Quantitative Economics

Applied Microeconomics II - The Returns to Schooling

Ian Crawford
(Material provided by Climent Quintana-Domeque)

Department of Economics

Nuffield College

Trinity Term, 2018

Outline

- Controlling for ability (omitted variable bias, attenuation bias, “bad controls”).
- Twins
- Exploiting variation in birth quarter (IV/2SLS).

Educashun

- There is a strong positive correlation between education and earnings.
- Two main explanations with diametrically opposing causal stories:

1. Human capital theory: education makes you more productive.
2. Signalling theory: productive people acquire education

Education as Human Capital

- Schooling is an investment (in human capital), with a monetary payoff similar to that of a financial investment.
- Investment decisions depend on NPV calculations. Are the foregone earnings and tuition costs associated with a degree worthwhile?
- Causal effect of education on earnings: **"the returns to schooling"**

Education as a Signal

- Schooling does not increase productivity
- Productivity is unobserved
- More productive individuals acquire more schooling to signal their ability type
- Schooling as an expensive ability test.

The Economic Returns to Education

- Treatment: having a degree

$$S_i = \begin{cases} 1 & \text{if degree} \\ 0 & \text{if no degree} \end{cases}$$

- Outcome of interest: ln earnings or ln wage, $\ln Y_i$
- Potential outcomes: what would have earned someone who has a degree if she had not had a degree and vice versa

$$\ln Y_i = \begin{cases} \ln Y_i(1) & \text{if } S = 1 \\ \ln Y_i(0) & \text{if } S_i = 0 \end{cases}$$

The Economic Returns to Education

- Causal effect of a degree on ln earnings for individual i:

$$\ln Y_i(1) - \ln Y_i(0)$$

- Observed ln earnings can be written as

$$\ln Y_i = \ln Y_i(0) + [\ln Y_i(1) - \ln Y_i(0)]\, S_i$$

- If D_i is "as good as randomly assigned" the population regression of $\ln Y_i$ on S_i

$$\ln(Y_i) = \alpha + \rho S_i + e_i$$

will give us a parameter ρ with a causal interpretation.

The Economic Returns to Education

- ρ measures the (approximate) percentage increase in earnings due to a degree

$$\rho = \ln Y_i(1) - \ln Y_i(0) = \ln \left(1 + \underbrace{\frac{Y_i(1) - Y_i(0)}{Y_i(0)}}_{\text{\% change in earnings}} \right)$$

- Now if % change in earnings is small, then

$$\rho \simeq \frac{Y_i(1) - Y_i(0)}{Y_i(0)}$$

The Economic Returns to Education

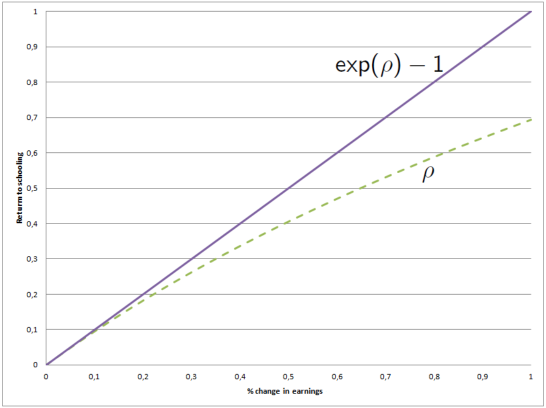
- $\exp(\rho) - 1$ measures the (exact) percentage increase in earnings due to a degree

$$\exp(\rho) = \exp\left(\ln\left(\frac{Y_i(1)}{Y_i(0)}\right)\right) = \frac{Y_i(1)}{Y_i(0)}$$

- Hence

$$\exp(\rho) - 1 = \frac{Y_i(1) - Y_i(0)}{Y_i(0)}$$

The Economic Returns to Education



The Economic Returns to Education

- Since S_i is not (as good as) randomly assigned, we will not learn about the causal effect fo interest by running a regression of $\ln Y_i$ on S_i.
- We can still run the regression, we will still get a parameter, but that parameter will not have a causal interpretation.
- We can attempt to identify the causal; effect by using the conditional independence assumption and adding regressors which account for ("control for") non-random assignment of schooling.

