MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 12

FE and ME, Mastered by IV

This note recounts a 'metrics drama in three acts (see also MM Section 6.2 and the appendix to Chapter 6). First, we learn how data on siblings can be used to control for omitted variables bias in estimates of the economic returns to schooling. The key idea here is to use *panel data* to control for *unobserved individual effects*, also known as "fixed effects" (FEs). Their invisibility notwithstanding, the assumed fixedness of these effects allows us to control for them. Act II reveals, however, that the news is not all good: *attenuation bias* due to measurement error (ME) tends to shrink regression coefficients towards zero, and attenuation bias is greatly aggravated in models with fixed effects. Fixed-effects models may therefore suggest the returns to schooling are low simply because schooling is mismeasured. Finally, Act III shows how instrumental variables methods resolve the FE and ME conundrum.

1 Fixed Effects: Twins Double the Fun

Twinsburg (Ohio) embraces its zygotic heritage with an annual Twins Festival. Not wanting to miss the party, labor economists use exotic zygotic data from the Twins Festival to mitigate OVB in estimates if the economic returns to schooling.

- The long regression that motivates a twins analysis of the economic returns to schooling can be written:

$$\ln Y_{if} = \alpha^l + \rho^l S_{if} + \lambda A_{if} + e^l_{if}. \tag{1}$$

  Here, subscript f stands for family, while subscript $i = 1, 2$ indexes twin siblings, say Karen and Sharon or Ronald and Donald

- Control variable A_{if} is a measure of ability, motivation, or talent; conditional on this, we'd be prepared to assume that schooling, S_{if}, is as good as randomly assigned
  - Alas, A_{if} is not collected in the Current Population Survey

- Since Ronald and Donald have the same parents, were mostly raised together, and may even have the same genes, we might reasonably assume $A_{if} = A_f$. In other words, ability is fixed within families. Given this fixedness, we can write:

$$\ln Y_{1f} = \alpha^l + \rho^l S_{1f} + \lambda A_f + e^l_{1f}$$
$$\ln Y_{2f} = \alpha^l + \rho^l S_{2f} + \lambda A_f + e^l_{2f}.$$

  Subtracting the equation for Donald from that for Ronald gives:

$$\ln Y_{1f} - \ln Y_{2f} = \rho^l (S_{1f} - S_{2f}) + (e^l_{1f} - e^l_{2f}), \tag{2}$$

  a differenced regression model that captures the coefficient of interest and from which unobserved ability disappears!
  - From this we learn that when unobserved ability is constant within twin pairs, a regression of the *difference* in twins' earnings on the *difference* in their schooling recovers the long regression coefficient, ρ^l

- Column 1 in MM Table 6.2 reports estimates of a short regression in levels ("short" because the model omits A_{if}; "levels" because the model isn't differenced):

$$\ln Y_{if} = \alpha^s + \gamma' X_f + \rho^s S_{if} + e^s_{if}. \tag{3}$$

  This model controls for age, race, and sex in covariate vector X_f. Estimates of the differenced equation (2) appear in column 2 (why does X_f disappear from equation 2?)

