ECON 594: Applied Economics

Measurement Error and Other Empirical Concerns

Sam Norris
University of British Columbia

Today

A grab-bag of practical concerns that come up once you start running regressions

- Measurement error
- Signing the bias
- Bad controls
- Logging and functional form
- Binary outcomes
- Multiple testing

These are review topics

- I assume you've seen most of this before
- Focus is on the decisions you'll actually make in your thesis

Measurement error

Running example: the return to schooling

- We want the effect of years of education X^* on log earnings Y
- But we observe self-reported education, transcripts with errors, rounding

The observed regressor is the truth plus an error:

$$X_i = X_i^* + w_i$$

Classical measurement error assumes the error is pure noise

- w_i uncorrelated with the truth X_i^* and with the equation error e_i
- Mean zero, just adds static around the true value

Measurement error in X attenuates the slope

True model $Y_i = \alpha + \beta X_i^* + e_i$, but we regress Y on the mismeasured X:

$$\text{plim } \hat{\beta} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} = \beta \cdot \underbrace{\frac{\text{Var}(X^*)}{\text{Var}(X^*) + \text{Var}(w)}}_{\lambda \in (0,1)}$$

The estimate is biased toward zero by the reliability ratio λ

- λ is the share of observed variance that is real signal
- Half the variance is noise $\Rightarrow \lambda = 0.5 \Rightarrow$ you recover half the true slope

This is attenuation bias

- Noise in a regressor makes its effect look smaller than it is
- A signed bias: you know the truth is further from zero

Error in Y is harmless; error in X is not

Put the error in the outcome instead: observe $Y_i = Y_i^* + v_i$

$$Y_i = \alpha + \beta X_i + (e_i + v_i)$$

The noise just joins the residual

- $\hat{\beta}$ stays unbiased
- Standard errors get bigger, because the residual is noisier

The asymmetry is worth remembering

- Mismeasured outcome: less precision, no bias
- Mismeasured regressor: attenuation toward zero

Fixed effects/controls make attenuation worse

Attenuation depends on signal relative to noise

- Much of a regressor's variation is across units, not within
- Fixed effects and differencing throw away the across-unit variation

The within transformation shrinks the signal but not the noise:

$$\lambda_{\text{within}} = \frac{\text{Var}(\tilde{X}^*)}{\text{Var}(\tilde{X}^*) + \text{Var}(\tilde{w})} < \lambda$$

The fix for one problem can deepen another

- Fixed effects remove the omitted variable, but amplify measurement error
- Controls or FEs can kill a result just via the attenuation channel

IV undoes the attenuation

The trouble is that the mismeasured X is correlated with its own error w

Find an instrument Z that moves with the truth but not the noise

- Relevant: Z correlated with the true regressor X^*
- Valid: Z uncorrelated with the measurement error w (and with e)

Then the attenuation factor cancels out of the IV estimator:

$$\text{plim}\ \hat{\beta}_{IV} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)} = \frac{\beta\,\text{Cov}(Z, X^*)}{\text{Cov}(Z, X^*)} = \beta$$

A second, independent measurement of X^* is the natural instrument

- Correlated with the truth, because both measure X^*
- Its error is independent of the first measurement's error

The twins, and the IV fix

Ashenfelter and Krueger (1994) set up at the Twinsburg twins festival

- Differencing within a twin pair removes family background and genes
- But siblings have similar schooling, so the true gap is small relative to reporting error
- Each twin reports their own and their sibling's education

Variable	OLS (i)	GLS (ii)	GLS (iii)	IV[a] (iv)
Own education	0.084	0.087	0.088	0.116
	(0.014)	(0.015)	(0.015)	(0.030)
Sibling's education	—	—	−0.007	−0.037
			(0.015)	(0.029)
Age	0.088	0.090	0.090	0.088
	(0.019)	(0.023)	(0.023)	(0.019)
Age squared (÷100)	−0.087	−0.089	−0.090	−0.087
	(0.023)	(0.028)	(0.029)	(0.024)
Male	0.204	0.204	0.206	0.206
	(0.063)	(0.077)	(0.077)	(0.064)
White	−0.410	−0.417	−0.424	−0.428
	(0.127)	(0.143)	(0.144)	(0.128)
Sample size:	298	298	298	298
R^2:	0.260	0.219	0.219	—