Omitted Variable Bias

- Suppose that years of schooling S_i are “as good as randomly assigned” conditional on ability A_i
- We can think of invoking the conditional independence assumption and running the following “long regression”

$$Y_i = \alpha^L + \rho^L S_i + \gamma A_i + e_i^L$$

- If ability is unobserved we can only run the “short regression”

$$Y_i = \alpha^S + \rho^S S_i + e_i^S$$

- What do we get?

Omitted Variable Bias

- The OLS population regression coefficient ρ^S is given by

$$\rho^S = \frac{Cov(Y_i, S_i)}{Var(S_i)} = \rho^L + \underbrace{\gamma \frac{Cov(A_i, S_i)}{Var(S_i)}}_{Ability\ Bias}$$

- *Ability Bias* $\neq 0$ if and only if

1. $\gamma \neq 0$: ability is correlated with earnings
2. $\frac{Cov(A_i, S_i)}{Var(S_i)} \neq 0$: ability is correlated with schooling

- We expect $\gamma > 0$: ability is positively correlated with earnings
- How about the relationship between ability and schooling?

Omitted Variable Bias

- What is the correlation between ability and schooling?
  - If they are substitutes, $\frac{Cov(A_i,S_i)}{Var(S_i)} < 0$
  - If they are complements, $\frac{Cov(A_i,S_i)}{Var(S_i)} > 0$
- How does ρ^S compare to ρ^L?
  - If they are substitutes, $\rho^S < \rho^L$
  - If they are complements, $\rho^S > \rho^L$

Omitted Variable Bias

Table 1. OLS regressions of (natural) log earnings: Short vs. Long[1]

	Short Regression	Long Regression
Years of schooling	0.068	0.059
	(0.003)	(0.003)
IQ score in high school		0.0028
		(0.0005)

Note: Standard errors in parentheses.

This comparison suggests

$$\frac{Cov(A_i, S_i)}{Var(S_i)} > 0$$

[1] These results are from: Griliches, Z. (1977) "Estimating the Returns to Schooling," *Econometrica*, 45(1):1-22

Omitted Variable Bias

- Intriguing results, but hard to see them as conclusive...
  1. Ability is a multidimensional concept
  2. IQ is likely to be measured with error
  3. Education can be measured with error too
- We will focus here on (3) if only because...

Measurement error

- Many economic variables are mismeasured
- We use proxy variables for economic concepts
- What is the role of measurement error in schooling?

Measurement error - attenuation bias

1. **Measurement error in the Short Regression**

- Let the SR be

$$Y_i = \alpha^S + \rho^S S_i^* + e_i^S$$

where S_i^* is unobserved.

- What we observe is S_i

$$S_i = S_i^* + m_i$$

where m_i is classical measurement error (CME)

$$\mathbb{E}\left[m_i\right]) = 0$$

$$Cov(m_i, S_i^*) = 0$$

$$Cov(m_i, e_i^S) = 0$$

Measurement error - attenuation bias

- The return to schooling in the SR is

$$\rho^S = \frac{Cov(Y_i, S_i^*)}{Var(S_i^*)}$$

- The return to schooling in the SR with measurement error is

$$\rho^{S'} = \frac{Cov(Y_i, S_i)}{Var(S_i)} = \frac{Cov(\alpha^S + \rho^S S_i^* + e_i^S, S_i^* + m_i)}{Var(S_i^* + m_i)} =$$
$$= \rho^S \frac{Var(S_i^*)}{Var(S_i^*) + Var(m_i)} = \rho^S \lambda$$

- λ is the proportion of the variation in schooling unrelated to the measurement error = "reliability" of S_i

$$\rho^{S'} - \rho^S = \rho^S \lambda - \rho^S = -(1 - \lambda)\rho^S$$

- Ashenfelter and Krueger (1994) find a reliability ratio of 0.9 for schooling.

Measurement error - attenuation bias

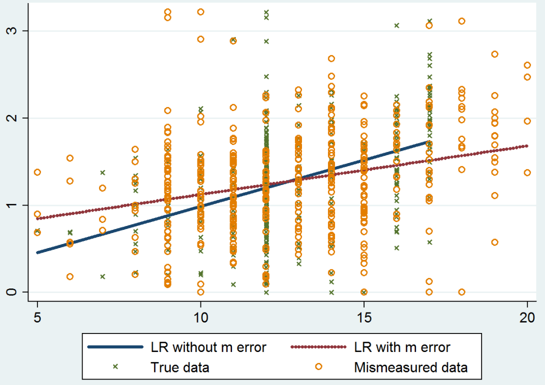
Table 2. OLS regressions of (natural) log hourly earnings: Attenuation bias[2]

	No measurement error	Measurement error
Years of schooling	0.106	0.056
	(0.012)	(0.009)
N	411	411

Note: Standard errors in parentheses. The observations, from the Panel Study of Income Dynamics (PSID), are working married women with an hourly wage of at least 1 US dollar per hour. I have generated measurement error in years of schooling as $m \sim U[-2, 2]$.

