + m_{t_1} + E(\eta_i | D_{it_1} = 1) - [\alpha + m_{t_0} + E(\eta_i | D_{it_1} = 1)] = m_{t_1} - m_{t_0} + E(\Delta_i | D_{it_1} = 1)$$

$$E(Y_{it_1} - Y_{it_0} | D_{it_1} = 1) = \alpha + m_{t_1} + E(\eta_i | D_{it_1} = 0) - [\alpha + m_{t_0} + E(\eta_i | D_{it_1} = 0)] = m_{t_1} - m_{t_0}$$

2<sup>nd</sup> "diff" across group removes time FE

$$E(Y_{it_1} - Y_{it_0} | D_{it_1} = 1) - E(Y_{it_1} - Y_{it_0} | D_{it_1} = 0) = E(\Delta_i | D_{it_1} = 1)$$

# DID SPECIFICATION

## BASELINE

$$y_{it} = \alpha + X_{it}\beta + \delta T_{it} \times P_{st} + \gamma T_{it} + \lambda P_{st} + v_{it}$$

$$\alpha = E(y_{it} | D_i = 0, Z_i = 0)$$

$$\alpha + \gamma = E(y_{it} | D_i = 1, Z_i = 0)$$

$$\alpha + \lambda = E(y_{it} | D_i = 0, Z_i = 1)$$

$$\alpha + \delta + \gamma + \lambda = E(y_{it} | D_i = 1, Z_i = 1)$$

$$\alpha = E(\eta_i | D_i = 0) + m_{t_0}$$

$$\gamma = E(y_{it} | D_i = 1, Z_i = 0) - E(y_{it} | D_i = 0, Z_i = 0) = E(\eta_i | D_i = 1) - E(\eta_i | D_i = 0) \text{ entity FE}$$

$$\lambda = E(y_{it} | D_i = 0, Z_i = 1) - E(y_{it} | D_i = 0, Z_i = 0) = m_{t_1} - m_{t_0} \text{ time FE common to all entities}$$

$$\delta = ATET = E(y_{it} | D_i = 1, Z_i = 1) - E(y_{it} | D_i = 1, Z_i = 0)$$

$$- \{ E(y_{it} | D_i = 0, Z_i = 1) - E(y_{it} | D_i = 0, Z_i = 0) \}$$

## FIXED EFFECTS

$$y_{it} = X_{it}\beta + \delta T_{it} \times P_{st} + \gamma_j \sum_j D_{ij} + \lambda_s \sum_s Z_{st} + v_{it}$$

Entity FE  $D_{ij} = \mathbb{1}(i=j)$   
Year FE  $Z_{st} = \mathbb{1}(s=t)$

## LEADS & LAGS

$$y_{it} = X_{it}\beta + \sum_{s=0}^{t_0-1} \gamma_s T_{it} \times Z_{st} + \sum_{s=t_0}^T \delta_s T_{it} \times Z_{st} + \gamma_j \sum_j D_{ij} + \lambda_s \sum_s Z_{st} + v_{it}$$

leads (anticipation)      contemporaneous effect  
                                  + lags (diffusion of TEs)

Test  $H_0: \sum_{s=0}^{t_0-1} \gamma_s = 0$  for pretrends.

Rejection need not invalidate study if it's some interesting anticipatory effect

## STAGGERED DID

$$y_{it} = X_{it}\beta + \delta T_{it} \times P_{st} + \gamma_j \sum_j D_{ij} + \lambda_s \sum_s Z_{st} + v_{it}$$

## LINEAR TRENDS

$$y_{it} = X_{it}\beta + \delta T_{it} \times P_{st} + \gamma_j \sum_j D_{ij} + \lambda_s \sum_s Z_{st} + \eta_i t \sum_j D_{ij} + v_{it}$$

$\eta_i t \sum_j D_{ij}$  control for entity-specific linear trends (not the same as FE!)

## EVENT STUDY

$$y_{it} = X_{it}\beta + \delta_t T_{it} \times P_{st} + \gamma_j \sum_j D_{ij} + \lambda_s \sum_s Z_{st} + v_{it}$$

★ Technically one  $D_{ij}$ , one  $Z_{st}$  and one  $\delta_t$  should be omitted as the reference group — link to incidental parameters problem (LT)

# REGRESSION DISCONTINUITY DESIGN

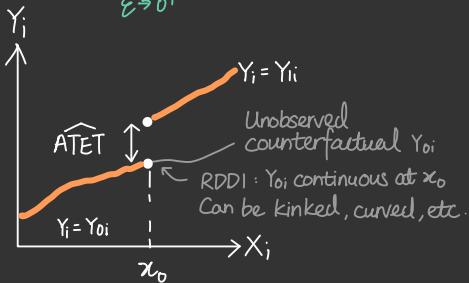
$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} = Y_{i0} + \Delta_i D_i$$

## IDENTIFYING ASSUMPTION

(RDD1)  $E(Y_{oi}|X_i=x)$  is continuous at  $x_0$

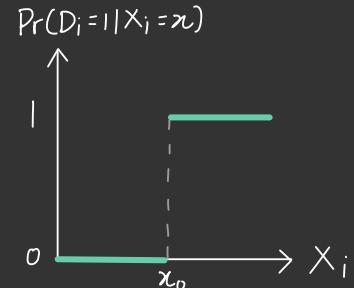
- $Y_{oi}$  may be affected by many individual characteristics  $\alpha_i$ . Those  $\alpha_i$  should also be continuous at  $x_0$ .
- Precludes manipulation around threshold — always check for balance of observable baseline characteristics for people just above and below  $x_0$ .
- If not balanced, it suggests that unobserved parts of  $\alpha_i$  may also be unbalanced.
- Don't need to know the precise function linking  $Y_i$  to  $X_i$  (can always approximate). Just need to assume that the function is continuous at  $x_0$ .