Instrument own-reported education with the sibling's report of it

- Two independent reports of the same truth
- Their common error is the signal; reporting noise washes out
- OLS 0.084 $\rightarrow$ IV 0.116: attenuation undone

Non-classical measurement error

Classical error is the clean case

- The error has to be pure noise, uncorrelated with the truth

Real errors often correlate with the truth, and then anything goes

- Top earners under-report income; low earners round up
- Misreporting can be larger for larger true values
- Bias can go either direction, and is no longer a tidy attenuation

A misclassified binary variable is always non-classical

- If the truth is 1, the error can only be 0 or -1; if it's 0, only 0 or $+1$
- So the error is mechanically correlated with the truth

What to do

- Better data first: a second measurement, an administrative source
- Argue the likely direction

Omitted variable bias has a formula

Leave a confounder A out of the regression of Y on X:

$$\hat{\beta}_{\text{short}} = \underbrace{\beta_{\text{long}}}_{\text{true effect of } X} + \underbrace{\gamma \cdot \pi}_{\text{bias}}$$

The bias is a product of two pieces

- γ: the effect of the omitted A on Y (in the long regression)
- π: the slope of A regressed on the included X

Returns to schooling, omitting ability

- Ability raises earnings: $\gamma > 0$
- Ability is positively correlated with schooling: $\pi > 0$
- Bias is positive: OLS overstates the return

Signing the bias

You usually can't observe *A*, but you can often sign γ and π

- Multiply the two signs to get the sign of the bias
- Example: $\gamma > 0$ and $\pi < 0$ give a negative bias, so OLS understates the effect

The payoff: a signed bias is a bound

- If the bias is negative and your estimate is still positive, the true effect is larger than your estimate
- "My estimate is conservative"

Bad controls

A control either blocks a confounding path or opens a new bias

- Good control: a confounder that affects both treatment and outcome
  - Age, baseline test score, characteristics fixed before the policy
- Bad control: a variable that is itself an outcome of the treatment
  - Graduation, absences, classes taken

A timing rule that almost always works

- Measured before treatment or instrument assignment: safe to include
- Determined after: it may be a consequence of treatment, so be careful

Controlling for a consequence reintroduces bias

Estimating the return to schooling, should you control for occupation?

- Occupation is partly how schooling raises earnings
- It is determined after schooling, not before

Conditioning on it changes the comparison

- Might inadvertently affect the comparison group
- E.g., education moves person from relatively well-paid teacher $\rightarrow$ relatively poorly paid doctor
- Increases earnings, but moves person from high- to low-earnings within occupation!

Even random assignment doesn't save you here

- The treatment was randomized; the post-treatment control was not

Good questions can lead to bad controls

The temptation comes from a reasonable-sounding question

- "Is the wage gap really about schooling, or just the jobs people end up in?"
- Answering it by adding occupation creates the bad control

More controls is not safer

- The kitchen-sink regression quietly mixes good and bad controls
- A control highly correlated with treatment also inflates your standard errors

Decide each control on purpose

- Ask what it is for: blocking confounding, or absorbing an outcome
- If you want the channel, estimate it as its own outcome instead

Logs and functional form

The functional form should follow the quantity you care about

- A dollar change in Y per unit of X?
- A percent change? An elasticity?

Pick the form to deliver that number, then read the coefficient off directly

- Don't run a regression and then hunt for an interpretation
- Logs buy percent and elasticity readings; levels buy unit readings

Reporting nonlinearities

- Quadratics or polynomials when the slope changes over the range of X
- Report the slope at a meaningful value(s?), not the raw coefficients
- Show graphically

The interpretation toolkit

Form	Model	A change in X …
Level–level	$Y = \alpha + \beta X$	raises Y by β units per unit of X
Log–level	$\ln Y = \alpha + \beta X$	raises Y by 100β percent per unit of X
Level–log	$Y = \alpha + \beta \ln X$	raises Y by $\beta/100$ units per 1 percent of X
Log–log	$\ln Y = \alpha + \beta \ln X$	raises Y by β percent per 1 percent of X

