

Revision Notes by Sally Yang

# INTRODUCTION TO ECONOMETRICS

MICHAELMAS TERM

DR CANH THIEN DANG  
PROFESSOR STEVE PISCHKE  
LONDON SCHOOL OF ECONOMICS 2020/21

# CAUSATION

Correlation does not imply causation

DIRECT

$$A \longrightarrow B$$

BIDIRECTIONAL

$$A \iff B$$

REVERSE

$$A \longleftarrow B$$

INDIRECT

$$A \rightarrow C \rightarrow B$$

CONFOUNDER

$$C \begin{matrix} \nearrow & \searrow \\ A & B \end{matrix}$$

COINCIDENTAL

$$A \qquad B$$

All statistical techniques only establish **associations**.

Causation requires **interpretation**.

Causation does not imply correlation



$\xrightarrow{x \rightarrow Y}$  Google it!  
Directed acyclic graphs (DAGs) are often used  
with the usual potential outcomes framework  
( $Y_{10}, Y_{11}, E(\cdot)$ ) to represent and explain causation !!

# COUNTERFACTUALS

Explained using the potential outcomes framework

For person  $i$

$$\boxed{\text{TREATMENT}} \quad D_i = \{0, 1\} \quad \boxed{\text{OUTCOME}} \quad Y_i$$

**AIM** Find effect of treatment on  $i = Y_{ii} - Y_{oi}$

## COUNTERFACTUALS

Before treatment is assigned,

$$\text{Potential (counterfactual) outcomes} = \begin{cases} Y_{ii} & \text{if } D_i = 1 \\ Y_{oi} & \text{if } D_i = 0 \end{cases}$$

★ All potential outcomes are counterfactual before the assignment decision is made.

Once treatment is assigned / conditional statement made,

one of the potential (counterfactual) outcomes becomes the actual (observed) outcome.

$$\text{Observed outcome for person } i \text{ is } Y_i = \begin{cases} Y_{ii} & \text{if } D_i = 1 \\ Y_{oi} & \text{if } D_i = 0 \end{cases} = Y_{oi} + (Y_{ii} - Y_{oi}) D_i$$

can only observe one or the other for person  $i$

$$\text{Counterfactual outcome for person } i \text{ is } = \begin{cases} Y_{oi} & \text{if } D_i = 1 \leftarrow \text{If the treated person were untreated (unobservable)} \\ Y_{ii} & \text{if } D_i = 0 \leftarrow \text{If the untreated person were treated (unobservable)} \end{cases}$$

$$E(Y_i | D_i = 1) = E(Y_{ii} | D_i = 1)$$

Average outcome of the treated group      Average treated outcome of the treated group  
should hopefully be intuitive ..

$$\text{BUT } E(Y_{ii} | D_i = 1) \neq E(Y_{ii}).$$

Average treated outcome of the treated group      Average treated outcome of everyone in population

Why?

If treated group differs systematically from the rest of the (untreated) population ...

Take EC333 for further elaboration

## SELECTION PROBLEM

Observed difference in average health

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

(unobservable)  
untreated outcomes  
of those who were treated

(observed)  
untreated outcomes  
of those who were not treated

$$= E(Y_{ii} | D_i = 1) - E(Y_{oi} | D_i = 0)$$

$$= E(Y_{ii} | D_i = 1) - E(Y_{oi} | D_i = 1) + E(Y_{oi} | D_i = 1) - E(Y_{oi} | D_i = 0)$$

$$= E(Y_{ii} - Y_{oi} | D_i = 1) + E(Y_{oi} | D_i = 1) - E(Y_{oi} | D_i = 0)$$

selection bias

Something else cause them to "start out" differently

Average treatment  
effect on the treated  
(ATET)

## SELECTION BIAS

Difference in potential untreated outcomes between treatment / control groups.  
Caused by assignment rules (self-selection, in this case)

## RANDOM ASSIGNMENT of treatment $D_i$

Makes  $D_i$  statistically independent of potential outcomes  $Y_{ii}$  and  $Y_{oi}$

$$E(Y_{oi} | D_i = 1) = E(Y_{oi} | D_i = 0)$$

$$\Rightarrow E(Y_i | D_i = 1) - E(Y_i | D_i = 0) = E(Y_{ii} - Y_{oi} | D_i = 1) = E(Y_{ii} - Y_{oi})$$

- Can mitigate selection bias
- Allows for inference of causality assuming LLN / same internal characteristics
- Only say "can make causal inference" in exam if treatment is (as good as) randomly assigned

## LAW OF LARGE NUMBERS

As  $n \rightarrow \infty$ ,  $\bar{Y}_i \rightarrow \mu$

## BALANCE

No systematic differences in characteristics across all groups

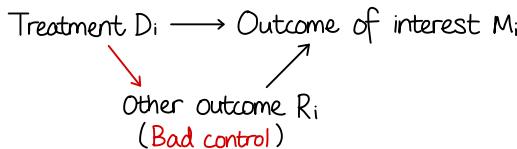
- Achievable with large  $n$  and randomisation  $\rightarrow$  LLN
- Can check for balance amongst pre-treatment observed characteristics
- If randomisation works, should also work for unobserved characteristics (motivations, incentives, etc.)
  - ★ Cannot test. Must assume  $\rightarrow$  External validity concerns
  - ★ Balance amongst observed characteristics good indicator of balance amongst unobserved
- Makes counterfactual outcomes comparable  $E(Y_{oi} | D_i = 1) = E(Y_{oi} | D_i = 0)$

## REPLICATION

- Richer inference about effect
- Boost external validity
- Results may differ based on population / experimental method

# GOOD, BAD, NEUTRAL CONTROLS

## EG OF A BAD CONTROL

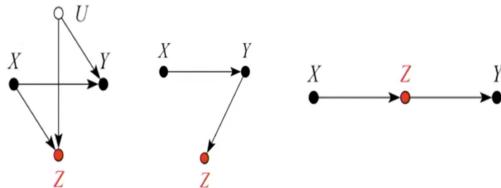


$$\begin{aligned} E(M_i|R_i=r, D_i=1) - E(M_i|R_i=r, D_i=0) &= E(M_{ii}|R_i=r, D_i=1) - E(M_{0i}|R_i=r, D_i=0) \\ &= E(M_{ii}|R_{ii}=r) - E(M_{0i}|R_{0i}=r) \\ &= E(M_{ii} - M_{0i}|R_{ii}=r) + E(M_{0i}|R_{ii}=r) - E(M_{0i}|R_{ii}=r) \\ &\quad \text{Causal effect conditional on small class and reading test score} \end{aligned}$$

Selection bias

## BAD CONTROLS

When  $Z$  is outcome of  $X$  and correlated with  $Y$

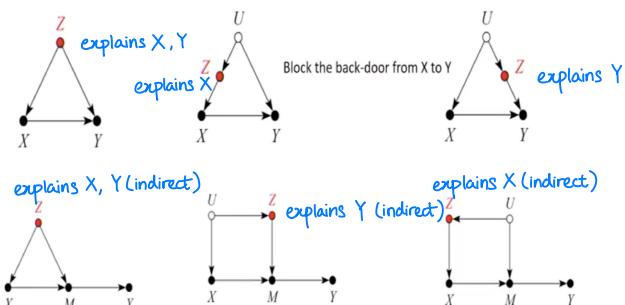


Even with perfect randomisation,  $Z$  will introduce Selection bias since it is (an outcome of  $X$  and so) not balanced across treatment and control groups

## GOOD CONTROLS

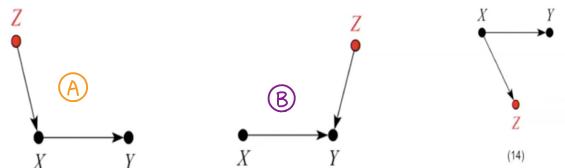
- Either determined prior to the process that leads from the treatment and outcomes
- Or immutable characteristics of the units of observation
- Basically, not an outcome of (not affected by)  $X_i$ .