(RDD2)  $\lim_{\epsilon \rightarrow 0^+} E(\Delta_i|X_i=x+\epsilon)$  is well-defined



SHARP DESIGN  $D_i = f(X_i)$

- $D_i$  is fixed given  $X_i$
- Selection on observables only
- Point  $x_0$  at which  $f$  is discontinuous is known
- $\Pr(D_i=1|X_i=x) \in \{0, 1\}$ , violating CSA
- Analogous to simple linear regression with truly exogenous treatment / a valid IV with full compliance, since the determinant of  $D_i$  is fully observed



$$\begin{aligned} & \lim_{\epsilon \rightarrow 0^+} [E(Y_i|X_i=x_0+\epsilon) - E(Y_i|X_i=x_0-\epsilon)] \\ & \stackrel{\text{Well-defined (RDD2)}}{=} \lim_{\epsilon \rightarrow 0^+} E(\Delta_i|X_i=x_0+\epsilon) \stackrel{=1}{\lim_{\epsilon \rightarrow 0^+} E(D_i|X_i=x_0+\epsilon)} - \lim_{\epsilon \rightarrow 0^+} E(\Delta_i|X_i=x_0-\epsilon) \stackrel{=0}{\lim_{\epsilon \rightarrow 0^+} E(D_i|X_i=x_0-\epsilon)} \\ & \quad + \lim_{\epsilon \rightarrow 0^+} E(Y_{i0}|X_i=x_0+\epsilon) - \lim_{\epsilon \rightarrow 0^+} E(Y_{i0}|X_i=x_0-\epsilon) \\ & = E(\Delta_i|X_i=x_0) \quad \text{ATE on people with } X_i=x_0 \end{aligned}$$

RDD1 and 2 are implicitly used in all following proofs

$$\Pr(D_i=1|X_i=x)$$

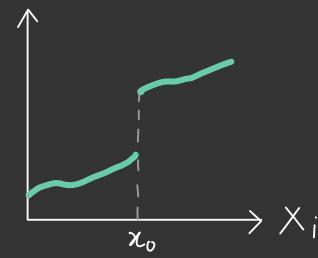
## FUZZY DESIGN

$D_i$  is random given  $X_i$

$\mathbb{1}(X_i > x_0)$  is like an IV for  $\Pr(D_i=1|X_i=x)$

$\Pr(D_i=1|X_i=x)$  is discontinuous at  $x_0$  ('Relevance')

### ① Homogeneous TE $\Delta_i = \Delta$



$$\lim_{\varepsilon \rightarrow 0^+} [E(Y_i|X_i = x_0 + \varepsilon) - E(Y_i|X_i = x_0 - \varepsilon)]$$

$$= \Delta \lim_{\varepsilon \rightarrow 0^+} [E(D_i|X_i = x_0 + \varepsilon) - E(D_i|X_i = x_0 - \varepsilon)]$$

$$\Delta = \frac{\lim_{\varepsilon \rightarrow 0^+} [E(Y_i|X_i = x_0 + \varepsilon) - E(Y_i|X_i = x_0 - \varepsilon)]}{\lim_{\varepsilon \rightarrow 0^+} [E(D_i|X_i = x_0 + \varepsilon) - E(D_i|X_i = x_0 - \varepsilon)]}$$

### ② Heterogeneous TE, but $D_i \perp\!\!\!\perp \Delta_i | X_i \quad \forall i \text{ s.t. } X_i \in \mathcal{N}(x_0)$ in the neighbourhood of $x_0$

- $\mathbb{1}(X_i > x_0)$  is basically an IV with one-sided noncompliance

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0^+} [E(Y_i|X_i = x_0 + \varepsilon) - E(Y_i|X_i = x_0 - \varepsilon)] \\ &= \lim_{\varepsilon \rightarrow 0^+} E(\Delta_i|X_i = x_0 + \varepsilon) \lim_{\varepsilon \rightarrow 0^+} E(D_i|X_i = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0^+} E(\Delta_i|X_i = x_0 - \varepsilon) \lim_{\varepsilon \rightarrow 0^+} E(D_i|X_i = x_0 - \varepsilon) \\ &= E(\Delta_i|X_i = x_0) \lim_{\varepsilon \rightarrow 0^+} [E(D_i|X_i = x_0 + \varepsilon) - E(D_i|X_i = x_0 - \varepsilon)] \end{aligned}$$

$$\text{ATE at } x_0 \quad E(\Delta_i|X_i = x_0) = \frac{\lim_{\varepsilon \rightarrow 0^+} [E(Y_i|X_i = x_0 + \varepsilon) - E(Y_i|X_i = x_0 - \varepsilon)]}{\lim_{\varepsilon \rightarrow 0^+} [E(D_i|X_i = x_0 + \varepsilon) - E(D_i|X_i = x_0 - \varepsilon)]}$$

### ③ $D_i \perp\!\!\!\perp \Delta_i | X_i$ not satisfied (due to self selection, etc.), but

$(\Delta_i, D_i(x)) \perp\!\!\!\perp X_i \quad \forall i \text{ s.t. } X_i \in \mathcal{N}(x_0)$  ('exclusion restriction')

$\forall \varepsilon > 0 \text{ s.t. } D_i(x_0 + \varepsilon) \geq D_i(x_0 - \varepsilon) \quad \forall i$  (monotonicity)

- $\mathbb{1}(X_i > x_0)$  is basically an IV with general noncompliance

$$\text{'Local' LATE} \quad \lim_{\substack{\varepsilon \rightarrow 0^+ \\ \text{Compliers near cutoff}}} E(\Delta_i|X_i = x_0 + \varepsilon, D_i(x_0 + \varepsilon) - D_i(x_0 - \varepsilon) = 1) = \frac{\lim_{\varepsilon \rightarrow 0^+} [E(Y_i|X_i = x_0 + \varepsilon) - E(Y_i|X_i = x_0 - \varepsilon)]}{\lim_{\varepsilon \rightarrow 0^+} [E(D_i|X_i = x_0 + \varepsilon) - E(D_i|X_i = x_0 - \varepsilon)]} \quad \text{Similar working as IV estimator}$$

## Summary

The same estimator identifies different TEs under different assumptions

External validity concern: ATE identified only for people with  $X_i$  near  $x_0$  (unless  $\Delta_i = \Delta$ )

## NONPARAMETRIC ESTIMATOR

Can estimate each limit using one-sided kernel regressions

$$\hat{y}^+ = \frac{\sum_{\substack{i: x_i > x_0 \\ i: x_i < x_0}} y_i K(x_i - x_0)}{\sum_{i: x_i < x_0} K(x_i - x_0)} \quad \hat{d}^+ = \frac{\sum_{\substack{i: x_i > x_0 \\ i: x_i < x_0}} d_i K(x_i - x_0)}{\sum_{i: x_i < x_0} K(x_i - x_0)}$$

Weighted average of  $y_i/d_i$  for  $i$  above cutoff  
similar for  $\hat{y}^-/\hat{d}^-$

- Basically comparing local averages just above & below cutoff

- Imposes no structure, thus has lower precision than parametric

- Kernel need not be the same/symmetric for above and below. Kernel need not be smooth

- Can e.g. apply equal weight to obs within self-defined range and zero weight outside

- Only doable when running variable is continuous.

