In-Class Activity: Instrumenting Education — Solutions

A labor economist wants to estimate the **causal effect of years of education on log hourly wages**. She uses data on 3,010 U.S. men born between 1930 and 1939 from the 1980 Census. After hanging out with the ECON3500 crowd for too long, she is appropriately worried that OLS estimates are biased because **ability** is unobserved: more able individuals get more education *and* earn higher wages, even holding education constant.

Her proposed instrument: **quarter of birth**. U.S. compulsory schooling laws require students to stay in school until age 16. But students' age when they first enter school is staggered because of age cut-offs. Hence, students born earlier in the year reach age 16 earlier in the school year, so they can legally drop out with less total education. The idea is that quarter of birth thus affects years of education completed but should not directly affect wages.

(Stylized example in the spirit of Angrist and Krueger, 1991 — numbers constructed to prove my specific pedagogical points.)

Regression Output

	(1)	(2)	(3)	(4)
Dep. variable:	log(wage)	educ	log(wage)	log(wage)
Method:	OLS	First stage	Reduced form	2SLS
Years of education	0.0700***			0.142**
	(0.0035)			(0.061)
Born Q1 (Jan–Mar)		$-$0.152**	$-$0.0216**	
		(0.067)	(0.0093)	
Black	$-$0.236***	$-$1.47***	$-$0.342***	$-$0.133*
	(0.018)	(0.12)	(0.022)	(0.075)
Married	0.121***	0.38***	0.129***	0.075*
	(0.016)	(0.11)	(0.020)	(0.043)
Region (South)	$-$0.098***	$-$0.54***	$-$0.137***	$-$0.060
	(0.015)	(0.10)	(0.018)	(0.038)
Constant	5.02***	12.84***	5.93***	4.11***
	(0.052)	(0.078)	(0.062)	(0.78)
N	3,010	3,010	3,010	3,010
R^2	0.132	0.087	0.041	—
First-stage F	—	5.15	—	—

Standard errors in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Column (3) is OLS of log(wage) on Born Q1 and controls. Column (4) instruments for years of education using Born Q1.

Questions + Answers

1. Interpret the OLS coefficient on years of education. Be precise about units and magnitude.

   Answer. *One additional year of education is associated with approximately 7.0% higher hourly wages, controlling for race, marital status, and region. (Technically: a one-year increase in education is associated with a 0.070 increase in log wages, or about 7.0% higher wages.)*

2. Interpret the coefficient on "Born Q1" in the first-stage regression. What does it tell us about the relationship between quarter of birth and years of education?

   Answer. *Being born in Q1 (January–March) is associated with* ***0.152 fewer years of education*** *than being born in Q2–Q4, controlling for race, marital status, and region. Consistent with compulsory schooling laws: Q1 students reach the legal dropout age earlier and can exit with less total schooling.*

3. The first-stage F-statistic on the excluded instrument is $F = 5.15$. What does this tell us? Should we be concerned? Why or why not?

   Answer. $F = 5.15$ *is* ***below the conventional threshold of 10*** *for strong instruments — a* ***weak instrument*** *concern. Consequences: (1) 2SLS estimates are biased toward OLS in finite samples; (2) standard errors are unreliable (CIs have incorrect coverage); (3) the 2SLS point estimate may not be trustworthy. This is a well-known limitation of the Angrist-Krueger design with a single QOB indicator. BJB (1995) formalized the critique.*

4. Interpret the 2SLS coefficient on years of education. How does it compare to the OLS estimate?

   Answer. *The 2SLS estimate implies that an additional year of education causes approximately a 14.2% increase in hourly wages —* ***twice as large*** *as the OLS estimate of 7.0%. Standard errors are also much larger (0.061 vs. 0.0035), which is typical of IV: we use less variation (only the part driven by the instrument), so estimates are noisier.*

5. With a single instrument and the same controls in all regressions, the 2SLS coefficient equals the ratio of the **reduced-form** coefficient on Born Q1 (column 3) to the **first-stage** coefficient on Born Q1 (column 2):

$$\hat{\beta}_1^{\text{2SLS}} \approx \frac{\hat{\gamma}_{Q1}^{\text{RF}}}{\hat{\pi}_{Q1}^{\text{FS}}}.$$

   Verify this numerically using the table. Then explain why, when the first-stage coefficient (the denominator) is small, the IV estimate gets amplified — and why a weak first stage is so dangerous.

   Answer. *Numerical check:* $\hat{\gamma}_{Q1}^{RF}/\hat{\pi}_{Q1}^{FS} = (-0.0216)/(-0.152) \approx 0.142$, *exactly the 2SLS coefficient in column (4). So 2SLS isolates the slice of education variation that tracks with the instrument, then asks: how big is the* Y*-response per unit of that variation?*

   Intuition for weak-instrument danger: the denominator is small (Q1 births only shift education by 0.15 years), so any wage gap induced by Q1 gets divided by a small number, *amplifying both the point estimate* ***and*** *the noise. That's why a small first stage (low F) makes IV estimates large AND unreliable.*

6. Given that the OLS estimate (0.070) is *smaller* than the 2SLS estimate (0.142), what does this imply about the direction of bias in OLS? Is this surprising? Why or why not?

   Hint: Think about what you would expect if ability bias drives OLS upward. What else could be going on?

   ANSWER. ***Surprising*** *if ability bias is the main problem, because ability bias should push OLS up, not down. Three possible explanations:*

   (a) ***Measurement error in education.*** *If education is measured with error (self-reported years), OLS is attenuated toward zero. IV corrects for classical measurement error because the instrument is correlated with true education, not the noise. Attenuation can dominate the upward ability bias. (Leading explanation in the returns-to-schooling literature.)*

   (b) ***LATE $\neq$ ATE.*** *The IV estimate applies to compliers — marginal dropouts. Returns to schooling may be higher for this group than the population average (heterogeneous treatment effects). So IV $>$ OLS could reflect higher returns for compliers, not a downward OLS bias.*

   (c) ***Weak-instrument bias.*** *With $F = 5.15$, the 2SLS estimate is itself biased toward OLS in finite samples; the true IV (with a strong instrument) might be even larger, or the current estimate is simply unreliable.*

7. The IV estimate is a **Local Average Treatment Effect (LATE)**. In this context, who are the *compliers* — the group whose behavior is actually affected by the instrument?

   ANSWER. *Individuals whose total years of education were affected by compulsory schooling laws — those who would have dropped out earlier if they could have (i.e., if they had reached 16 before the end of the school year). These are people at the margin of the dropout decision. The IV estimate does* ***not*** *apply to always-takers (college-bound students who would have stayed regardless) or never-takers (students who dropped out before 16 regardless).*