Two readings worth memorizing

- Log–level coefficient: a semi-elasticity, “percent per unit”
- Log–log coefficient: an elasticity, “percent per percent”

Logging is scale-free

- Measuring income in dollars or thousands shifts only the intercept
- The slope, and its interpretation, is unchanged

The trouble with $\log(1 + Y)$

Logs need positive values, so people reach for $\log(1 + Y)$ when Y has zeros

- It looks innocent, and Stata won't complain

But there is no such thing as a percent change from a base of zero

- The elasticity reading quietly depends on the units of Y
- Rescale Y and the coefficient changes meaning, with zeros doing the work
- The convenient interpretation you wanted is exactly the one you lose

Better options when the outcome has zeros

- Estimate in levels and report a unit effect
- Poisson / PPML for a multiplicative model that handles zeros honestly
- Model the extensive margin directly with $\mathbb{1}[Y > c]$

Binary outcomes

For effects, use the linear probability model

- A 0/1 outcome regressed with OLS is the linear probability model
- The coefficient is the change in $P(Y=1)$ in percentage points
- Directly interpretable, no transformation needed

Fitted values can fall outside $[0, 1]$: fine, you wanted a slope, not a prediction

- The error is heteroskedastic by construction: use robust standard errors

Default to the LPM for treatment effects

- It works nicely with fixed effects, IV, and interactions
- A 17 percentage point effect needs no further translation

Reach for logit or probit when you need a probability

The case for a nonlinear model is when the prediction itself has to be valid

- Probabilities must land in $[0, 1]$
- Propensity scores for matching or reweighting
- Out-of-sample prediction or simulation

If you do use one, report marginal effects

- Raw logit/probit coefficients are not interpretable as effects
- The marginal effect at the mean is usually close to the LPM coefficient anyway

The choice is about the goal

- Want the effect of X? LPM
- Want a well-behaved predicted probability? logit or probit

Run enough tests and something turns up

Test one true null at the 5 percent level: 5 percent chance of a false positive

- Test 20 independent true nulls: expect one "significant" result by luck
- $P(\text{at least one false positive}) = 1 - 0.95^{20} \approx 0.64$

Theses are full of multiple tests

- Several outcomes, several subgroups, several specifications
- Each star you hunt for raises the chance one is spurious

Two ways to discipline this, controlling two different error rates

- Family-wise error rate: chance of any false positive
- False discovery rate: expected share of your findings that are false

Family-wise error rate: protect against any false positive

The FWER is the probability of at least one false rejection across the family

Bonferroni is the blunt instrument

- Test each of m hypotheses at α / m
- Guarantees FWER $\leq \alpha$, but assumes the worst about dependence
- Very conservative when your tests are correlated, which they usually are

Romano–Wolf is the one to actually use

- Resamples the data to learn the joint distribution of your test statistics
- Accounts for correlation across outcomes, so it is far less conservative
- In Stata: `rwolf` (and `rwolf2`)

False discovery rate: tolerate a known share of mistakes

When you screen many outcomes, “zero false positives” is too strict

- The FDR controls the expected fraction of rejections that are false
- Reject more, accept that some share q of them are wrong

Benjamini–Hochberg is the standard procedure

- Rank the p-values and compare each to a step-up threshold
- Reports a q-value for each test, alongside your p-values
- More power than FWER methods when you have many hypotheses

Which one to use depends on the cost of a mistake

- One wrong headline finding is costly $\Rightarrow$ control the FWER
- Screening many outcomes to flag candidates $\Rightarrow$ control the FDR

Recap

Measurement error in a regressor attenuates; in the outcome it just adds noise

- Fixed effects make attenuation worse; a second report can instrument it away

When you can't remove a bias, sign it

- A signed bias turns into a bound, and a bound can carry the argument

Choose controls and functional form on purpose

- Don't control for outcomes of treatment
- Pick the form that gives the number you want to report

Use the LPM for effects, and correct for multiple testing when you test a lot

Coming up

We're done with the methods

- You now have the toolkit: panel, DD, event studies, RD, IV, and today's concerns

The rest is about communicating the work

- Next: how to present
- After that: how to write