If discrete (e.g. whole numbers  $\dots, x_{-1}, x_0, \dots$ ), then the sharp RD

$$\begin{aligned} \lim_{\epsilon \rightarrow 0^+} [E(Y_i | X_i = x_0 + \epsilon) - E(Y_i | X_i = x_0 - \epsilon)] &= \lim_{\epsilon \rightarrow 0^+} [E(Y_{i1} | X_i = x_0 + \epsilon) - E(Y_{i0} | X_i = x_0 - \epsilon)] \\ &= \lim_{\epsilon \rightarrow 0^+} [E(\Delta_i | X_i = x_0 + \epsilon) + E(Y_{i0} | X_i = x_0 + \epsilon) - E(Y_{i0} | X_i = x_0 - \epsilon)] \\ &= E(\Delta_i | X_i = x_0) + \underline{E(Y_{i0} | X_i = x_0) - E(Y_{i0} | X_i = x_{-1})} \neq 0!! \text{ Bias!} \end{aligned}$$

## PARAMETRIC ESTIMATOR

Necessary when running variable is discrete (e.g. whole nos)

$$Y_i = \alpha + \beta \Pr(D_i = 1 | X_i) + f(X_i) + \varepsilon_i$$

$f(X_i)$  : flexible function of running variable

approximates relationship between  $Y_i$  and  $X_i$  with a continuous function

$\uparrow$  flexibility  $\Rightarrow \downarrow$  power

$$\text{Sharp RD example of } \Pr(D_i = 1 | X_i) = \begin{cases} 1 & \text{if } X_i \geq x_0 \\ 0 & \text{if } X_i < x_0 \end{cases}$$

Generally,  $\Pr(D_i = 1 | X_i)$  just needs a large jump at  $x_0$  (not necessarily  $0 \rightarrow 1$ ) and it can be anything when not around  $x_0$ .

## SELECTIVITY BIAS

Tradeoff between bandwidth (bias) and precision/power (sample size)

- Check for robustness to bandwidth (choice of kernel)

## OTHER THREATS

### Spillovers

- Confounding unobservables also change around cutoff (i.e. RDDI)

- Cutoff ideally should be arbitrary rather than e.g. land borders

- Visualise covariates against running variable (should be smooth around  $x_0$ )

- Visualise density of running variable (should be smooth around  $x_0$ )

- Mistaking a non-linear relationship around cutoff as discontinuity

- Try more flexible specifications for  $f(X_i)$ , e.g. quadratic spline

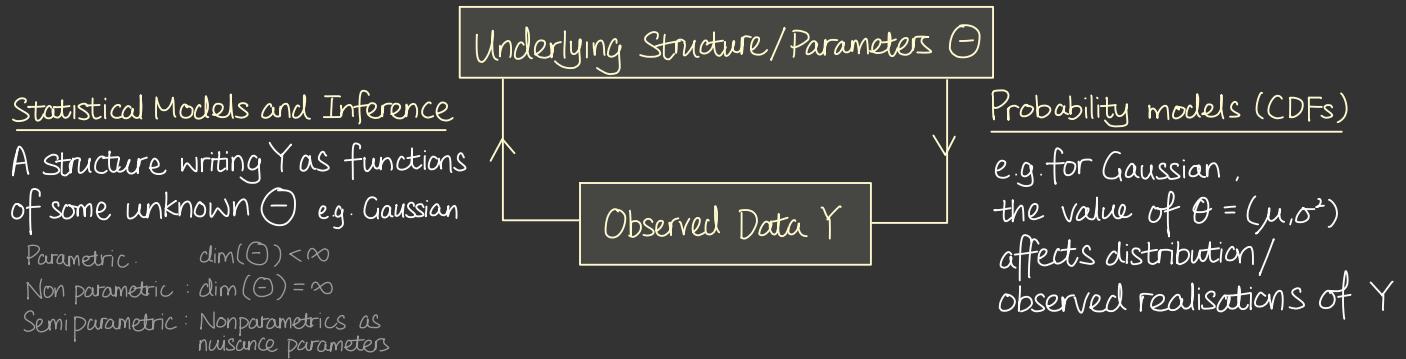
Revision Notes by Sally Yang

# PROBLEMS OF *Applied* ECONOMETRICS

LENT TERM

DR RACHAEL MEAGER  
LONDON SCHOOL OF ECONOMICS 2021/22

# INTRODUCTION



Weaker (not fewer) assumptions = Less structure  
 But less structure is not always better (possibly more noise)

Random variable  $Y$  has a set of possible outcomes  $\mathcal{Y}$  ('event space').  
 $y \in \mathcal{Y}$  is a particular outcome.

## Structure/Assumption

Draws  $Y_1, \dots, Y_N$  from population  $Y$  are random and i.i.d. with mean  $\mu$   
 each unit has equal chance to be selected

+  $Y_i$  has finite variance  $\sigma^2 < \infty$   
 counterexample: profits  
 if sample var is too large, pop var may not exist

## Asymptotic Result/Property

**LLN**  $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i \rightarrow \mu$  as  $N \rightarrow \infty$

**CLT**  $\sqrt{N}(\bar{Y} - \mu) \xrightarrow{N \rightarrow \infty} N(0, \sigma^2)$

## EXTERNAL VALIDITY

Sample may not be representative of population

Economic research doesn't insist on high  $R^2$

- Fitting the relationship in the sample too well may cause overfitting wrt pop<sup>1</sup>

## PROBABILITY MOMENT

$$E_f(y)(Y^i) = \int_y y^i f(y) dy$$

$E(Y) = \mu_Y$	Population mean
$E(Y^2) = \sigma_Y^2 + \mu^2$	since $\text{var}(Y) = E[Y - E(Y)]^2 = E(Y^2) - [E(Y)]^2$
$E(Y^3)$	Skewness/asymmetry
$E(Y^4)$	Kurtosis, how heavy the distribution's tails are (relative to Gaussian) High kurtosis $\Rightarrow$ extreme events more common. Mean/median no longer useful!

## LAW OF ITERATED EXPECTATIONS

$$\begin{aligned}
 E[E(Y|X)] &= E_{f(x)}[E_{f(y|x)}(Y|X)] = \int_X \left[ \int_y y f(y|x) dy \right] f(x) dx \\
 &= \int_X \left[ \int_y y f(y|x) f(x) dy \right] dx \stackrel{*}{=} \int_X \left[ \int_y y f(x|y) f(y) dy \right] dx \\
 &= \int_y y f(y) \left[ \int_X f(x|y) dx \right] dy = \int_y y f(y) dy = E(Y)
 \end{aligned}$$

By Bayes' Theorem,  $f(y|x)f(x) = f(x|y)f(y) = f(x,y)$

# POINT ESTIMATION

Guessing  $\theta$ , the estimand

The estimate  $\hat{\theta}(Y)$  will depend on data  $Y$  and have error (e.g.  $|\hat{\theta} - \theta|$ )

Estimators are functions of random variables (data  $Y$ ) so they are also random variables and have means, variances (which we also often want to estimate, e.g. for hypothesis testing)