U: Confounder (always unobservable)  
Z: Control (to approximate confounder)  
→: Correlation



★ Good control explains (blocks causality pathway from confounder to)  $X$  and/or  $Y$ , but is not a result of  $X$

## NEUTRAL CONTROLS



Possibly bad for precision  
(increase the variance of residual)

Possibly good for precision  
(reduce the variance of residual)

Possibly bad for precision,  
but does not harm our  
treatment effect since does  
not correlate with  $Y$

SE

(14)

## A MULTICOLLINEARITY

When you include a covariate that is highly correlated with ('almost as good as') an existing regressor,  $\text{se}(\hat{\beta}) \uparrow$

- $\text{Var}(\tilde{D}_i) \ll \text{Var}(D_i)$  Very little noise left in  $D_i$  since regressor nearly perfectly explains  $D_i$
- $\text{Var}(e_i^t) \approx \text{Var}(e_i^s)$
- Since  $\text{se}(\hat{\beta}) = \sqrt{\frac{\text{Var}(e_i)}{n\text{Var}(\tilde{D}_i)}}$ ,  $\text{se}(\hat{\beta}) \uparrow$
- Cannot estimate a model if regressors are collinear
- If near-multicollinear, will artificially ↑ se (↓ precision) and unstable est
- Either: get more/better data, drop some controls, or accept limitations
- No need to care about multicollinearity between controls

## B UNCORRELATED COVARIATE

$$Y_i = \alpha^L + \beta^L D_i + \gamma X_i + e_i^t \quad \text{vs} \quad Y_i = \alpha^S + \beta^S D_i + e_i^s$$

- If  $\text{Cov}(D_i, X_i) = 0$ ,  $\gamma \neq 0$ , then  $E(\beta_L) = E(\beta_S)$ .

Having or not having  $X_i$  doesn't affect the  $\beta$  estimate's value

- By ANOVA theorem,  $\text{Var}(e_i^t) < \text{Var}(e_i^s)$  so  $\text{se}(e_i^t) < \text{se}(e_i^s)$  'soak up' residual var esp if  $X_i$  is strong predictor of  $Y_i$
- $\text{se}(\hat{\beta}) \downarrow \uparrow$  precision

**POPULATION MEAN**  $\mu = E(Y_i)$

- A parameter (fixed number)

**SAMPLE MEAN**  $\bar{Y}_n = \text{Avg}_n(Y_i) = \frac{1}{n} \sum_{i=1}^n Y_i$

- An estimator of  $\mu$
- Estimator: a type of sample statistic: a function of sample data used to infer the parameter
- Estimate: The value the estimator ( $\bar{Y}_i$ ) takes on for a particular sample  
a function

**SAMPLING VARIABILITY**  $\bar{Y}_n - \mu$

**UNBIASEDNESS**  $E(\bar{Y}_n) = E(Y_i) = \mu$ , so  $E(\bar{Y}_n - \mu) = 0$

- Holds for all  $n$
- Unbiasedness means in many repeated random samples, the average of each sample's average will equate the population mean.

**VARIANCE**  $\text{Var}(Y_i) = E(Y_i - \mu)^2$

$$\begin{aligned}\text{Var}(\bar{Y}_i) &= E(\bar{Y}_i - E(\bar{Y}_i))^2 = E(\bar{Y}_n - \mu)^2 \text{ if } \bar{Y}_n \text{ is unbiased} \\ &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) \text{ since } Y_i \text{ are randomised (independent)}\end{aligned}$$

Assume  $\text{Var}(Y_i) = \sigma^2$  constant (homoscedastic). Then  $\text{Var}(\bar{Y}_n) = \frac{\sigma^2}{n}$

**SAMPLING S.D.**  $SE(\bar{Y}_n) = \sqrt{\text{Var}(\bar{Y}_n)} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{E(Y_i - \mu)^2}}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}}{\sqrt{n}}$

- A population parameter

**STANDARD ERROR**  $\hat{SE}(\bar{Y}_n) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}{\sqrt{n}}$

- estimator of  $SE(\bar{Y}_n)$

Treatment group  
↓  
control group  
↓

\* We do not use Bessel's correction as  
1. still not unbiased 2. Doesn't minimise MSE

**SAMPLING SD FOR THE DIFFERENCE IN MEANS**  $\text{Var}(\bar{Y}_n^1 - \bar{Y}_m^2) = \frac{\sigma_n^2}{n} + \frac{\sigma_m^2}{m}$ , so  $SE(\bar{Y}_n^1 - \bar{Y}_m^2) = \sqrt{\frac{\sigma_n^2}{n} + \frac{\sigma_m^2}{m}}$

\* Assumes independence (due to random assignment) and homoscedasticity

# CONDITIONAL EXPECTATION FUNCTION

$E(Y_i | X_1, \dots, X_K)$  for a CEF with K conditioning variables

$E(Y_i | X_1 = x_1, \dots, X_K = x_K)$  is one point in the range of the CEF

CEF is an estimator of  $Y_i$  for a particular  $X_i$  value

Summarises bivariate relationship nonparametrically (no functional form)

Best linear predictor to the CEF and to the data are the same

$$\min_{a,b} E (E(Y_i | X_i) - a - bX_i)^2 = \min_{a,b} E(Y_i - a - bX_i)^2$$



# REGRESSION

CEF line can be overfit, so instead use  $\min_{a,b} E \left\{ \underbrace{(E(Y_i | X_i) - a - bX_i)^2}_{\text{Residual Sum of Squares (RSS)}} \right\}$

We find  $(\alpha, \beta) = (\arg \min_{a,b} E \left\{ (E(Y_i | X_i) - a - bX_i)^2 \right\})$

Diff wrt a and b  
set = 0, solve

OLS for Bivariate :

$$\beta = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)}$$

if uncorrelated, no slope  
Need variability in  $X_i$

$$\alpha = E(Y_i) - \beta E(X_i)$$

Interpret:  $\alpha$  = average outcome when all regressors = 0  
(when  $X_1, \dots, X_n = 0$ ,  $Y_i = \alpha + e_i$  and  $E(Y_i) = \alpha$ )

# REGRESSION

$Y$	$x_1, x_2, \dots, x_k$
Dependent variable	Independent variables
Explained variable	Explanatory variables
Response variable	Control variables
Predicted variable	Predictor variables
Regressand	

\* Should not call  $e_i$  'error term' as error term is undesirable.

## TERMS IN A REGRESSION

Outcome/Dependent variable

Regressor/Covariate / Indep. var  
(Any var on RHS)

Residual

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 PI_i + e_i$$

Control variables (Regressors used to hold all confounders constant)

Causal variable of interest

→ Causal effect of interest

## PROPERTIES

$$Y_i = \alpha + \beta X_i + e_i$$

$$\beta = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)}$$

$$\alpha = E(Y_i) - \beta E(X_i)$$

- $E(e_i) = 0$  so  $E(Y_i) = \alpha + \beta E(X_i)$ .
- So regression line passes through the mean  $(E(X_i), E(Y_i))$
- Can be used to check regression
- $\text{Cov}(e_i, X_i) = 0$ . Residual is uncorrelated with regressors
  - $E(X_i | e_i) = E(X_i)$  Mean independence
  - No more info can be extracted from  $e_i$
  - Variation in  $Y_i$  partitioned into :
    - Variation related to  $X_i$ : the regression line
    - Variation unrelated to (can't be explained by)  $X_i$ :  $e_i$
- Note: "fitted value",  $\hat{Y}_i = \alpha + \beta X_i$

## ADVANTAGES

- OLS is BLUE of  $\beta$ , best linear approximation of  $E(Y|X)$
- For large samples, OLS often produces normally-distributed residual
- For finite samples, OLS often produces normally/t-distributed estimates
- If covariates are normally distributed, coefficients can be interpreted as average derivatives of outcome
- Tractable (easier to interpret)

# REPARAMETERISATION

$$\text{Given } Y_i = \alpha + \beta X_i + e_i$$

	Action	Results
Add constant	Replace $X_i$ by $X_i + A$	$\alpha^* = \alpha - A\beta \quad \beta^* = \beta$
	Replace $Y_i$ by $Y_i + A$	$\alpha^* = \alpha + A \quad \beta^* = \beta$
Multiply regressor by a constant	Replace $X_i$ by $AX_i$	$\alpha^* = \alpha \quad \beta^* = \frac{\beta}{A}$ $SE(\beta^*) = \frac{SE(\beta)}{A}$
	Replace $Y_i$ by $AY_i$	$\alpha^* = A\alpha \quad \beta^* = A\beta$ $SE(\beta^*) = A SE(\beta)$

★ Easy to manipulate  $\beta$ , so  
remember to report unit  
when re-scaling

T-stat unchanged  
★ assuming population scaled by  $\frac{1}{A}$

**NORMALISATION**  $Y_{\text{new}} = \frac{Y_{\text{old}}}{\text{sd}(Y_{\text{old}})}$

Interpretation: 1 unit  $\uparrow X$  associated with a change in  $Y$  by  $\beta$  standard deviations

- 68% of obs fall within 1 s.d. of mean. 95% fall with 2 s.d.

