FE and ME, Mastered by IV

This note recounts a 'metrics drama in three acts. First, we see 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). Invisibility notwithstanding, it's these effects fixedness that 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 regression models with fixed effects. Models with fixed effects may therefore suggest the returns to schooling are low simply because schooling is measured poorly. Finally, Act III shows how instrumental variables methods resolve the FE/ME conundrum.

1 Fixed Effects: Twins Double the Fun

Twinsburg (Ohio) embraces its zygotic heritage with an eponymous annual Twins Festival. Not wanting to miss the party, labor economists use exotic zygotic data from the Twins Festival to control for OVB.

- 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 which we imagine schooling, S_{if}, is as good as randomly assigned.
  - Alas, A_{if} is not part of 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 $A_{if} = A_f$. 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 regression model that captures the coefficient of interest and from which unobserved ability disappears!
  - From this we learn that when 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}):

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

  This model includes controls for age, race, and sex in vector X_i (why do these disappear in equation 2?), alongside estimates of differenced equation (2) in column 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 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 much ability bias in ρ^s!

2 Measurement Error Messes Things Up

Of 340 twin pairs interviewed for the Ashenfelter and Rouse (1998) study, about half report *identical* educational attainment.

- If my brothers and I are so similar, then why should our schooling differ? Good question! 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.
- The problem of mistakes in regressors is known as *measurement error*. The fact that a few people report their schooling incorrectly sounds unimportant, yet, when it comes to regression, the consequences of such measurement error may be major.
- Mismeasured schooling affects (2) more than (1)

Interlude: Attenuation Bias

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, are 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:

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

  These assumptions are said to describe "classical measurement error"

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

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

  Alas, we must work with mismeasured regressor, S_i, instead of S_i^*. This yields slope coefficient:

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

- Therefore,

  $$\beta_b = r\beta, \tag{9}$$

  where

  $$r = \frac{V(S_i^*)}{V(S_i)} = \frac{V(S_i^*)}{V(S_i^*) + V(u_i)},$$

  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

- Assume measurement error, u_i, is "pure noise," and so uncorrelated with covariate X_i. The pure noise hypothesis implies:

$$\widetilde{S}_i = \widetilde{S}_i^* + u_i, \tag{11}$$

  where u_i and $\widetilde{S}_i^*$ are uncorrelated. We therefore have:

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

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

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

  where

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

  Covariates reduce the variance of the signal in S_i, while leaving the variance of the noise unchanged. ***The resulting reduction in signal aggravates attentuation bias.***

- ***Fixed effects are likely to be a worst-case version of this***
  - To see why, replace (10) with a panel model

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

    where $\alpha_f = \alpha^l + \lambda A_f$ is an unobserved fixed-within-family "ability" effect

– 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 changes are mostly noise. Paralleling (11), we have

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

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

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

  More generally, we expect the differencing transformation to kill more signal than noise.

  – In practice, $S_{1f} - S_{2f}$ is probably not *all* noise
  – But it gets worse: if measurement errors is uncorrelated across siblings, then the variance of the noise in the sibling schooling difference is twice the variance of the noise in levels (compare the variance of measurement errors in 15 and 16)

- This bodes ill for OLS estimates of equation (2) and provides an alternative explanation (besides ability bias) for the sharp decline in schooling coefficients as we move from column 1 to column 2 in Table 6.2

3 IV to the Rescue

We've seen that with a mismeasured regressor, OLS estimation fails to produce the coefficient we're after. But all is not lost.

- Recall from the previous IV notes 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:

$$\beta_{IV} = \frac{C\left(Y_i, Z_i\right)}{C\left(S_i, Z_i\right)}, \tag{17}$$

  where the instrumental variable is Z_i. In a measurement error story, we use Z_i to instrument for mismeasured S_i, an estimation strategy justified by assuming Z_i *is uncorrelated with both measurement error and the residual*, e_i.

- To see what this accomplishes, use (4) and (5) to substitute for Y_i and S_i in (17):

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

- Since $C\left(e_i, Z_i\right) = C\left(u_i, Z_i\right) = 0$, we have:

$$\beta_{IV} = \beta\frac{C\left(S^*_i, Z_i\right)}{C\left(S^*_i, Z_i\right)} = \beta.$$

  Attenuation bias begone!

- IV solutions to measurement error problems often exploit multiple measures of the same underlying construct. If only, we had two measures of schooling! We do: the Twinsburg sample survey asked 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.
- Assuming the measurement errors in self- and sibling-reports are uncorrelated (i.e., the 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)
  - Translating this notation to equation (2), the variable to be instrumented is $S_i \equiv (S_{1f} - S_{2f})$
  - The instrument is $Z_i \equiv \left(S^2_{1f} - S^1_{2f}\right)$ where S^j_{if} 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 due to ME rather than OVB.**

  And so the curtain falls on our story of ability bias.