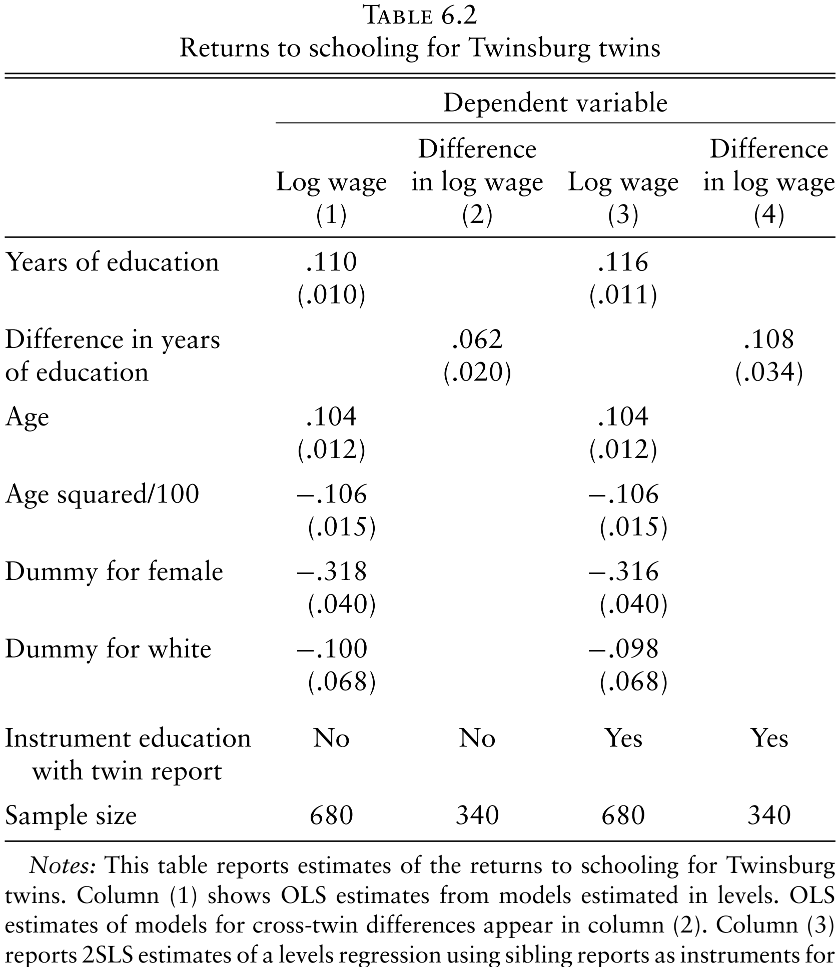

TABLE 6.2
Returns to schooling for Twinsburg twins

	Dependent variable			
	Log wage (1)	Difference in log wage (2)	Log wage (3)	Difference in log wage (4)
Years of education	.110 (.010)		.116 (.011)	
Difference in years of education		.062 (.020)		.108 (.034)
Age	.104 (.012)		.104 (.012)	
Age squared/100	−.106 (.015)		−.106 (.015)	
Dummy for female	−.318 (.040)		−.316 (.040)	
Dummy for white	−.100 (.068)		−.098 (.068)	
Instrument education with twin report	No	No	Yes	Yes
Sample size	680	340	680	340

Notes: This table reports estimates of the returns to schooling for Twinsburg twins. Column (1) shows OLS estimates from models estimated in levels. OLS estimates of models for cross-twin differences appear in column (2). Column (3) reports 2SLS estimates of a levels regression using sibling reports as instruments for

- The estimate schooling returns of just over 6% in the differenced equation (reported in column 2 of Table 6.2) is substantially below the estimate of 11% in column 1. This decline suggests the short-regression estimate of ρ^s indeed suffers from substantial ability bias

2 Measurement Error Messes Things Up

Of 340 twin pairs interviewed for the Ashenfelter and Rouse (1998) study (posted in Module E on Canvas), about half report *identical* educational attainment.

- If my brother and I are so similar, why then should our schooling differ? Good question! (My middle brother, Misha, has a Ph.D. just like me - and we're not twins.)

- Yet, if most twins really have the same schooling, then a fair number of the non-zero differences in *reported* schooling may reflect mistaken reports (Misha may not *tell* you he has a Ph.D. -- he doesn't remember his graduate work fondly)

- Masters of 'metrics refer to mistakes and misreporting in data as *measurement error*. Most people probably report their schooling correctly, but a few get it wrong. The fact that a few people report their schooling incorrectly sounds unimportant. Yet, when it comes to regression, the consequences of even minor mismeasurement can be major.

- And then, there's this: Mismeasured schooling affects (2) much more than it does (1).

Interlude: Attenuation Bias

Let's simplify for a moment: forget ability bias and twins, focus only on measurement. Suppose you've dreamed of running the regression:

$$Y_i = \alpha + \beta S_i^* + e_i, \tag{4}$$

but data on S_i^*, the regressor of your dreams, is unavailable.