**STANDARDISATION**  $Y_{\text{new}} = \frac{Y_{\text{old}} - E(Y_{\text{old}})}{\text{sd}(Y_{\text{old}})}$

Same interpretation as above.

In this case  $E(Y_{\text{new}}) = 0$

# OMITTED VARIABLE BIAS

**CONFOUNDER** An unobserved variable affecting both treatment and outcome

Once included in regression, becomes a control and eliminates selection bias.

But we can only control for observables.

"Long" (correct) regression:  $Y_i = \alpha + \beta X_i + \gamma D_i + e_i$  omitted variable

Consider auxiliary regression:  $D_i = \pi_0 + \pi_1 X_i + u_i$

Substitute to get "short" (wrong) regression:  $Y_i = (\alpha + \gamma \pi_0) + (\beta + \gamma \pi_1) X_i + (\gamma u_i + e_i)$   
 $= \alpha^s + \beta^s X_i + e_i^s$ , so  $\beta^s = \beta + \gamma \pi_1$  bias!!

$$OVB = \text{Coef in short} - \text{Coef in long} = \beta^s - \beta = \gamma \pi_1$$

Can estimate the sign (direction) of OVB if you can estimate signs of these hypothesis  $\Rightarrow$  Relationship between omitted variable  $A_i$  and var of interest  $Y_i$  ( $\gamma$ )  $\times$  Effect of omitted in long regression ( $\pi_1$ )  
w/o running regression

$\beta_{\text{short}}$	$\beta_{\text{long}}$	OVB	True causal effect of interest $\beta$ is ...
>0	>0	>0	Overstated/Overestimated
>0	>0	<0	Understated/Underestimated
<0	<0	>0	Understated/Underestimated
<0	<0	<0	Overstated/Overestimated
<0	>0	$\equiv 0$	Shouldn't use the above terms.
>0	<0		Just describe in detail, be explicit

Bias causes estimate to switch signs!

★ Just controlling for more variables to remove OVB typically not enough for us to conclude causality

1. There are always more omitted variables out there!
2. Reverse causality and measurement error doesn't get mitigated
3. Usually need either a truly random treatment (lottery) or more sophisticated methods (e.g. IV) for us to be convinced

★ By OVB, only correlated covariates matter in establishing causality

That said, uncorrelated covariates that explain the outcome can be included to  $\downarrow$  se

**OVB INCREASES VARIANCES**

Given  $\begin{cases} Y_i = \alpha + \beta X_{1i} + \beta_2 X_{2i} + e_i \\ Y_i = \alpha^s + \beta^s X_{1i} + e_i^s \end{cases}$ , Assuming  $X_{1i}$  and  $X_{2i}$  are uncorrelated,  
 $\text{Var}(e_i) \leq \text{Var}(e_i^s)$  thus  $R^2_{\text{long}} \geq R^2_{\text{short}}$

## ADDING CONTROLS MAY ↓ OVB

Instead of  $Y_i = \alpha + \beta X_i + \gamma D_i + e_i$  (auxiliary:  $D_i = \pi_0 + \pi_1 X_i + u_i$ )  
we use  $Y_i = \alpha + \beta X_i + \beta_1 A_i + \beta_2 D_i + e_i$  (auxiliary:  $D_i = \pi_0' + \pi_1' X_i + \pi_2' A_i + u_i'$ )

OVB = Effect of  $D_i$  in long regression  $\times$  effect of  $D_i$  on included (i.e.  $X_i$ , or  $A_i$  and  $X_i$ )  
 $= \gamma \pi_1$ , initially, but  $\beta_2 \pi_1'$  after control was added

If  $|OVB|$  falls, this means the control captures part of the original OVB  
(i.e. there is some overlap in what  $D_i$  and  $A_i$  explain)

# HYPOTHESIS TESTING

## TWO-SAMPLE HYPOTHESIS TESTING

- Test if treatment group and control group have different outcomes In this case  $\mu_0 = 0$
- Checking for balance in observable characteristics

Test  $H_0: \mu^1 - \mu^2 = \mu_0$  vs  $H_1: \mu^1 - \mu^2 \neq \mu_0$

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \mu_0}{\hat{SE}} \stackrel{\text{CLT}}{\sim} N(0, 1)$$

e.g. to test  $\beta = 0$  vs  $\beta \neq 0$ , get  $t = \frac{\hat{\beta}}{SE}$ . Reject if  $t > t_{\alpha/2, n}$   
 $t_{0.05, \infty} = 1.645, t_{0.025, \infty} = 1.96, t_{0.01, \infty} = 2.326, t_{0.005, \infty} = 2.576$

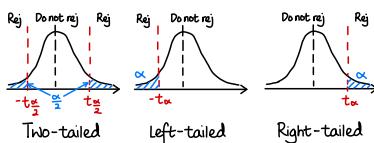
sig. level if  $n$  is large, just " $\infty$ " will do

**PAIRED** Take difference between two values in a pair and use  $\mu_0 = 0$  for a one-sample t-test  
e.g. before and after treatment for each treated individual (initial health status as counterfactual)

**UNPAIRED** The two populations are independent (e.g. treatment and control group in an RCT)

## CRITICAL VALUE APPROACH

A critical value  $t_{\text{crit}}$  ( $-t_{\alpha/2}, t_{\alpha/2}$ ) is set from the sig. level. Reject when t stat is further away from zero than  $t_{\text{crit}}$ .



	$H_0$ True	$H_1$ False
$H_0$ not rejected	✓	Type II error
$H_0$ rejected	Type I error	✓

significance level  
 $= P(\text{Type I error})$   
 power  
 $= 1 - P(\text{Type II error})$

★ Over the years  $\downarrow$  to account for the  $\uparrow$  in by better experimental methods / data collection

## P-VALUE

Probability that when  $H_0$  is true, the absolute value of the sample statistic would be  $\geq$  the actual estimate

- Lower p-value  $\rightarrow$  more confidence in rejecting  $H_0$ .
- Report p-values/t-stats and let readers decide

More informative than the rej/don't rej decision as it makes a continuous parametrisation of rej/don't rej

★ Smallest p-value  $\neq$  largest/most impt (just most precisely measured)

★  $\beta$  is insignificant  $\neq X$  doesn't matter to  $Y$  (maybe  $\uparrow$  data/Var( $X_i$ ); Causation may not imply correlation)

★  $\beta$  is significant  $\neq X$  matters to  $Y$  (with  $\uparrow$  n, p-value can go as small as you like)

★ Do not delete vars with insig coeffs and re-run. Choice of variables determined only by intuition/economic theory.