[2] The data come from: Mroz, T. (1987) "The Sensitivity of An Empirical Model Of Married Women's Hours of Work To Economic And Statistical Assumptions," *Econometrica*, 55, (4), 765-799

Measurement error - attenuation bias



Measurement error - attenuation bias

2. Measurement error in the Long Regression

- Let the LR be

$$Y_i = \alpha^L + \rho^L S_i^* + \gamma A_i + e_i^L$$

  where S_i^* is unobserved.

- What we observe is S_i

$$S_i = S_i^* + m_i$$

  where m_i is classical measurement error

$$E(m_i) = 0$$

$$Cov(m_i, S_i^*) = 0$$

$$Cov(m_i, e_i^L) = 0$$

$$Cov(m_i, A_i) = 0$$

Measurement error - attenuation bias

- The return to schooling in the LR is

$$\rho^L = \frac{Cov(Y_i, \widetilde{S}_i^*)}{Var(\widetilde{S}_i^*)}$$

where

$$S_i^* = \pi_0 + \pi_1 A_i + \widetilde{S}_i^*$$

- The return to schooling in the LR with measurement error is

$$\rho^{L'} = \frac{Cov(Y_i, \widetilde{S}_i)}{Var(\widetilde{S}_i)}$$

where

$$S_i = \delta_0 + \delta_1 A_i + \widetilde{S}_i$$

- CME in S_i translates into CME in $\widetilde{S}_i$:

$$S_i = S_i^* + m_i \Rightarrow \widetilde{S}_i = \widetilde{S}_i^* + m_i$$

Measurement error - attenuation bias

- The OLS estimand of the return to schooling in the LR with measurement error is

$$\rho^{L'} = \frac{Cov(Y_i, \widetilde{S}_i)}{Var(\widetilde{S}_i)} = \frac{Cov(\alpha^L + \rho^L S_i^* + \gamma A_i + e_i^L, \widetilde{S}_i^* + m_i)}{Var(\widetilde{S}_i^* + m_i)} =$$

$$= \rho^L \frac{Var(\widetilde{S}_i^*)}{Var(\widetilde{S}_i^*) + Var(m_i)} = \rho^L \widetilde{\lambda}$$

- Note that

$$Var(\widetilde{S}_i) = Var(\widetilde{S}_i^*) + Var(m_i)$$

- Note that

$$Var(S_i^*) = Var(\pi_0 + \pi_1 A_i + \widetilde{S}_i^*) = \pi_1^2 Var(A_i) + Var(\widetilde{S}_i^*) > Var(\widetilde{S}_i^*$$

$$\lambda = \frac{Var(S_i^*)}{Var(S_i^*) + Var(m_i)} > \frac{Var(\widetilde{S}_i^*)}{Var(\widetilde{S}_i^*) + Var(m_i)} = \widetilde{\lambda}$$

Lessons for empirical practice

- Adding (innate or pre-determined) ability controls...
    - reduces ability bias (pro)
    - aggravates attenuation bias (con)
- Net effect?

$$\rho^{S'} = \rho^S \lambda \leq \text{or} \geq \rho^L \widetilde{\lambda} = \rho^{L'}$$

- Empirical question
    - Bottom line: adding controls is not necessarily a good thing
    - Adding correlated regressors makes attenuation bias worse
- Indeed, things can become even worse with "bad controls".