- You see only a mismeasured version, S_i:

  $$S_i = S_i^* + u_i, \tag{5}$$

  where u_i is the measurement error in S_i

- Assume that measurement error is mean-zero and uncorrelated with S_i^* and e_i:

  $$E\left[u_i\right] = 0 \tag{6}$$
  $$C\left(S_i^*, u_i\right) = C\left(e_i, u_i\right) = 0 \tag{7}$$

  These assumptions are said to describe "classical measurement error" . Note that the first part of (7) implies:

  $$V\left(S_i\right) = V\left(S_i^*\right) + V\left(u_i\right).$$

- The regression coefficient we're after, β in (4), is given by:

  $$\beta = \frac{C\left(Y_i, S_i^*\right)}{V\left(S_i^*\right)}. \tag{8}$$

  But we don't know S_i^*. Regressing Y_i on mismeasured S_i instead yields slope coefficient:

$$\begin{aligned}
\beta_b &= \frac{C\left(Y_i, S_i\right)}{V\left(S_i\right)} \\
&= \frac{C(\alpha + \beta S_i^* + e_i, S_i^* + u_i)}{V\left(S_i\right)} \\
&= \frac{C(\alpha + \beta S_i^* + e_i, S_i^*)}{V\left(S_i\right)} = \beta \frac{V\left(S_i^*\right)}{V\left(S_i\right)}
\end{aligned}$$

The 3rd equals sign above uses the classical assumptions, (7); be sure you can see how.

- We can now write:
$$\beta_b = r\beta, \tag{9}$$
  where
$$r = \frac{V\left(S_i^*\right)}{V\left(S_i\right)} = \frac{V\left(S_i^*\right)}{V\left(S_i^*\right) + V\left(u_i\right)},$$
  is a number between zero and one
  - Fraction r is called the *reliability* of S_i
  - Reliability reveals the extent of proportional *attenuation bias* in β_b:
$$\frac{\beta_b}{\beta} = r$$
  - β_b is closer to zero than β unless $r = 1$ (in which case, there's no measurement error after all)

Covariates and Differencing Aggravate Attenuation Bias

The addition of covariates to a model with mismeasured regressors exacerbates attenuation bias.

- Suppose the regression of interest is:
$$Y_i = \alpha + \gamma X_i + \beta S_i^* + e_i, \tag{10}$$
  where X_i is a control variable, perhaps IQ or a test score. Regression anatomy says:
$$\beta = \frac{C(Y_i, \widetilde{S}_i^*)}{V(\widetilde{S}_i^*)},$$
  where $\widetilde{S}_i^*$ is the residual from a regression of S_i^* on X_i

- Replacing S_i^* with S_i in (10), the coefficient on S_i becomes:
$$\beta_b = \frac{C(Y_i, \widetilde{S}_i)}{V(\widetilde{S}_i)},$$
  where $\widetilde{S}_i$ is the residual from a regression of S_i on X_i

- In models with covariates, it's customary to assume measurement error and covs are uncorrelated, that is, $E[X_i u_i] = 0$ (we've already assumed $E[S_i^* u_i] = 0$, so this seems a natural extension). Given this assumption, we have:
$$\widetilde{S}_i = \widetilde{S}_i^* + u_i, \tag{11}$$
  where u_i and $\widetilde{S}_i^*$ are uncorrelated with each other (show this). We therefore have:
$$V(\widetilde{S}_i) = V(\widetilde{S}_i^*) + V(u_i).$$
  Note also that $V(\widetilde{S}_i^*) < V\left(S_i^*\right)$ when covariates predict true schooling (as seems likely)

- Applying the same logic used to establish (9), we get:

$$\begin{aligned}\beta_b = \frac{C(Y_i, \widetilde{S}_i)}{V(\widetilde{S}_i)} &= \frac{C(Y_i, \widetilde{S}^*_i + u_i)}{V(\widetilde{S}^*_i + u_i)} \\ &= \frac{V(\widetilde{S}^*_i)}{V(\widetilde{S}^*_i) + V(u_i)}\beta = \tilde{r}\beta, \end{aligned} \tag{12}$$

where

$$\tilde{r} = \frac{V(\widetilde{S}^*_i)}{V(\widetilde{S}^*_i) + V(u_i)} < \frac{V(S^*_i)}{V(S^*_i) + V(u_i)} = r.$$

Because covariates reduce the variance of the signal in S_i, while leaving the variance of the noise unchanged, ***covariates aggravate attenuation bias.***

Fixed effects are a worst-case scenario for covariate-aggravated attenuation bias

- To see why, consider a panel model for the effects of true schooling:

$$Y_{if} = \alpha_f + \beta S^*_{if} + e_{if}, \tag{13}$$

where $\alpha_f = \alpha^l + \lambda A_f$ and A_f is unobserved ability, as before

  - As noted in Section 1, we can eliminate the fixed effect by differencing:

$$Y_{1f} - Y_{2f} = \beta\left(S^*_{1f} - S^*_{2f}\right) + e_{1f} - e_{2f}, \tag{14}$$

- In this scenario, we might imagine that true schooling is also similar within families, so that within-twin differences are mostly noise. We can describe this extreme scenario by modeling observed schooling in the twins panel as:

$$S_{if} = S^*_f + u_{if} \tag{15}$$

where S^*_f is true schooling, fixed within families, and u_{if} is twin-specific reporting error.