**CONFIDENCE INTERVAL** Set of all values of  $\mu_0$  that cannot be rejected in a t-test at the  $\alpha\%$

$$[(\bar{Y}_n^1 - \bar{Y}_n^2) - t_{\frac{\alpha}{2}} \widehat{SE}, (\bar{Y}_n^1 - \bar{Y}_n^2) + t_{\frac{\alpha}{2}} \widehat{SE}]$$

**COMPARING COEFFICIENTS** Given  $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$ ,  $H_0: \beta_1 = \beta_2$  vs  $H_1: \beta_1 \neq \beta_2$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)}}$$

**TRANSFORMING THE REGRESSION** Let  $\theta = \beta_1 - \beta_2$ .

Then  $Y_i = \alpha + \theta X_{1i} + \beta_2(X_{1i} + X_{2i}) + e_i$ . Rerun regression to get  $t = \frac{\hat{\theta}}{se(\hat{\theta})}$

**JOINT HYPOTHESIS**  $H_0: \beta_1 = \beta_2 = 0$  vs  $H_1: (H_0)^c$

$$F = \frac{1}{2} \left( \frac{t_1^2 + t_2^2 - 2\rho_{t_1 t_2} t_1 t_2}{1 - \rho_{t_1 t_2}^2} \right) \sim \chi^2_2$$

NOT t-test! We cannot just combine the t-stats because

- ① We'll reject too often
- ②  $\beta_1$  and  $\beta_2$  may often be correlated

# REGRESSION ANOVA

THEOREM

$$\begin{aligned} \text{Var}(Y_i) &= \text{Var}(\hat{Y}_i) + \text{Var}(e_i) + \text{Cov}(\hat{Y}_i, e_i) = \text{Cov}(X_i, e_i) = 0 \\ &= \text{Var}(\alpha + \beta X_i) + \text{Var}(e_i) \\ &= \beta^2 \text{Var}(X_i) + \text{Var}(e_i) \end{aligned}$$

REGRESSION R<sup>2</sup>

$$R^2 = \frac{\text{Var}(\hat{Y}_i)}{\text{Var}(Y_i)} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(Y_i)} = \frac{\beta^2 \text{Var}(X_i)}{\text{Var}(Y_i)} = \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}} \text{ (in Stata)}$$

↑ sum of squares  
unexplained variation

= part of the variation of  $Y_i$  explained by model (because of  $\beta$ )  
not covariate!

- Higher  $R^2 \neq$  better fit of data.
  - $R^2$  can be increased by just including more variables (unless  $\hat{\beta}_{\text{new regressor}} = 0$ )
  - $R^2$  can be arbitrarily close to 1 when model is totally wrong
- $R^2$  reflects the combined explanatory power of all variables used, not sum of their individual explanatory power
  - e.g. if you get  $R_x^2$  from  $\hat{Y}_i = \hat{\beta}_1 X_i$  and  $R_s^2$  from  $\hat{Y}_i = \hat{\beta}_1 S_i$ , you don't get  $R_s^2 + R_x^2$  from  $\hat{Y} = \hat{\beta}_1 X_i + \hat{\beta}_2 S_i$

CORR COEFF  $\rho$

$$\rho = \frac{\text{Cov}(Y_i, X_i)}{\sqrt{\text{Var}(Y_i) \text{Var}(X_i)}}$$

- In bivariate regression,  $\rho^2 = R^2$
- Can approximate how much one variable explains the other by looking at their correlation
- Is symmetrical ( $\rho_{yx} = \rho_{xy}$ )

# HOMOSCEDASTICITY

## SAMPLING SD OF OLS SLOPE

$$se(\hat{\beta}) = \sqrt{\frac{1}{n} \frac{Var(e_i)}{Var(X_i)}}$$

↓ Variation in  $e_i$ , ↑ estimate precision  
 ↓ Variation in the regressor, ↑ estimate precision  
 ↓ Sample size, ↑ estimate precision

## ESTIMATED SD OF OLS SLOPE

$$\widehat{se}(\hat{\beta}) = \sqrt{\frac{1}{n} \frac{\widehat{Var}(e_i)}{\widehat{Var}(X_i)}}$$

Obs far from the mean in the  $x$ -direction are particularly informative (check for outliers)

## HETEROSCEDASTICITY

Dispersion (variance) in residuals **correlated** with regressor. Use **RSE**

→ Not necessarily independent

## HOMOSCEDASTICITY

Dispersion (variance) in residuals **uncorrelated** with regressor,

or  $E(e_i^2 | X_i) = Var(e_i) = \text{a constant}$ . Or  $Var[(X_i - E(X_i)) e_i] = Var(e_i) Var(X_i)$

## ROBUST SD OF OLS SLOPE

$$RSE(\hat{\beta}) = \sqrt{\frac{1}{n} \frac{Var[(X_i - E(X_i)) e_i]}{(Var(X_i))^2}}$$

\* Under homoscedasticity,  $RSE(\hat{\beta}) = se(\hat{\beta})$

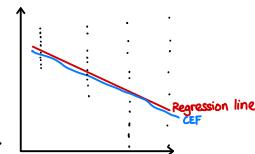
\* In practice, use RSE by default.

\* Even drawn from the same pop<sup>n</sup>, RSE is lower when sample homoscedastic

\* We can have a homoscedastic CEF but heteroscedastic data →

Even if CEF residual is homoscedastic, regression residual (sum of CEF residuals) may not be.

$$\hookrightarrow \varepsilon_i \equiv Y_i - E(Y_i | X_i)$$



# REGRESSION ANATOMY FORMULA

If we have a multivariate regression  $Y_i = \alpha + \sum_{k=1}^n \beta_k X_{ki} + e_i$   
Then  $\beta_i = \frac{\text{Cov}(Y_i, \tilde{X}_{ii})}{\text{Var}(\tilde{X}_{ii})} = \frac{\text{Cov}(Y_i, \tilde{X}_{ii})}{\text{Var}(\tilde{X}_{ii})}$  and  $X_{ii} = \pi_1 + \sum_{k=2}^n \pi_k X_{ki} + \tilde{X}_{ii}$

## PROOF

$$\begin{aligned}\frac{\text{Cov}(Y_i, \tilde{X}_{ii})}{\text{Var}(\tilde{X}_{ii})} &= \frac{\text{Cov}(\alpha + \beta_i X_{ii} + \sum_{k=2}^n \beta_k X_{ki} + e_i, \tilde{X}_{ii})}{\text{Var}(\tilde{X}_{ii})} \\ &= \frac{\text{Var}(\tilde{X}_{ii})}{\text{Var}(\tilde{X}_{ii})} \underbrace{\beta_i}_{= \beta_i \text{Cov}(X_{ii}, \tilde{X}_{ii}) + \sum_{k=2}^n \beta_k \text{Cov}(X_{ki}, \tilde{X}_{ii}) + \text{Cov}(e_i, \tilde{X}_{ii})} \underbrace{+ \sum_{k=2}^n \beta_k \text{Cov}(X_{ki}, \tilde{X}_{ii}) + \text{Cov}(e_i, \tilde{X}_{ii})}_{=0 \text{ since } \tilde{X}_{ii} \text{ is a function of } X_{ii} \text{ and } X_{ii}} \\ &= \beta_i \text{Cov}(X_{ii}, \tilde{X}_{ii}) + \sum_{k=2}^n \beta_k \text{Cov}(X_{ki}, \tilde{X}_{ii}) + \text{Cov}(e_i, \tilde{X}_{ii}) \underbrace{= 0}_{\text{while } e_i \text{ is a residual from a regression with } X_{ii} \text{ and } X_{ii}}\end{aligned}$$

Instead of a multivariate regression, we run smaller bivariate regressions

## ADDING CONTROLS HAS UNCERTAIN EFFECT ON SE

$$\text{s.e.}(\hat{\beta}_i) = \sqrt{\frac{\text{Var}(e_i)}{n \text{Var}(\tilde{X}_{ii})}} \quad \text{while} \quad \text{s.e.}(\hat{\beta}_i^s) = \sqrt{\frac{\text{Var}(e_i^s)}{n \text{Var}(X_{ii})}}$$

Since  $\text{Var}(e_i^s) \geq \text{Var}(e_i)$  and  $\text{Var}(X_{ii}) \geq \text{Var}(\tilde{X}_{ii})$ ,  $\text{se}(\hat{\beta}_i^s) \leq \text{se}(\hat{\beta}_i)$

- Don't know if an estimate becomes more precise if we add more regressors
- If new regressor is uncorrelated with  $X_{ii}$  ( $\text{Var}(X_{ii}) = \text{Var}(\tilde{X}_{ii})$ ) but a strong predictor of  $Y_i$  ( $\text{Var}(e_i) \ll \text{Var}(e_i^s)$ ), it will  $\downarrow$  se