**UNBIASEDNESS**  $E_{f(Y|\Theta)}[\hat{\theta}(Y)] = \theta$

**CONSISTENCY**  $\text{plim} [\hat{\theta}(Y)] = \theta \text{ as } N \rightarrow \infty$

**MINIMUM VARIANCE**  $\hat{\theta} = \underset{\hat{\theta} \in C}{\operatorname{argmin}} \operatorname{var}(\hat{\theta})$   
"efficient in the class  $C$ "

## EXAMPLE: OLS

Model relationship between  $Y$  and  $X$  as linear  $E(Y_i|X_i) = X_i'\beta \Leftrightarrow Y_i = X_i'\beta + \varepsilon_i$

Estimator  $\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - X_i'\beta)^2 \stackrel{EC221}{=} (X'X)^{-1}X'y$  is linear in  $y$

**GAUSS MARKOV THEOREM** If  $\begin{cases} A1 & y_i = X_i'\beta + \varepsilon_i \\ A2 & E(\varepsilon_i|X) = 0 \quad \forall i \\ A3 & \operatorname{Var}(\varepsilon_i|X) = \sigma^2 \quad \forall i \quad (\text{Homoscedasticity}) \\ A4 & \operatorname{Corr}(\varepsilon_i, \varepsilon_j|X) = 0 \quad \forall i \neq j \quad (\text{No autocorrelation}) \end{cases}$  NOTE:  $E(\varepsilon|X) = 0$  is sufficient for unbiasedness  
 $E(X'\varepsilon) = 0$  is necessary

Then  $\hat{\beta}_{OLS}$  is  $\underset{\substack{\text{Min var} \\ \hat{\beta} \text{ is a linear function of } Y}}{\underset{\uparrow}{\text{Best Linear}}} \text{ Unbiased Estimator}$

## INTERVAL ESTIMATION

✗ "The probability that  $\theta \in [a, b]$  is  $(1-\alpha) \cdot 100\%$ " ★  $\theta$  is not a random variable.

✓ "The probability that, upon repeated sampling of the data,  $\frac{(1-\alpha) \cdot 100\%}{(1-\alpha) \cdot 100\%}$  of interval estimates constructed this way contains the true  $\theta$ " coverage

## EXAMPLE: CONFIDENCE INTERVAL

$N$  i.i.d draws from  $Y$ ,  $Y_i \sim N(\theta, \sigma^2)$

By Gaussian property,  $\hat{\theta} = \bar{Y} \sim N(\theta, \frac{\sigma^2}{N})$   $\operatorname{se}(\hat{\theta}) \equiv \sqrt{\operatorname{var}(\hat{\theta})} = \frac{\sigma}{\sqrt{N}}$ .  

$$\frac{\hat{\theta} - \theta}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$
 If we don't know  $\sigma$ , we don't know  $\operatorname{se}(\hat{\theta})$  and will have to find an estimator for it  $\hat{\sigma}$ .

$$\Pr \left\{ -z_{1-\frac{\alpha}{2}} < \frac{\hat{\theta} - \theta}{\frac{\sigma}{\sqrt{N}}} < z_{1-\frac{\alpha}{2}} \right\} = \Pr \left\{ \hat{\theta} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} < \theta < \hat{\theta} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \right\} = 1-\alpha$$

$\uparrow$   
 $1-\frac{\alpha}{2}$ th quantile of  $N(0, 1)$

Confidence interval  $(\hat{\theta} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}})$  contains  $\theta$   $(1-\alpha) \times 100\%$  of the time.

# NULL HYPOTHESIS TESTING

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \neg H_0 \text{ or } \theta \neq \theta_0, \dots$$

**SIZE/CONFIDENCE**  $\Pr(\text{Type I error}) = \Pr(\text{Reject } H_0 | H_0) = \alpha$

When  $H_0$  is true, we want to fail to reject it  $(1-\alpha) \times 100\%$  of the time

**POWER**  $\Pr(\text{Type II error}) = \Pr(\text{Do not reject } H_0 | \neg H_0) = 1 - \beta$  Typically increases with  $\theta - \theta_0$

**FALSE DISCOVERY RATE** In multiple hypothesis testing

$$\Pr(H_0 \mid \text{Reject } H_0) = \frac{\Pr(H_0 \text{ and reject } H_0)}{\Pr(\text{Reject } H_0)} \neq \frac{\Pr(H_0 \text{ and reject } H_0)}{\Pr(H_0)} = \alpha$$

Expected no. of rejected null hypotheses that are actually true

People often conflate FDR with size and think  $FDR = \alpha$ .

But even if we limit our set of nulls to just the rejected ones / even if all nulls are rejected (i.e.  $\Pr(\text{Reject } H_0) = 1$ ), FDR  $\neq \alpha$  unless all nulls in the set are true

Even if  $\alpha = 0.05$ , FDR may be high if  $\begin{cases} \text{a large \% of tested nulls are true (high } \Pr(H_0) \\ \text{and low \% of tested nulls are rejected (low } \Pr(\text{reject } H_0)) \end{cases}$

★ z/t-testing does not tell us  $\Pr(H_0)$ ,  $\Pr(H_1)$  or  $\frac{\Pr(H_0)}{\Pr(H_1)}$   
z/t-stats always conditional on  $H_0$

BUT the LR test does give  $\frac{\Pr(\text{observing given data} | H_0)}{\Pr(\text{observing given data} | H_1)}$

## P-VALUE

Probability of observing data  $Y$  at least as extreme (i.e. adverse to the null) as the data that was actually observed,  $y$ , when  $H_0$  is true ★ always include

Lower p-value  $\Rightarrow$  more reasonable to think "no way that  $y$  was generated using  $\theta_0$ !"

1. Assume  $H_0$  true, so  $\theta = \theta_0$
2. Observe the implications of this assumption. Usually this means  $E(\hat{\theta}) = \theta_0$ .
3. Using what you know about  $f(\hat{\theta})$  and  $\hat{s}_e(\hat{\theta})$ , calculate  
 $\Pr(\hat{\theta} \text{ is at least as extreme as what you observe} | \theta = \theta_0)$

Again, cannot draw conclusions about how likely  $H_0$  is true ( $\Pr(H_0)$  or  $\Pr(H_1)$ ), no matter how low p-value is!  $\Pr(H_0 | \text{reject } H_0) \neq \Pr(\text{reject } H_0 | H_0)$

$\Pr(H_0 | \text{reject } H_0)$  or  $\Pr(H_0 | \text{data})$  in general is a posterior probability, requires a prior to calculate (Bayesian inference...). p values are not posteriors.