“Bad Controls”

Good vs. Bad controls

- What makes a good control variable?
- Variables correlated with both education and earnings, as long as they are fixed or pre-determined characteristics with respect to schooling (innate ability), may help in reducing ability bias
- However, when schooling is mismeasured, attenuation bias gets worse
- Variables uncorrelated with schooling but correlated with earnings may help in reducing standard errors (increasing the precision of our estimates)
- What makes a bad control variable? Variables which are outcomes of schooling (e.g., occupation)

“Bad Controls”

- Treatment: having a degree

$$D_i = \begin{cases} 1 & \text{if degree} \\ 0 & \text{if no degree} \end{cases}$$

- Outcomes: log earnings Y_i (note the change in notation)
- Occupation status W_i

$$W_i = \begin{cases} 1 & \text{if white collar job} \\ 0 & \text{if blue collar job} \end{cases}$$

- Key point: having a degree affects both occupation and earnings

“Bad Controls”

- Observed occupational status and earnings can be written as

$$W_i = W_i(0) + [W_i(1) - W_i(0)]\, D_i(0)$$

$$Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]\, D_i(0)$$

- Two causal effects:
  1. $W_i(1) - W_i(0)$ is the causal effect of having a degree on occupational status
  2. $Y_i(1) - Y_i(0)$ is the causal effect of having a degree on earnings

"Bad Controls"

- Suppose that D_i is randomly assigned:

$$\{Y_i(1), W_i(1), Y_i(0), W_i(0)\} \perp\!\!\!\perp D_i$$

- We can identify two average causal effects
  1. Average causal effect of having a degree on earnings

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

  2. Average causal effect of having a degree on occupational status

$$\mathbb{E}[W_i|D_i = 1] - \mathbb{E}[W_i|D_i = 0] = \mathbb{E}[W_i(1) - W_i(0)]$$

“Bad Controls”

- Can we identify the causal effect of a degree on earnings for those in a white collar job?

$$
\begin{aligned}
&\mathbb{E}[Y_i|W_i = 1, D_i = 1] - \mathbb{E}[Y_i|W_i = 1, D_i = 0] = \\
&= \mathbb{E}[Y_i(1)|W_i(1) = 1, D_i = 1] - \mathbb{E}[Y_i(0)|W_i(0) = 1, D_i = 0] = \\
&\underbrace{=}_{RA\ of\ D_i} \mathbb{E}[Y_i(1)|W_i(1) = 1] - \mathbb{E}[Y_i(0)|W_i(0) = 1] = \\
&= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|W_i(1) = 1]}_{(1)} + \\
&+ \underbrace{\mathbb{E}[Y_i(0)|W_i(1) = 1] - \mathbb{E}[Y_i(0)|W_i(0) = 1]}_{(2)}
\end{aligned}
$$

(1) is the average causal effect of a degree on workers who have a white collar job because they have a degree.
(2) is a compositional bias: university education changes the composition of white collar workers.

“Bad Controls” (Intuition)

- Random assignment of degrees balances mean characteristics of individuals with and without a degree
- Conditioning on occupation breaks down the balancing:
  - Individuals with a degree are more likely to end up in a white collar job...
  - Individuals without a degree who end up in a white collar job are likely to have better-than-average $Y_i(0)$

The perils of control variables

- We do not observe all relevant variables
- Measurement error
- Bad controls

Summary

- We cannot (should not) always randomise.
- **Conditional Independence Assumption**: Conditional on ability, education is "as good as randomly assigned"
- Main limitations of this approach:
  - Ability is multidimensional: $\mathbf{A}_i$ is a vector
  - Finding one ability measure (IQ) is difficult
  - Finding the whole relevant vector of abilities is probably impossible
- Measurement error: attenuation bias
- Attenuation bias gets worse with control variables
- Bad control(s)? A variable that is an outcome of schooling (e.g., occupation)

Twins

- Twin siblings have much in common:
  - most grow up in the same family (same environment, *nurture*)
  - some are genetically identical (same genes, *nature*)
- Types of twins:
  - **Monozygotic ("identical") twins**: they result from the splitting of a fertilised egg and are considered to be genetically identical
  - **Dizygotic ("fraternal") twins**: they result from the fertilisation of separate eggs and lead only to siblings that are genetically similar, as are non-twins brothers and sisters.

Twins

- Focus on **identical twins** (Monozygotic (MZ) twins)
- Assumptions:
  1. Twins are identical: One twin provides a good control for the other.
  2. The fact that one twin gets more schooling than his or her twin sibling is due mostly to serendipitous forces/events.
  3. Twin differences in schooling are random.