# NON-LINEAR FUNCTIONAL FORMS

**QUADRATIC**  $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + e_i$ , treating  $X_i^2$  as an additional explanatory variable  
Minimising  $e_i^2$ , the effect of a unit change in  $X_i$  is  $\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2\beta_2 X_i$

**LEVEL-LOG / LINEAR-LOG**  $Y_i = \alpha + \beta \ln X_i + e_i$   
 $\Rightarrow \frac{\partial Y_i}{\partial X_i} = \frac{\beta}{X_i} \Rightarrow \hat{\beta} = \frac{\frac{\partial Y_i}{\partial X_i}}{\left(\frac{\partial X_i}{X_i}\right)} \frac{\text{abs } \Delta \text{ in } Y_i}{\% \Delta \text{ in } X_i}$

Interpretation: Change in  $Y_i = \hat{\beta}$  (% change in  $X_i$ )

**LOG-LEVEL / LOG-LINEAR**  $\ln Y_i = \alpha + \beta X_i + e_i$   
 $\Rightarrow \beta = \frac{\partial \ln(Y_i)}{\partial X_i} = \frac{\left(\frac{\partial Y_i}{Y_i}\right)}{\left(\frac{\partial X_i}{X_i}\right)} \frac{\% \Delta \text{ in } Y_i}{\text{abs } \Delta \text{ in } X_i}$

Interpretation: % Change in  $Y_i = \hat{\beta}$  (change in  $X_i$ )

\* When  $Y_i$  is large (e.g. GDP) or  $\text{Var}(Y_i)$  is large

**LOG-LOG**  $\ln Y_i = \alpha + \beta \ln X_i + e_i$   
 $\Rightarrow \beta = \frac{\partial \ln(Y_i)}{\partial \ln(X_i)} = \frac{\left(\frac{\partial Y_i}{Y_i}\right)}{\left(\frac{\partial X_i}{X_i}\right)} \frac{\% \Delta \text{ in } Y_i}{\% \Delta \text{ in } X_i}$

\* Can be interpreted as an elasticity

**POLYLOG?!**  $Y_i = \alpha + \beta_1 \ln X_i + \beta_2 (\ln X_i)^2 + e_i$   
★ Even if a more complicated model fits the data better,  
it may lose clarity / be harder to interpret

- Approximative % change for logs not good for large impacts; For  $\beta > 0.2$ , use  $e^\beta - 1$

# DUMMY VARIABLE

or indicator/categorical variable

- Can break continuous variable  $X_i$  into

$$D_i = \begin{cases} 1 & \text{if } X_i < 20 \\ 0 & \text{if } X_i \geq 20 \end{cases}$$

$$D_{1i} = \begin{cases} 1 & \text{if } X_i < 20 \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if } X_i \geq 20 \\ 0 & \text{otherwise} \end{cases}$$

**CONSTANT ONLY**

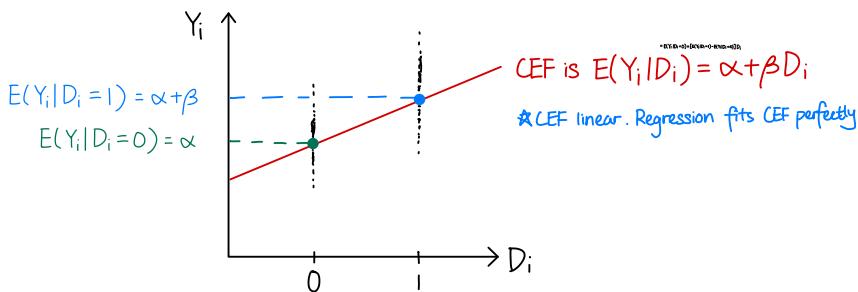
$$Y_i = \alpha + e_i$$

Then  $E(Y_i) = \alpha$

**I DUMMY**  $Y_i = \alpha + \beta D_i + e_i$

mean in subgroup with  $D_i=0$  ( $X_i \geq 20$ )

difference in means of  $Y_i$  for the two subgroups



**SPLITTING I INTO 2 DUMMIES**  $Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + e_i$ , with  $D_{1i} = D_i$

mean in subgroup with  $D_{1i}=1$  ( $D_i=1$ ) ( $X_i < 20$ )

mean in subgroup with  $D_{1i}=0$  ( $D_i=0$ ) ( $X_i \geq 20$ )

Since  $D_{1i} + D_{2i} = 1$ ,  $Y_i = \alpha + \beta D_i + e_i = (\alpha + \beta) D_{1i} + \alpha D_{2i} + e_i = \beta_1 D_{1i} + \beta_2 D_{2i} + e_i$

So  $\alpha + \beta = \beta_1$ ,  $\alpha = \beta_2$ ,  $\beta = \beta_1 - \beta_2$

**DUMMY VARIABLE TRAP**

If we have K MECE dummies, we:

- Can include all K dummy variables and no constant ★ if you care about means of each category
- Can include constant and any K-1 of the dummies ★ if you care about difference in means of each category relative to omitted category
- Cannot include constant and all K dummy variables - perfectly collinear

## SATURATION

**Saturated**: As many <sup>dummies</sup> parameters as there are unique combinations of <sup>possible categories</sup> covariates

- e.g.  $Y_i = \alpha + \beta D_i + e_i$ ,  $Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + e_i$   
MEEC
- Completely unrestricted - regression fits group means defined by dummies perfectly. Fits CEF

**Not saturated**: No. of categories/subgroups  $\neq$  no. of parameters

- e.g.  $Y_i = \alpha + \beta_1 D_i + \beta_2 G_i + e_i$ ,  $D_i$  and  $G_i$  not MEEC ( $3$  parameters,  $2^2 = 4$  subgroups)
- Imposes a restriction ( $D_i + G_i \neq 1$ )

# INTERACTIONS

\* This is saturated (4 parameters, 4 categories)

Expected value of  $Y_i$  given

	$D_i = 0$	$D_i = 1$
$G_i = 0$	$\alpha$	$\alpha + \beta_1$
$G_i = 1$	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + \beta_3$

$E(Y_i|D_i=1, G_i=0) - E(Y_i|D_i=0, G_i=0) = \beta_1$  ← Effect of change in  $D_i$  on  $Y_i$  for  $G_i=0$

$E(Y_i | D_i=1, G_i=1) - E(Y_i | D_i=0, G_i=1) = \beta_1 + \beta_3$  ← Differential effect of change in  $D_i$  on  $Y_i$  for  $G_i=1$

Allows for different treatment effect/slopes for different groups (**heterogeneity**)

- Have to account for heterogeneity to avoid OVB

## HYPOTHESIS TESTING

$$(A) \quad Y_i = \alpha + \beta_1 D_i + \beta_2 G_i + \beta_3 (D_i \times G_i) + e_i$$

Class size effect for  $C_{1i} = 0$  group is  $\beta_1$

Class size effect for  $G_i = 1$  group is  $\beta_1 + \beta_3$

\*  $\beta > 0 \Rightarrow$  effect stronger for  $G_i = 1$

To test whether class size effect is same in  $G_i=0$  and  $G_i=1$  groups:  $H_0: \beta_3 = 0$

$$(B) Y_i = \alpha + \beta_1 (D_i \times (1 - G_i)) + \beta_2 G_i + \beta_3 (D_i \times G_i) + e_i$$

Class size effect for  $G_i = 0$  group is  $\beta_1$

Class size effect for  $G_i = 1$  group is  $\beta_3$

To test whether class size effect is same in  $G_i=0$  and  $G_i=1$  groups:  $H_0: \beta_1 = \beta_3$

# DATA

## Sources of Variation

### EXPERIMENTAL

When we ourselves generate the variations in  $X_i$  (e.g. through random allocation) to determine  $X$ 's causal effect on  $Y_i$

### OBSERVATIONAL

We observe the sample, but researcher has no direct influence.  
So we don't know why  $X$  varies — could be due to omitted variables, reverse causation or even measurement error. May be non-random

★ Natural/quasi-experiments used to identify causal effects

## Types

### CROSS-SECTIONAL

- Many subjects
- One point in time / without regard for time

### TIME-SERIES

- Only one subject, but tracked over time (e.g. yearly UK CPI)
- Data typically gathered over time periods at equal length (quarterly, daily)

### PANEL/LONGITUDINAL

- Many subjects, each tracked over multiple periods
- Different from repeated/pooled cross section (a new sample drawn independently every time)

Thus, a dataset with "10 observations" might actually be .