# THREATS TO IDENTIFICATION

- Functional form misspecification (A1)
- OVB, reverse causality, measurement error (A2)
  - Data not missing at random (MNAR) — selection bias
    - If  $\Pr(\text{data})$  is correlated with  $Y_i$ ,  $\varepsilon_i$  no longer i.i.d.
    - Selecting on  $X_i$  (MAR) affects external, not internal, validity
- Nonstationarity

# EXAM ANSWER TEMPLATES

# INTERPRET $\beta$ (structural equation)

The effect of a 1 unit of X increase in treatment on the ★ outcome for level of observation holding all other determinants of outcome constant (all else constant)

$$\begin{aligned} \text{Y on X: } & \Delta Y = \beta \Delta X \\ \ln Y \text{ on X: } \% \Delta Y &= 100\% \times (e^{\beta \Delta X} - 1) \approx 100\% \times \beta \Delta X \\ \text{Y on } \ln X: & \Delta Y = \beta \ln(1 + \% \Delta X) \approx \beta \% \Delta X \\ \ln Y \text{ on } \ln X: & \% \Delta Y = (1 + \% \Delta X)^{\beta} - 1 \approx \beta \% \Delta X \end{aligned}$$

# INTERPRET $\hat{\beta}$ (reading regression output)

In sample restriction (e.g. Indian villages), for level of observation,  
a 1 unit of X increase in treatment increases outcome  
by  $\star$  unit of Y, holding all other determinants of outcome constant.

# SIZE VS FALSE DISCOVERY RATE

$$\Pr(\text{Type I error}) = \Pr(\text{Reject } H_0 | H_0) = \frac{E(\text{No. rejected and true})}{E(\text{No. true})}$$

Expected number of true hypotheses we reject

$$\Pr(H_0 \mid \text{Reject } H_0) = \frac{E(\text{No. rejected and true})}{E(\text{No. rejected})}$$

Expected number of rejected hypotheses that were true

FDR is unknown and depends on prior probability of any hypothesis being true

For example, if all hypotheses were true, then all rejections are false, but the size is still  $\alpha$ .

# P-VALUE

Probability of observing data at least as extreme  
(i.e. adverse to the null) as data that  
was actually observed when  $H_0$  is true

# POWER

$$1 - \Pr(\text{Type II error}) = 1 - \Pr(\text{Do not reject } H_0 | \neg H_0)$$

- Increases when population variance/measurement error is low, sample size is high,  $|\theta - \theta_0|$  is large  
 $\uparrow$  size will  $\uparrow$  power (trading off Type I and II errors)

# SPILLOVERS

The treatment status  $D_i$  of  $i$  affects outcome  $Y_j(D_j)$  of  $j \neq i$ . Makes  $\hat{\beta}$  biased and inconsistent.

$\hat{\beta}$  will capture (direct effect - indirect effect), which may over- or under-state the direct effect and social effect (direct effect + indirect effect)

Changing level of analysis will protect us from bias, but reduce power as estimating at a higher level of variation introduces variance.

- True model is linear  
For non-linear see other page
- Unobserved entity-specific time-invariant OVB  $\alpha_i$

## FIXED EFFECTS

Allows  $\alpha_i$  and  $X_{it}$  to be correlated  
No assumptions on  $E(\alpha_i | X_{it})$ , unlike RE

$T=2$	Identical
$T > 2$	

Generally different in finite sample:  

- FE gives weighted mix of first and lagged diff under a parallel trends assumption
- Note that FE is more efficient than FD under FE3, as FE uses info from more lags

If too different:

- functional form misspecified
- or FE1 violated  $\Rightarrow \hat{\beta}_{FE}$  and  $\hat{\beta}_{FD}$  will generally have different ptms (ntrg)

## FIRST DIFFERENCES

$$\Delta Y_{it} = \Delta X_{it} + \Delta u_{it}$$

Allows  $\alpha_i$  and  $X_{it}$  to be correlated  
No assumptions on  $E(\alpha_i | X_{it})$ , unlike RE

### Assumptions

- FE1  $E(u_{it} | X_{i1}, \dots, X_{iT}, \alpha_i) = 0 \quad \forall t$  Strict exogeneity Violated by LDV ( $y_{it-1}$  on RHS)  
 FE2  $\text{Rank}(\tilde{X}'\tilde{X}) = K$  or  $N+K$   $\tilde{X}$  is transformed  $X$  to check  $X_{it+1}$  should not explain  $y_{it}$   
 FE3  $E(u_i u_i' | X_{i1}, \dots, X_{iT}) = \sigma^2 I_{T \times T}$  Stationarity:  $f(\dots, x_{t-1}, y_{t-1}, x_t, y_t, \dots)$  doesn't depend on  $t$   
 $\downarrow \text{FEI} \rightarrow \text{Var}(u_i | X_{i1}, \dots, X_{iT}) = \sigma^2 I_{T \times T}$  Violation: unit root, time trends, structural breaks  
 $\rightarrow u_{it}$  cannot be autocorrelated If violated, do FD, but FE still better than FD if  $\exists \theta$  s.t.  $Y - \theta X = C$   
i.e. cointegration/stationary residuals

### Approach

Same  $\hat{\beta}$   
different df

want  $\hat{\alpha}_i$

Within Regression  $\hat{\beta}$  unbiased, consistent

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + \varepsilon_{it} \xrightarrow[\substack{\text{Stack } N.T \\ NT \times 1}]{\substack{\text{NT} \\ K \times 1}} Y = X\beta + \varepsilon \quad df = NT - K$$

Dummies Regression  $\hat{\beta}$  unbiased, consistent if  $\hat{\beta}$  and  $\hat{\alpha}_i$  additively separable (linear)

$$y_{it} = \sum_{i=1}^N \alpha_i D_{ni} + X_{it}\beta + \varepsilon_{it} \xrightarrow[\substack{\text{Stack } N.T \\ = 1 \text{ iff } i=n \\ NT \times 1}]{\substack{\text{NT} \\ N \times 1 \\ K \times 1}} Y = D\alpha + X\beta + \varepsilon \quad df = NT - (N+K)$$

$\hat{\alpha}_i$  unbiased but inconsistent for small  $T$  (only consistent if  $T \rightarrow \infty$ ); has high var in short panels

### Problems

$\hat{\beta}, \hat{\alpha}_i$

Number of entity dummies  $\alpha_i$  increases with  $N$ , but each  $\alpha_i$  is estimated from  $T$  obs

$\hat{\alpha}_i$  inconsistent in  $N$  for fixed  $T$ , though unbiased

$\hat{\beta}, \hat{\alpha}_i$

Interactions of  $\alpha_i$  with  $X_{it}$   $\frac{\partial Y}{\partial X} = f(\alpha_i)$   $\hat{\beta}$  and  $\hat{\alpha}_i$  not additively separable