Twins

- Where do we get data on Twins?
- The Annual Twins Days Festival in Twinsburg, Ohio. The largest gathering of twins in the world
- In 1991 Orley Ashenfelter and Alan Krueger collected twins data in the 16th Annual Twins Days Festival

Twins

16th Annual Twins Days Festival: `http://www.twinsdays.org/`

Twins

Table 1: Descriptive Statistics[3]

	Identical twins	Fraternal twins	Population
Self-reported education	14.1	13.7	13.1
Sibling-reported education	14.0	13.4	–
Hourly wage	\$13.3	\$12.0	\$11.1
Age	36.6	35.6	38.9
White	0.94	0.93	0.87
Female	0.54	0.48	0.45
Twins report same education	0.49	0.43	–
N	298	92	164,085

Note: Population data from the 1990 Current Population Survey.

[3] These results are from: Ashenfelter, O., and A. B. Krueger (1994) "Estimates of the Economic Returns to Schooling from a New Sample of Twins," *American Economic Review*, 84(5):1157-1173.

Twins

Sample of twins from 16th Annual Twins Days Festival (1991) vs. CPS sample (1990)

- Compared to individuals in the CPS Sample, twins are
  - better educated
  - more highly paid
  - younger
  - more likely to be female and white
- Identical twins in Ohio tend to have similar education levels
- **Identical** twins are more similar than **fraternal** twins: 49% vs. 43% report the same education

Twins

The Long Regression for Twins

- The LR for a MZ twin *i* of family *f* can be written as:

$$Y_{i,f} = \alpha^L + \rho^L S_{i,f} + \gamma \mathbf{A}_{i,f} + e^L_{i,f}$$

where $\mathbf{A}_{i,f}$ contains IQ, charisma, perseverance, creativity, etc.

$$\gamma \mathbf{A}_{i,f} = \gamma_1 A_{1i,f} + \gamma_2 A_{2i,f} + ... + \gamma_Q A_{Qi,f}$$

- We cannot run the LR
- We can run the SR

Twins

The Short Regression for Twins

- The SR for a MZ twin i of family f can be written as:

$$Y_{i,f} = \alpha^S + \rho^S S_{i,f} + e^S_{i,f}$$

- Hence, ρ^S is given by:

$$\rho^S = \frac{Cov(Y_{i,f}, S_{i,f})}{Var(S_{i,f})} = \rho^L + \sum_{q=1}^{Q} \gamma_q \frac{Cov(A_{qi,f}, S_{i,f})}{Var(S_{i,f})}$$

- Running the SR leaves us with ability bias

Twins

The SR and LR with demographic controls for twins

- The LR with demographic controls:

$$Y_{i,f} = \alpha^L + \rho^L S_{i,f} + \gamma \mathbf{A}_{i,f} + \delta^L \mathbf{X}_{i,f} + e^L_{i,f}$$

  where $\mathbf{X}_{i,f} = \{age_{i,f}, age^2_{i,f}, male_{i,f}, white_{i,f}\}$.

- The SR with demographic controls:

$$Y_{i,f} = \alpha^S + \rho^S S_{i,f} + \delta^S \mathbf{X}_{i,f} + e^S_{i,f}$$

- Hence, ρ^S is given by:

$$\rho^S = \frac{Cov(Y_{i,f}, \widetilde{S}_{i,f})}{Var(\widetilde{S}_{i,f})} = \rho^L + \sum_{q=1}^{Q} \gamma_q \frac{Cov(A_{qi,f}, \widetilde{S}_{i,f})}{Var(\widetilde{S}_{i,f})}$$

  where

$$S_{i,f} = \pi_0 + \boldsymbol{\pi} \mathbf{X}_{i,f} + \widetilde{S}_{i,f}$$

Twins

Table 2: Regressions of log wage for identical twins

	OLS in levels
Self-reported education	0.084
	(0.014)
Age	0.088
	(0.019)
Age squared / 100	−0.087
	(0.023)
Male	0.204
	(0.063)
White	−0.410
	(0.127)
N	298

Note: Regression includes an intercept term. Numbers in parentheses are standard errors.

Twins

The Long Regression in differences

- We can write the LR for each twin *i* (1 and 2) in each family *f*

$$Y_{1,f} = \alpha^L + \rho^L S_{1,f} + \gamma \mathbf{A}_{1,f} + \delta^L \mathbf{X}_f + e^L_{1,f}$$

$$Y_{2,f} = \alpha^L + \rho^L S_{2,f} + \gamma \mathbf{A}_{2,f} + \delta^L \mathbf{X}_f + e^L_{2,f}$$

- We can subtract the second from the first equation

$$Y_{1,f} - Y_{2,f} = \rho^L (S_{1,f} - S_{2,f}) + \gamma (\mathbf{A}_{1,f} - \mathbf{A}_{2,f}) + (e^L_{1,f} - e^L_{2,f})$$

- This is the LR in differences.