- 10 subjects  $\times$  1 period cross section
- 1 subject  $\times$  10 periods time series
- 2 subjects  $\times$  5 periods panel
- 5 subjects  $\times$  2 periods panel
- 3 subjects  $\times$  2 periods unbalanced panel (with attrition)  
+ 4 subjects  $\times$  1 period oh no! But it's not tested, haha

## Random Sampling $\neq$ Random Assignment

### **RANDOM SAMPLING**

Generates sample representative of population characteristics

- Sample characteristics  $\{x_i\}$  have the same distribution as population characteristics  $\{X_i\}$
- Able to say something about distributions and correlations in whole pop<sup>n</sup>  
i.e. your results have external validity
- Able to use statistical inference to infer pop<sup>n</sup> parameters from sample characteristics
- Cannot remove selection bias in pop<sup>n</sup> (internal validity concerns)

### **RANDOM ASSIGNMENT**

Actually assigns treatment/characteristics randomly in sample

- Able to say something about causal effects of that treatment on that set of data  
i.e. your results have internal validity
- But sample can still be non-representative (external validity concerns)

# MEASUREMENT ERROR

## CLASSICAL MEASUREMENT ERROR

$Y_i = \alpha + \beta X_i^* + e_i$  measurement error of  $X_i^*$   
 We don't observe  $X_i^*$  but  $X_i = X_i^* + w_i$   
 $\text{Cov}(X_i^*, w_i) = 0$  and  $\text{Cov}(e_i, w_i) = 0$   $\star$   $w_i$  has nothing to do with outcome

$$\beta = \frac{\text{Cov}(Y_i, X_i^*)}{\text{Var}(X_i^*)}$$
 but the estimator based  
 on mismeasured data is  $\hat{\beta} = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)} = \beta \frac{\text{Var}(X_i^*)}{\text{Var}(X_i^*) + \text{Var}(w_i)}$

## ATTENUATION BIAS

Let attenuation error be  $\lambda = \frac{\text{Var}(X_i^*)}{\text{Var}(X_i^*) + \text{Var}(w_i)}$ .  
 "signal" "noise"

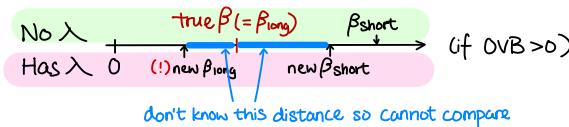
Since  $0 < \lambda \leq 1$ ,  $\hat{\beta}$  is biased towards 0 by  $\lambda$  (underestimate)

$$\text{Multivariate: } \tilde{\lambda} = \beta \frac{\text{Var}(\tilde{X}_{ii}^*)}{\text{Var}(\tilde{X}_{ii}^*) + \text{Var}(w_i)}$$

Since  $\text{Var}(\tilde{X}_{ii}^*) < \text{Var}(X_{ii}^*)$  <sup>ANOVA</sup>, adding correlated regressors worsens attenuation error  
 $\star$  but it also  $\downarrow$  OVB!

## RELATIONSHIP WITH OVB

With attenuation bias, sign of OVB doesn't change  
 but now we don't know if  $\beta_{short}$  or  $\beta_{long}$  is closer to the true effect,  $\beta$



- $\star$  Attenuation bias can cause good controls to appear insignificant
- $\star$  Implies you cannot check for OVB simply by regressing and checking sig

## CLASSICAL MEASUREMENT ERROR - DEPENDENT VAR

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

We don't observe  $Y_i^*$  but  $Y_i = Y_i^* + w_i$  measurement error of  $Y^*$

$$\text{Cov}(Y_i^*, w_i) = 0 \text{ and } \text{Cov}(X_i, w_i) = 0 \quad \star w_i \text{ has nothing to do with outcome}$$

Thus, the regression becomes

$$Y_i = \alpha + \beta X_i + \underbrace{w_i}_{w_i \text{ becomes a random part of residual}} + e_i$$

↑ precision ↑ se

but does not lead to bias

## AVERAGING - NONCLASSICAL MEASUREMENT ERROR

We have observations  $X_{ij}$  for  $i = 1, 2, \dots$  but instead of  $Y_i = \alpha + \beta X_{ij} + e_{ij}$   
we take the averages  $\bar{X}_j = \sum_{i=1}^n X_{ij}$  and run  $Y_i = \alpha + \beta \bar{X}_j + e_{ij}$

- We can regress  $X_{ij}$  on  $\bar{X}_j$ :  $X_{ij} = \bar{X}_j + w_{ij}$

- $\text{Cov}(\bar{X}_j, w_{ij}) = 0$

- $\text{Cov}(X_{ij}, w_{ij}) = \text{Var}(w_{ij})$

- $\text{Cov}(X_{ij}, \bar{X}_j) = \text{Var}(\bar{X}_j)$

- No attenuation bias as  $\hat{\beta} = \beta$

- Same coefficient and correct se ✓ can prove but not required  
any noise due to averaging 'cancels out'

- se for group averages will be larger as we have not used all variations in  $X_{ij}$

## OTHER NONCLASSICAL MEASUREMENT ERRORS

- In dependent variables

- Systematically over/underreport

- Definition: Error correlated with true values (esp. self-reported)

# INTERNAL VALIDITY

Can we interpret the estimate (for this particular setting and question) as a **causal effect**?

① Estimator for the causal effect must be **unbiased** and **consistent**

② Statistical inference valid

- Hypothesis tests should have desired sig level & power
- Valid se (e.g. by accounting for heteroscedasticity)

## RESEARCH DESIGN

- Randomised experiment (experimental data)
  - Check for balance — if ✓, no need to worry about controls
- Non-experimental data
  - Have we controlled for all confounders (OVB)
    - Avoid bad controls
    - Use uncorrelated controls to use ↑ precision
  - Specified the right functional form? — Plot data and check
  - Mismeasured data? — IV may fix attenuation bias in classical case
  - Missing data/ Sample selection?
    - If missing at random — No bias, but less efficient estimators due to ↓n
    - If missing randomly based on the value of a regressor (e.g. only count obs with  $X_i < x$ )
      - No bias, but less efficient estimators due to ↓n
    - If missing randomly based on the value of dep var (e.g. only count obs with  $Y_i < y$ )
      - Sample selection bias (whether data is observed depends on outcome)
  - Simultaneity bias (reverse causality bias)? — IV?

# EXTERNAL VALIDITY

Is an estimate we obtained in this setting **valid in other settings**?

Differences between pop<sup>n</sup> studied and pop<sup>1</sup> of interest (heterogeneity)/in timing/setup

# PREDICTION

- Threats to internal validity negligible if just want prediction (no causal int<sup>n</sup> needed)
- Prediction in Machine Learning called a 'black box' (can predict but don't know why)

## WHAT MAKES A GOOD PREDICTION

Certainty Stable relationship in data → Consistently predict result

Accuracy High R<sup>2</sup> (Good fit of functional form) ★ Does not tell us about causality

- Avoid overfitting - can't predict anything if exactly fit the data

## STATA TABLE

. reg lifesatisf child, robust

Linear regression

Outcome		$\beta$	se	t-stat for testing $\beta=0$	$t = \frac{\beta}{se}$	P-value for $\beta=0$	[95% Conf. Interval]
Regressors	child	.0751444	.0396867	1.89	0.059	-.0027574	.1530461
Constant	cons	.5450237	.0343242	15.88	0.000	.4776481	.6123993

p-value for f-test on all coefficients  
 $\beta_1 = \dots = \beta_k = 0$  (i.e. "nothing is going on here")  
equals p-value for t-test on  $\beta=0$  if just one regressor

Number of obs	=	806
F( 1, 804)	=	3.59
Prob > F	=	0.0587
R-squared	=	0.0045
Root MSE	=	.48929

# INSTRUMENTAL VARIABLES

**PURPOSE** Dealing with reverse causation, OVB, or non-random measurement error

Instrument eliminates bias by acting as a randomiser for the instrumented variable.