  - In this extreme case, the observed difference in schooling is *entirely* noise:

$$S_{1f} - S_{2f} = u_{1f} - u_{2f} \tag{16}$$

  (what, then, will we get from estimation of differenced equation (2)?)

  - In practice, $S_{1f} - S_{2f}$ is probably not *all* noise. More realistically, we have:

$$S_{1f} - S_{2f} = (S^*_{1f} - S^*_{2f}) + (u_{1f} - u_{2f}). \tag{17}$$

  Even so, because S^*_{1f} and S^*_{2f} are so similar, the difference between them, $S^*_{1f} - S^*_{2f}$, has variance well below that of the difference in measured schooling, $S_{1f} - S_{2f}$

- Attenuation bias in differenced equation (2) is likely much worse than attenuation bias in the levels equation (3). Aggravated attenuation bias provides an alternative explanation (besides ability bias) for the sharp decline in schooling coefficients as we move from column 1 to column 2 in MM Table 6.2

3 IV to the Rescue

Attenuation bias may make the differencing cure for ability bias worse than the disease. But all is not lost.

- Recall from LN11 that the IV estimator of the coefficient on S_i in a bivariate regression of Y_i on S_i is the sample analog of:

  $$\frac{C(Y_i, Z_i)}{C(S_i, Z_i)}, \tag{18}$$

  where the instrumental variable is Z_i.

  - In the measurement error version of the IV story, we use Z_i to instrument for mismeasured S_i
  - In the context of an equation like (4), this works when Z_i *is correlated with* S_i^**, but uncorrelated with both measurement error,* u_i*, and the residual* in the equation of interest, e_i

- To see how IV gives us what we want, use (4) and (5) to substitute for Y_i and S_i in (18):

  $$\begin{aligned}\frac{C(Y_i, Z_i)}{C(S_i, Z_i)} &= \frac{C(\alpha + \beta S_i^* + e_i, Z_i)}{C(S_i^* + u_i, Z_i)} \\ &= \frac{\beta C(S_i^*, Z_i) + C(e_i, Z_i)}{C(S_i^*, Z_i) + C(u_i, Z_i)}.\end{aligned}$$

- Assuming $C(e_i, Z_i) = C(u_i, Z_i) = 0$, we then have:

  $$\frac{C(Y_i, Z_i)}{C(S_i, Z_i)} = \beta \frac{C(S_i^*, Z_i)}{C(S_i^*, Z_i)} = \beta.$$

Attenuation bias begone!

- IV solutions to measurement error problems often exploit multiple measures of the variable of interest. If only we had two measures of schooling!
  - We do: the Twinsburg survey asks each twin to report not only his or her own schooling but also that of their sibling. We therefore have two measures of schooling for each twin, one self-report and one sibling-report.

- This extra info is especially valuable for the measurement-error-afflicted differenced equation. Assuming the measurement error in self- and sibling-reports is uncorrelated (i.e., mistakes I make in reporting my own schooling are uncorrelated with mistakes my sibling makes in reporting my schooling), the difference in sibling reports can be used to instrument the difference in self-reports in equation (2)
  - In the IV formula, (18), the variable to be instrumented is

    $$S_i \equiv (S_{1f} - S_{2f}),$$

    while $Y_i \equiv (\ln Y_{1f} - \ln Y_{2f})$,
  - The instrument is

    $$Z_i \equiv \left(S_{1f}^2 - S_{2f}^1\right),$$

    where S_{if}^j is sibling j's report of sibling i's schooling

- **The resulting IV estimates, reported in cols 3-4 in Table 6.2, suggest the decline in returns to schooling from columns 1 to 2 is indeed due to ME rather than OVB.**

  And so the curtain falls on our story.