Twins

- The LR in differences

$$Y_{1,f} - Y_{2,f} = \rho^L (S_{1,f} - S_{2,f}) + \gamma(\mathbf{A}_{1,f} - \mathbf{A}_{2,f}) + (e^L_{1,f} - e^L_{2,f})$$

- Assuming that $\mathbf{A}_{1,f} = \mathbf{A}_{2,f}$, the LR in differences becomes

$$\underbrace{Y_{1,f} - Y_{2,f}}_{\Delta Y_f} = \rho^L \underbrace{(S_{1,f} - S_{2,f})}_{\Delta S_f} + \underbrace{(e^L_{1,f} - e^L_{2,f})}_{\Delta e^L}$$

- Hence, ρ^L is given by:

$$\rho^L = \frac{Cov(\Delta Y_f, \Delta S_f)}{Var(\Delta S_f)}$$

- ρ^L can be recovered without information on $\mathbf{A}_{i,f}$

Twins

- Assumption: ability is common to a pair of twin siblings $\mathbf{A}_{1,f} = \mathbf{A}_{2,f}$
- **Monozygotic (MZ) twins**:
  - same genetic ability (by construction, perhaps...)
  - same acquired ability (by assumption, really?)

Twins

Scatter plot of ΔY_f against ΔS_f

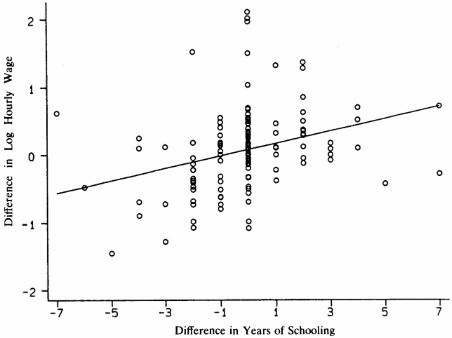


FIGURE 1. INTRAPAIR RETURNS TO SCHOOLING, IDENTICAL TWINS

Ashenfelter and Krueger (1994)

Twins

Table 3: Regressions of log wage for identical twins[4]

	OLS *in levels*	*OLS* *in differences*
Self-reported education	***0.084***	***0.092***
	(0.014)	*(0.024)*
Age	*0.088*	–
	(0.019)	
Age squared / 100	*−0.087*	–
	(0.023)	
Male	*0.204*	–
	(0.063)	
White	*−0.410*	–
	(0.127)	
N	*298*	*149*

Twins

- Twins are similar in many ways, including their schooling
- If most twins have the same schooling, then a fair number of nonzero differences in reported schooling may reflect mistaken reports by at least one of them...

Twins

- Let's first run the short regression

$$Y_{i,f} = \alpha^S + \rho^S S^*_{i,f} + e^S_{i,f}$$

  where $S^*_{i,f}$ is unobserved.

- What we observe is $S_{i,f}$:

$$S_{i,f} = S^*_{i,f} + m_{i,f}$$

  where $m_{i,f}$ is classical measurement error

$$\mathbb{E}\left[m_{i,f}\right] = 0$$

$$Cov(m_{i,f}, S^*_{i,f}) = 0$$

$$Cov(m_{i,f}, e^S_{i,f}) = 0$$

Twins

- The return to schooling in the SR is

$$\rho^S = \frac{Cov(Y_{i,f}, S^*_{i,f})}{Var(S^*_{i.f})}$$

- The return to schooling in the SR with measurement error is

$$\rho^{S'} = \frac{Cov(Y_{i,f}, S_{i,f})}{Var(S_{i,f})} = \frac{Cov(\alpha^S + \rho^S S^*_{i,f} + e^S_{i,f}, S^*_{i,f} + m_{i,f})}{Var(S^*_{i,f} + m_{i,f})} =$$

$$= \rho^S \frac{Var(S^*_{i,f})}{Var(S^*_{i,f}) + Var(m_{i,f})} = \rho^S \lambda$$

- Can we fix the attenuation bias?
- Can we find λ?