Instead of having the treatment randomly assigned, instrument randomly generates variations in the treatment variable that we can use to estimate the causal effect

## FIRST STAGE/RELEVANCE ASSUMPTION

Instrument has a (strong) causal effect on the regressor of interest.  $\text{Cov}(Z_i, D_i) \neq 0$

**INDEPENDENCE ASSUMPTION** Instrument randomly assigned or as good as randomly assigned

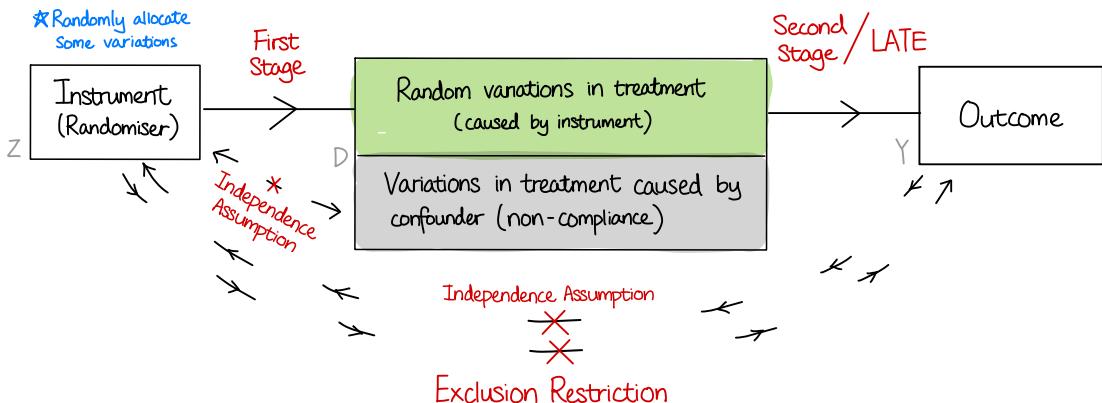
Unrelated to omitted variables  $\text{Cov}(Z_i, e_i) = \text{Cov}(Z_i, X_i) = 0$ .

There should be no variable affecting both Z and Y, and Y should not affect Z.

**EXCLUSION RESTRICTION** The instrument must affect the outcome only through a single channel — the treatment (the var it is instrumenting).

This means if we regress  $Y_i = \alpha + \beta D_i + \gamma X_i + \theta Z_i + e_i$ , we must have  $\theta = 0$

- $Z_i$  cannot directly explain  $Y_i$  once  $D_i$  is in equation
- Mustn't have weak first stage but strong reduced form ( $\Rightarrow$  another causal channel exists)
- Assumption can be empirically rejected/supported but never proven (We don't observe  $X_i$ )



## INTENTION TO TREAT

When treatment assigned differs from treatment delivered (*imperfect compliance*)  
 ITT analysis (the *reduced form*) captures causal effect of being *assigned treatment*.

Both causal effects since  $Z_i$  is randomised

$$\begin{array}{ccc} \text{Reduced form/ IOT effect} & \text{First stage} & \hat{\beta} \text{ Second stage} \hat{=} \\ \uparrow & \uparrow & \\ \text{Effect of } Z \text{ on } Y = \text{Effect of } Z \text{ on } D \times \text{Effect of } D \text{ on } Y & & \end{array}$$

## WALD (IV) ESTIMATOR

**LOCAL AVERAGE TREATMENT EFFECT (LATE)** Can be found using the **IV (Wald) estimator**

$$\begin{aligned} \text{Effect of } D \text{ on } Y, \beta_{IV} &= \frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D} = \frac{E(Y_i|Z_i=1) - E(Y_i|Z_i=0)}{E(D_i|Z_i=1) - E(D_i|Z_i=0)} \\ &\xrightarrow{\text{Rescaling ITT effect by compliance}} = \frac{\delta}{\phi} \quad \text{where} \quad \begin{aligned} Y_i &= \alpha_0 + \delta Z_i + e_{0i} && (\text{reduced form}) \\ D_i &= \alpha_1 + \phi Z_i + e_{1i} && (\text{first stage}) \end{aligned} \end{aligned}$$

## TWO-STAGE LEAST SQUARES

Allows for *multiple instruments* and *controls*

Instead of Wald, regress the first stage to get fitted values  $\hat{D}_i = \alpha_1 + \phi Z_i$   
 then run second stage with fitted values  $Y_i = \alpha_2 + \beta_{2SLS} \hat{D}_i + e_{0i}$  to get  $\beta_{2SLS}$

- Works because  $Y_i = \alpha_2 + \beta_{2SLS} \hat{D}_i + e_{0i} = \alpha_2 + \beta_{2SLS} (\alpha_1 + \phi Z_i) + e_{0i}$   
 $= \alpha_2 \beta_{2SLS} + \alpha_2 + (\beta_{2SLS} \phi) Z_i + e_{0i}$   
 $\text{Compare w reduced form} \quad = \alpha_0 + \delta Z_i + e_{0i} \quad \text{so} \quad \beta_{2SLS} = \frac{\delta}{\phi}$
- But manual calculation gives wrong se!

$$\beta_{2SLS} = \frac{\delta}{\phi} = \frac{\text{Cov}(Y_i, Z_i)/\text{Var}(Z_i)}{\text{Cov}(D_i, Z_i)/\text{Var}(Z_i)} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} \quad \text{vs OLS} \quad \beta = \frac{\text{Cov}(Y_i, D_i)}{\text{Var}(D_i)}$$

We see OLS is just special-case IV with treatment as instrument ( $D_i = Z_i$ )

## IV SOLVES OVB

treatment all unobserved confounders

compare

$$\text{OLS: } Y_i = \alpha + \beta D_i + \gamma X_i + e_i$$

$$\text{IV: } Y_i = \alpha + \beta_{IV} D_i + \eta_i$$

$\eta_i$  is a structural error term, not residual  
 $\eta_i = \gamma X_i + e_i$  since  $\beta_{IV} = \beta$

Then  $\beta_{IV} = \beta$

## IV SOLVES MEASUREMENT ERROR

If we have  $Y_i = \alpha + \beta X_i^* + e_i$ ,  $X_i = X_i^* + w_i$

$$\beta_{IV} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(X_i, Z_i)} = \frac{\text{Cov}(\alpha + \beta X_i^* + e_i, Z_i)}{\text{Cov}(X_i, Z_i)} = \frac{\beta \text{Cov}(X_i^*, Z_i) + \text{Cov}(e_i, Z_i)}{\text{Cov}(X_i^*, Z_i) + \text{Cov}(w_i, Z_i)} = \beta$$

$\stackrel{=0}{\rightarrow}$

either bc  $Z_i$  is specifically to solve  $w_i$  problem  
 or, even if  $Z_i$  is just for OVB, random assignment often implies  $\text{Cov}=0$

## ADD COVARIATES

To all equations, including second stage

Add  $A_i$  to

$\left\{ \begin{array}{l} \text{reduced form} \\ \text{first stage} \\ \text{fitted first stage} \end{array} \right.$	$\begin{array}{l} Y_i = \alpha_0 + \delta Z_i + \gamma_0 A_i + e_{0i} \\ D_i = \alpha_1 + \phi Z_i + \gamma_1 A_i + e_{1i} \\ \hat{D}_i = \alpha_1 + \phi Z_i + \gamma_1 A_i \end{array}$	$\Rightarrow$	$2\text{SLS second stage}$
			$Y_i = \alpha_0 + \beta_{2\text{SLS}} \hat{D}_i + \gamma_0 A_i + e_{0i}$

Adding covariates can control for confounders affecting both  $Z_i$  and  $Y_i$  to make  $Z_i$  as good as randomly assigned (intuition: only using good variation unrelated with  $X_{li}$  as instrument). Still Solves OVB.