$\hat{\beta}, \hat{\alpha}_i$

Time-varying individual-specific confounders  $v_{it}$

### Assumptions

### Robustness

If FDI violated

- FD1  $E(u_{it} | X_{i1}, \dots, X_{iT}, \alpha_i) = 0 \quad \forall t$  guarantees  $E(\Delta X' \Delta u_{it} | \Delta X) = 0$  necessary for consistency  
 FD2  $\text{Rank}(\Delta X' \Delta X) = K$  to check  $X_i$  should not explain  $\Delta y_{it}$   
 FD3  $E(\Delta u_i' \Delta u_i | X_{i1}, \dots, X_{iT}) = \sigma^2 I_{T \times T}$   $\Delta u_{it}$  should be autocorrelated, e.g. random walk  $u_{it} = u_{it-1} + v_{it}$   
If not,  $\text{Cov}(\Delta u_{it}, \Delta u_{it-1}) = -\sigma^2$  regression to the mean

Additionally include covariates in levels  $X_{t-1}$  in FD regression and do F test. Should be insignificant.  
If significant, base model may have omitted a linear trend  $X_{tt}$  ( $\Delta X_{tt} = x_t$ )

### ANDERSON-HSIAO IV

Assume sequential exogeneity  $E(u_{it} | X_{i1}, \dots, X_{it}, \alpha_i) = 0$

Use lag  $X_{it-1}$  to instrument for  $\Delta X_{it}$

Arellano-Bond: all lags valid, so add more!

✓ Validity:  $E(X_{it-1} \Delta u_{it}) = 0$

? Relevance:  $\text{Cov}(X_{it-1}, \Delta X_{it}) = \text{Cov}(X_{it}, X_{it-1}) - \text{Var}(X_{it-1})$

### Weak IV

- Different lags may have opposite effects on  $\Delta X_{it}$  and cancel out
- $\uparrow$  No. of lags used  $\Rightarrow \downarrow$  sample size
- If  $X_{it}$  and  $X_{it-1}$  highly correlated,  $\text{Cov}(X_{it}, X_{it-1}) \approx \text{Var}(X_{it-1})$

$y_i$  is not linear in  $\beta$

**TOBIT**  $y_i = \max(0, X_i\beta + u_i)$   $u_i \sim N(0, \sigma_u^2)$

OLS is biased and inconsistent since model is non-linear

Ignoring 0s creates sample selection bias  $E(y_i|y_i>0, X_i) = X_i\beta + \frac{\sigma_u^2 \varphi(\frac{X_i\beta}{\sigma_u})}{\Phi(\frac{X_i\beta}{\sigma_u})}$

↓ omitted variable: "some non-zero term"

## Censored $y_i$

e.g. WTP, demand (comes!?)

$$\Pr(y_i=0|X_i) = \Pr(u_i < -X_i\beta|X_i) = 1 - \Phi(\frac{-X_i\beta}{\sigma_u})$$

$$\Pr(y_i|y_i>0, X_i) = f(u_i|X_i) = \varphi(\frac{y_i - X_i\beta}{\sigma_u}) \quad L \stackrel{iid}{=} \prod_i^N \left[ \frac{1}{\sigma_u} \varphi(y_i - \frac{X_i\beta}{\sigma_u}) \right]^{1(y_i>0)} \left[ 1 - \Phi(\frac{X_i\beta}{\sigma_u}) \right]^{1(y_i=0)}$$

## Binary

Assume  $y_i$  is Bernoulli  
 $E(y_i|X_i) = \Pr(y_i=1|X_i)$

**LPM**  $E(y_i|X_i) = \Pr(y_i=1|X_i) = X_i\beta$

✓ MTE doesn't depend on  $X_i$  value ✓ May be a good approximation

✗ Inconsistent and biased unless the true  $X_i\beta \in [0, 1]$   $\forall i$  beyond that,  $\epsilon_i=0$  but  $\hat{\epsilon}_i \neq 0$   
 ! At least verify  $\hat{y}_i \in [0, 1] \quad \forall i$

**BOUNDED LPM**

$E(Y_i|X_i) = \Pr(Y_i=1|X_i) = \min[\max(X_i\beta, 0), 1]$



**Latent variable**

$$y_i = \mathbb{1}\{y_i^* = X_i\beta + u_i > 0\}$$

$$u_i|X_i \sim G \quad \forall i. \quad E(u_i|X_i) = 0$$

↑ Symmetric

$$\begin{aligned} \Pr(y_i=1) &= \Pr(y_i^*|X_i) \\ &= \Pr(u_i > -X_i\beta|X_i) \\ &= G(X_i\beta) \end{aligned}$$

## Models

**LOGIT**  $\Pr(y_i=1|X_i) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$   $L \stackrel{iid}{=} \prod_i^N \left[ \frac{e^{X_i\beta}}{1+e^{X_i\beta}} \right]^{y_i} \left[ \frac{1}{1+e^{X_i\beta}} \right]^{1-y_i}$

✓ In panel, can "delete"  $\{\alpha_i\}^N$  by conditioning on entity-specific means  $\{\sum_{t=1}^T y_{it}\}^N$

- $\{\bar{y}_{it}\}^N$  is a sufficient statistic for  $\{\alpha_i\}^N$
- A statistic  $S$  is sufficient for a parameter  $\theta$  if  $Y|S$  doesn't depend on  $\theta$
- Factorisation criterion:  $S(Y)$  sufficient for  $\theta$  iff  $\exists g_\theta, h$  s.t.  $f_\theta(Y) = g_\theta(S(Y)) h(Y)$

likelihood function

**PROBIT**  $\Pr(y_i=1|X_i) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i\beta)^2}{2}}$   $L \stackrel{iid}{=} \prod_i^N \left[ \Phi(X_i\beta) \right]^{y_i} \left[ 1 - \Phi(X_i\beta) \right]^{1-y_i}$

**TOBIT**  $\Pr(y_i=1|X_i) = \max(0, X_i\beta + u_i)$   $u_i \sim N(0, \sigma_u^2)$   $L \stackrel{iid}{=} \prod_i^N \left[ \frac{1}{\sigma_u} \varphi(y_i - \frac{X_i\beta}{\sigma_u}) \right]^{1(y_i>0)} \left[ 1 - \Phi(\frac{X_i\beta}{\sigma_u}) \right]^{1(y_i=0)}$

## Problems

### Latent heteroscedasticity

$\text{Var}(y_i^*|X_i) = \text{Var}(u_i|X_i) = \sigma_i^2$  must be explicitly modelled in (rescale  $X_i, y_i$  by  $\frac{1}{\sigma_i}$ )  
 If not,  $\hat{\beta}$  may accidentally capture effect of  $X_i$  on  $y_i$  through  $\sigma_i$

### OVB

Must control for all predictors of  $y_i$ , even if uncorrelated with  $X_i$  e.g. measurement error in  $y_i$   
 Any omitted  $w_i$  makes  $\hat{\beta}$  inconsistent ( $w_i$  and  $X_i$  not additively separable in nonlinear model)

✗ Unobserved heterogeneity  $\alpha_i$  in panel (Exception: logit)

- Since FE, FD allow  $E(\alpha_i|X_{it})$  to be anything,  $\alpha_i$  likely correlated with  $X_{it}$
- Since  $\beta$  now depends on  $X_{it}$ , likely also depends on  $\alpha_i$  FE, FD fail by incidental parameters

$$Y \sim f(y|\theta)$$

$$\mathcal{L}(\theta) = f(y|\theta)$$

$$l(\theta) = \log f(y|\theta)$$

## MAXIMUM LIKELIHOOD $\hat{\theta}_{MLE}$

$$\hat{\theta}_{MLE} \equiv \operatorname{argmax} \mathcal{L}(\theta)$$

$$= \operatorname{argmax} l(\theta)$$

Conditions	Univariate	Score function FOC: $\frac{\partial l(\hat{\theta}_{MLE})}{\partial \theta} = 0$	SOC: $\frac{\partial^2 l(\hat{\theta}_{MLE})}{\partial \theta^2} < 0$
	Multivariate	FOC: $\nabla l(\hat{\theta}_{MLE}) = J(\theta) = 0$ Jacobian	SOC: $\nabla^2 l(\hat{\theta}_{MLE}) = H(\theta) = 0$ Hessian

## Properties

Consistent

Efficient Minimum Variance in class of consistent estimators

Asymptotically Normal for doing inference

Under  $H_0: \theta = \theta_0$ ,  $\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \approx N(0, I^{-1}(\theta_0))$

- Inverse Fisher Information  $I^{-1}(\theta_0) = E\left[-\frac{\partial^2 l(\hat{\theta}_{MLE})}{\partial \theta^2}\right]^{-1}$
- Large 2<sup>nd</sup> derivative  $\rightarrow f$  very concave at  $\hat{\theta}_{MLE}$   
 $\rightarrow$  More certain that  $\hat{\theta}_{MLE}$  is maximiser  $\rightarrow$  Smaller  $\text{Var}(\hat{\theta}_{MLE})$

## Solving

If  $l$  is globally smooth and concave

If not

## NUMERICAL OPTIMISATION

Newton-Raphson procedure : easiest to describe when tested

1	Start at some $\theta_0$
2	Calculate $l'(\theta_0)$
3	If $l'(\theta_0) > 0$ , go to $\theta_1 > \theta_0$ where $\theta_1 = \theta_0 + A$ If $l'(\theta_0) < 0$ , go to $\theta_1 < \theta_0$ where $\theta_1 = \theta_0 - A$
4	Repeat ②-③. Stop when $ l'(\theta)  \leq B$

! May not converge (i.e. find solution); can get stuck

! Always check for sensitivity to starting point, tuning parameter, tolerance  
 $\hookrightarrow$  If sensitive, algorithm may be 'stuck' at a wrong solution (try changing tolerance or algorithm) or solution not unique (calculate  $L$  at each solution, choose largest)

Calculus  
Solve  $\hat{\theta}_{MLE}$  from FOC

Data  $Y$ , model with parameters  $\Theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$

characterised by  $R$  moment conditions

$$E[f(Y, \Theta)] = 0 \quad f(Y, \Theta) = \begin{pmatrix} f_1(Y, \Theta) \\ \vdots \\ f_R(Y, \Theta) \end{pmatrix}$$

For  $N$  draws,

$$g_N(\Theta) = \frac{1}{N} \sum_{i=1}^N f(Y_i, \Theta) = 0$$

## GMM

$$\hat{\Theta}_{GMM} = \underset{\Theta}{\operatorname{argmin}} g_N(\Theta)' W_N g_N(\Theta)$$

Minimise (weighted) sum of squares

! Minimiser must be unique

→ Minimises  $\text{Var}(\hat{\Theta}_{GMM})$

**Optimal weighting matrix**  $W_N^* = E[f(Y, \Theta) f(Y, \Theta)']^{-1}$

- Inverse of covariance of sample moments
- Highly weight / emphasise sample moments with low var (more precisely estimated)
- "Plug-in" estimator: Get  $\hat{\Theta}$  from unweighted GMM, then calculate  $\hat{W}_N^* = \frac{1}{N} \sum_{i=1}^N f(Y_i, \hat{\Theta}) f(Y_i, \hat{\Theta})'$

## Properties

Consistent

Asymptotically normal  $\sqrt{N}(\hat{\Theta}_{GMM} - \Theta) \xrightarrow{D} N(0, V)$  for doing inference

Efficient (Min var in class of asy. normal estimators)

## OVERIDENTIFICATION (SARGAN'S J) TEST

If we have more moments than needed to get a single solution (For linear system,  $R > K$ )

Test statistic:  $ng_N(\hat{\Theta}_{GMM})' W_N^* g_N(\hat{\Theta}_{GMM}) \sim \chi^2_{R-K}$  under  $H_0$ : All population moments are zero

- ! If the GMM obj fn value is too big to be compatible with  $H_0$ , discard suspicious moments  
Tests if moments are internally consistent, not if they're valid (they may be invalid in similar ways)

## IV ISSUES

$$Y = X\beta + \varepsilon$$

$$X = Z\pi + \epsilon$$

**BLUNT** For a given  $Y, Z$  can be a valid instrument for at most one  $X$



## WEAK

$\pi \neq 0$  (still valid) but almost 0. Conventional asymptotic approximations do not capture it properly and will generate errors not well-behaved.

$$\beta - \hat{\beta} = (Z'X)^{-1} Z'\varepsilon \xrightarrow{N \rightarrow \infty} (0)^{-1} 0$$

! Using many weak IV on RHS may actually make it worse

- Homoscedastic  
Heteroscedastic

## F>10

## ANDERSON-RUBIN CI

$AR(\beta_0) \sim \chi^2_K$  under  $H_0: \beta = \beta_0$

Doesn't use  $\pi$ ; Robust to weak IV

- Test a grid of  $\beta_0$ s to generate CI, keeping  $\beta_0$ s that aren't rejected
- Model may be misspecified if
  - CI is empty or unbounded

F stat is low but CI is very tight (nearly empty). Misleading!