## ADD INSTRUMENTS

Add  $W_i$  to

$\left\{ \begin{array}{l} \text{reduced form} \\ \text{first stage} \\ \text{fitted first stage} \end{array} \right.$	$\begin{array}{l} Y_i = \alpha_0 + \delta_s Z_i + \delta_s W_i + e_{0i} \\ D_i = \alpha_1 + \phi_s Z_i + \phi_s W_i + e_{1i} \\ \hat{D}_i = \alpha_1 + \phi_s Z_i + \phi_s W_i \end{array}$	$\Rightarrow$	$2\text{SLS second stage}$
			$Y_i = \alpha_0 + \beta_{2\text{SLS}} \hat{D}_i + e_{0i}$

Command: ivregress 2sls outcome controls (treatment= instruments), robust

More (strong) instruments is better (in E(220))

**2SLS SE** Second stage:  $Y_i = \alpha + \beta_{2\text{SLS}} \hat{D}_i + e_{2i}$ ,  $\text{Cov}(e_{2i}, \hat{D}_i) = 0$

Second stage that we're trying to estimate:  $Y_i = \alpha + \beta_{IV} D_i + \eta_i$

$\text{Cov}(\eta_i, D_i) \neq 0$  since we haven't estimated this by OLS

$\text{se}(\hat{\beta}_{IV}) = \sqrt{\frac{\text{Var}(\eta_i)}{n \text{Var}(\hat{D}_i)}}$ . Since  $\text{Var}(\hat{D}_i) \ll \text{Var}(D_i)$ , IV can be very imprecise  
 usually ( $Z_i$  predicts only part of  $D_i$ 's variance)

\* Manual 2SLS doesn't give right  $\text{se}(\hat{\beta}_{IV})$ ! Use Stata

# IV AND SIMULTANEITY

e.g. ss/dd

$$\begin{cases} Y_i = \alpha + \beta X_i + \eta_{1i} \\ X_i = \gamma + \delta Y_i + \eta_{2i} \end{cases}$$

## PROBLEM

Attempt to do OLS on equation (1) :

$$Y_i = \alpha + \beta X_i + \eta_{1i}$$

This clearly implies  $\text{Cov}(Y_i, \eta_{1i}) \neq 0$

$$= \alpha + \beta(\gamma + \delta Y_i + \eta_{2i}) + \eta_{1i}$$

$\uparrow$  OLS fails since  $\text{Cov}(X_i, \eta_{1i}) = \dots \text{Cov}(Y_i, \eta_{1i}) \dots \neq 0!$

Can solve with an instrument  $Z_i$

- As good as random ( $X_i \not\rightarrow Z_i$ )
- Affects  $X_i$  directly, but not  $Y_i$  OR Affects  $Y_i$  directly, but not  $X_i$



Supply equation

$$\ln(Q_i) = \alpha_S + \beta_S \ln P_i + \gamma_S Z_{is} + \eta_{1s}$$

Instrument acts as a "supply shifter" and keeps dd unchanged

Demand equation

$$\ln(Q_i) = \alpha_D + \beta_D \ln P_i + \eta_{1D}$$

$Z$  must not appear here

$\beta_S, \beta_D$  are ss/dd elasticities

If reverse causation makes it a self-reinforcing cycle (i.e.  $X_i \xrightarrow{+} Y_i, X_i \xleftarrow{-} Y_i$ )  
bias will magnify true effect (typical for PED)

If  $X_i \xleftarrow{+} Y_i$ , bias will oppose true effect

$$\begin{aligned} \beta_{D, IV} &= \frac{\text{Cov}(\ln Q_i, Z_{is})}{\text{Cov}(\ln P_i, Z_{is})} = \frac{\text{Cov}(\alpha_D + \beta_D \ln P_i + \eta_{1D}, Z_{is})}{\text{Cov}(\ln P_i, Z_{is})} \\ &= \frac{\beta_D \text{Cov}(\ln P_i, Z_{is}) + \text{Cov}(\eta_{1D}, Z_{is})}{\text{Cov}(\ln P_i, Z_{is})} = 0 \quad \text{Exogeneity of instrument} \\ &\quad \text{i.e. ss shifter uncorrelated w/ dd conditions} \\ &= \beta_D \text{ (PED)} \end{aligned}$$

Thus, the supply shifter  $Z_i$  instruments for PED,  $\beta_D$

## LINCO

- SS/dd equations are structural equations that rep. structure of the sim. eq's model
- They cannot be regression eq's since  $\eta_{1s}, \eta_{1D}$  are correlated with price
- $\ln P_i, \ln Q_i$  are endogenous variables determined by model
- $Z_{is}$  is an exogenous variable not determined by model (controls, too)

# IDENTIFICATION

- There is a reduced form for each endogenous variable
- Endo var on LHS; constants, exo var and error terms on RHS
- Can be estimated by OLS (exo var assumed uncorrelated with error terms)
- We can run OLS on  $\ln Q_i$  and  $\ln P_i$ 's reduced forms, directly manipulate reduced form coeffs to get  $\beta_D$
- A sim eq's model is identified when we can work back from reduced form coeffs to the structural parameters

e.g. take

$$\left\{ \begin{array}{ll} \text{Supply equation} & \ln(Q_i) = \alpha_S + \beta_S \ln P_i + \gamma_S Z_{is} + \eta_{is} \\ \text{Demand equation} & \ln(Q_i) = \alpha_D + \beta_D \ln P_i + \eta_{id} \end{array} \right.$$

Reduced forms are

$$\left\{ \begin{array}{l} \ln Q_i = \frac{\alpha_S \beta_D - \alpha_D \beta_S}{\beta_D - \beta_S} + \underbrace{\frac{\gamma_S \beta_D}{\beta_D - \beta_S}}_{\pi_{21}} Z_{is} + \frac{\eta_{is} \beta_D - \eta_{id} \beta_S}{\beta_D - \beta_S} \\ \ln P_i = \frac{\alpha_S - \alpha_D}{\beta_D - \beta_S} + \underbrace{\frac{\gamma_S}{\beta_D - \beta_S}}_{\pi_{11}} Z_{is} + \frac{\eta_{is} - \eta_{id}}{\beta_D - \beta_S} \end{array} \right.$$

can do OLS!

$$\text{Then PED, } \beta_D = \frac{\pi_{21}}{\pi_{11}}$$

If we have a supply shifter as instrument, can do the same to get  $\beta_S$

Just-identified: 1 instrument to each endo var

Not identified: no instrument for it (e.g. supply function since  $\nexists$  dd-shifters in e.g.)

Over-identified:  $>1$  instrument for it

# EXTRA · IV MONOTONICITY

		Lottery losers $Z=0$	
		Doesn't attend $D_i(0)=0$	Attends $D_i(0)=1$
Lottery winners $Z=1$	Doesn't attend $D_i(1)=0$	Never-takers	Defiers
	Attends $D_i(1)=1$	Compliers	Always-takers

- Instrument affects treatment status
- Instrument unrelated to treatment

## MONOTONICITY

The assumption that the instrument affects the variable in one direction only

- Monotonicity assumes no defiers, otherwise a serious problem for IV

IV only estimates treatment effect for ppl whose treatment status can be manipulated by instrument

- Randomise compliers to treatment control groups, and these work as counterfactuals

- % of lottery losers who enrolled =  $\Pr[D_i(0)=1]$  = % of always-takers
- % of lottery winners who did not enroll =  $\Pr[D_i(1)=0]$  = % of never-takers
- % of lottery winners who enrolled =  $\Pr[D_i(1)=1]$  = % of always-takers + % of compliers
- % of lottery losers who did not enroll =  $\Pr[D_i(0)=0]$  = % of never-takers + % of compliers

Randomisation means % of always-takers in ① and ③ are the same  
% of never-takers in ② and ④ are the same

③ - ① = % of compliers = effect of offers on attendance (denominator of LATE!!)

## INTENTION TO TREAT

ITT analysis captures the causal effect of being offered treatment

- Both always-takers and compliers will be analysed
- ITT is LATE if there are no always-takers  $E(Y_{1i} - Y_{0i} | Z_i=1) = E(Y_{1i} - Y_{0i} | D_i=1)$
- Policymakers often more interested on ITT