Non-linear IV

## FORBIDDEN REGRESSION

- Run  $X_i = \pi Z_i + \gamma_i$  to get  $\hat{X}_i$  and plug  $\hat{X}_i^2$  into second stage  $X_i^2$  ( $E(\hat{X}_i^2) \neq [E(\hat{X}_i)]^2$ )
- Can instrument using  $Z_i^2$  but likely weak
  - No reason why strong first stage for  $X_i, Z_i$  should hold for  $X_i^2, Z_i^2$
- Sometimes just fitting the linear IV still gives LATE (even if true model is non-linear)

# STANDARD ERRORS

N/N(cluster) > 50? ★		No	
Yes	Cross section	Homoscedastic	Bootstrap SEs $\frac{1}{N-K} \sum_i \hat{u}_i^2$ unbiased, consistent in N
	Heteroscedastic	Assume/know $\{\sigma_i^2\}_{i=1}^N$ ★ Assume/know structure of heteroscedasticity e.g. parameterise $\sigma_i^2 = f(X)$	Weighted Least Squares t-stats exact ( $\sim N, X^2$ ) Feasible Weighted Least Squares Fit f, get consistent $\hat{\sigma}_i^2$ . t-stats approximate ( $\sim t, F$ )
		Don't	White RSEs $\frac{1}{N-K} \sum_i \hat{u}_i^2 x_i' x_i$ consistent in N
Panel	No autocorrelation ★ additionally consider...		Arellano CSE by Entity $\frac{1}{N} \sum_i \hat{x}_i' \hat{u}_i \hat{u}_i' \hat{x}_i$ consistent in N for fixed T More conservative (gives larger SEs) if ICC > 0
	Autocorrelation	Assume/know structure of autocorrelation $\rho$ ★ (Parameterise) Do not	Newey-West HAC $\frac{1}{T} \frac{\sigma^2}{(\hat{x}'_x)^2} (1 + 2 \sum_{j=1}^{T-1} \frac{T-j}{T} p_j)$ formula is for TS Estimate $\sum_{j=1}^{T-1} \frac{T-j}{T} p_j$ with $\sum_{j=1}^{m-1} \frac{m-j}{m} \hat{p}_j$ where m grows slower than T e.g. $m = \frac{3}{4} T^{\frac{2}{3}}$ When $T \rightarrow \infty$ , $N(p_j) \rightarrow \infty$ (incidental parameters problem) so use less lags than T need large T and thin-tailed $x_t, u_t$ for consistency
			Arellano CSE by Entity $\frac{1}{N} \sum_i \hat{x}_i' \hat{u}_i \hat{u}_i' \hat{x}_i$ consistent in N for fixed T More conservative (gives larger SEs) iff ICC > 0

★ There is no finite sample unbiased SE estimator except in homoscedasticity/WLS/FWLS! Everything else works asymptotically (consistency)!

★ White RSEs  $\frac{1}{NT-N-K} \sum_i^N \hat{x}_i' \hat{x}_i \hat{u}_i \hat{u}_i'$  inconsistent in N for fixed T even without autocorrelation

★ Formulae shown are for "core component" of  $SE(\hat{\beta})$ , not  $SE(\hat{\beta})$

★ The more assumptions a procedure makes, the lower the SE value (e.g. HAC < CSE as HAC leverages the structure of stationarity)

## CLUSTERING

Cluster by group : SUTVA violation, Treatment assigned at higher level, Intra cluster correlation

Cluster by time : Spatial spillovers (across entities)

N(Cluster) < 50: Bootstrap performs better

- Extremum statistic (max, min)
- Statistic not a smooth function of data (median ( $X$ ) when  $X$  bunches at 0)
- Random variables in the data have unbounded variances kills LLN
- Statistic changes the support of the distribution of data

<p>Yes</p> <hr/> <p>Don't bootstrap</p>	<p><b>BOOTSTRAP</b></p> <p>Given a sample <math>(X, Y) = \{(X_i, Y_i)\}_{i=1}^N</math> i.e. <math>N</math> pairs of <math>(X_i, Y_i)</math></p> <p style="text-align: center;">↑ indirectly determines <math>Y_i = X_i\hat{\beta} + \hat{\varepsilon}_i</math></p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 10%;">1</td><td>Sample <math>&lt; N</math> pairs of <math>(X_i, \hat{\varepsilon}_i)</math> with replacement to make one bootstrapped sample <math>(X, \hat{\varepsilon})_b</math></td></tr> <tr> <td>2</td><td>Run estimation on <math>(X, \hat{\varepsilon})_b</math> to get <math>\hat{\beta}_b</math></td></tr> <tr> <td>3</td><td>Repeat <math>B</math> times to get <math>\{\hat{\beta}_b\}_{b=1}^B</math></td></tr> <tr> <td>4</td><td><math>\widehat{\text{Var}}(\hat{\beta}_{\text{original}}) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b - \hat{\beta}_{\text{original}})^2</math></td></tr> </tbody> </table>	1	Sample $< N$ pairs of $(X_i, \hat{\varepsilon}_i)$ with replacement to make one bootstrapped sample $(X, \hat{\varepsilon})_b$	2	Run estimation on $(X, \hat{\varepsilon})_b$ to get $\hat{\beta}_b$	3	Repeat $B$ times to get $\{\hat{\beta}_b\}_{b=1}^B$	4	$\widehat{\text{Var}}(\hat{\beta}_{\text{original}}) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b - \hat{\beta}_{\text{original}})^2$
1	Sample $< N$ pairs of $(X_i, \hat{\varepsilon}_i)$ with replacement to make one bootstrapped sample $(X, \hat{\varepsilon})_b$								
2	Run estimation on $(X, \hat{\varepsilon})_b$ to get $\hat{\beta}_b$								
3	Repeat $B$ times to get $\{\hat{\beta}_b\}_{b=1}^B$								
4	$\widehat{\text{Var}}(\hat{\beta}_{\text{original}}) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b - \hat{\beta}_{\text{original}})^2$								
<p>No</p> <hr/> <p><b>X is fixed</b> e.g. RCT</p> <p>Cluster/Serial correlation</p> <p>better account for cluster effects</p>	<p><b>Sample just <math>\hat{\varepsilon}_i</math></b></p> <p>Bootstrapped sample : <math>(X_i, Y_i)</math> where <math>Y_i = X_i\hat{\beta} + \hat{\varepsilon}_i</math> <math>\hat{\varepsilon}_i</math> must be i.i.d (no heteroscedasticity, etc)</p> <p><b>Block/Cluster Bootstrap</b></p> <p>Sample clusters with replacement Can also sample within clusters; usually not done Blocks must be independent</p> <p><b>Wild Cluster Bootstrap</b></p> <p>Additionally flip sign of residuals <math>\hat{\varepsilon}</math> with <math>\frac{1}{2}</math> probability for any cluster (then <math>\hat{Y} = \beta X + \hat{\varepsilon}_{\text{new}}</math>) ↑ <math>\widehat{\text{var}}(\hat{\beta})</math></p>								