Twins

- Suppose that we have two measures of schooling for each twin

$$\begin{aligned} S_{1,f} &= S^*_{1,f} + m_{1,f} && \text{report of twin 1 on her own education} \\ S^1_{2,f} &= S^*_{1,f} + m^1_{2,f} && \text{report of twin 2 on her twin education} \\ S_{2,f} &= S^*_{2,f} + m_{2,f} && \text{report of twin 2 on her own education} \\ S^2_{1,f} &= S^*_{2,f} + m^2_{1,f} && \text{report of twin 1 on her twin education} \end{aligned}$$

with

$$\begin{aligned} Cov(S^*_{1,f}, m_{1,f}) &= 0 \\ Cov(S^*_{1,f}, m^1_{2,f}) &= 0 \\ Cov(S^*_{2,f}, m_{2,f}) &= 0 \\ Cov(S^*_{2,f}, m^2_{1,f}) &= 0 \\ Cov(m_{1,f}, m^1_{2,f}) &= 0 \\ Cov(m_{2,f}, m^2_{1,f}) &= 0 \end{aligned}$$

Twins

- What is the $Cov(S_{i,f}, S^i_{j,f})$ for $i \neq j$, where $i, j = \{1, 2\}$?

  $Cov(S_{i,f}, S^i_{j,f}) = Cov(S^*_{i,f} + m_{i,f}, S^*_{i,f} + m^i_{j,f}) = Cov(S^*_{i,f}, S^*_{i,f}) = Var(S^*_{i,}$

- Why?

$$Cov(m_{i,f}, S^*_{i,f}) = 0$$

$$Cov(S^*_{i,f}, m^i_{j,f}) = 0$$

$$Cov(m_{i,f}, m^i_{j,f}) = 0$$

- Once we have $Var(S^*_{i,f})$, we can compute λ:

$$\frac{Var(S^*_{i,f})}{Var(S_{i,f})} = \frac{Var(S^*_{i,f})}{Var(S^*_{i,f}) + Var(m_{i,f})} = \lambda$$

- Ashenfelter and Krueger (1994) estimate λ to be 0.9

Twins

- Short regression:
- Ability bias
    - Attenuation bias
- Long regression:
    - Ability bias decreases
    - Attenuation bias increases
- Long regression in differences:
    - Ability bias decreases
    - Attenuation bias increases

Twins

- The return to schooling in the LR in differences with measurement error is

$$\rho^{L'} = \frac{Cov(\Delta Y_f, \Delta S_f)}{Var(\Delta S_f)} = \rho^L \frac{Var(\Delta S_f^*)}{Var(\Delta S_f^*) + Var(\Delta m_f)}$$

where

$$\lambda' = \frac{Var(\Delta S_f^*)}{Var(\Delta S_f^*) + Var(\Delta m_f)} < \frac{Var(S_{i,f}^*)}{Var(S_{i,f}^*) + Var(m_{i,f})} = \lambda$$

- Again, we can find λ'. If we recover λ', we can fix both: ability bias and attenuation bias.

Twins

- If schooling differences between twins are not random, these estimates of the economic return to schooling are biased
- **Economic rational** for non-random differences in schooling between twins?
  - If parents invest more in the “more talented/able” twin, they reinforce ability differences, the estimates of the return to schooling will be biased upward
  - If parents invest more in the “less talented/able” twin, they compensate/equalize ability differences, the estimates of the return to schooling will be biased downward.

Twins

- **Main advantage:**
  No need to “observe the unobservable” set of abilities, skills, etc. of an individual

Twins

- How about internal validity?
- **Measurement error** in schooling gets exacerbated when taking differences
  - The *Nature* vs. *Nurture* debate is an old one
  - Twins develop ability differences over time: **epigenetics**
  - **Parental behavior/investment** towards/on twins differs across families

Twins

- How about external validity?
  - Twins and singletons are **different**: they have different birth outcomes than singletons (e.g., lower birth weight), and this affects their schooling and earnings
  - Nowadays: the majority of twins are due to **Assisted Reproductive Technology** (families who use ART are self-selected)
  - Still many things can be learnt from studying twins.

Twins - Summary

- **Regression with identical twins** to estimate the returns to schooling
- **Assumption** Differences in schooling between identical twins are random
- Main limitations of this approach:
  - Internal validity? Identical twins develop differences over time.
  - External validity? Twins are different than singletons

Eye-Vee

- IV
- 2SLS
- Is Quarter of Birth a valid instrument?
- The return to schooling for whom?

Generating “random” variation

- Use variation in schooling unrelated to ability
- Using “birthday” as an instrumental variable (IV)

Generating “random” variation

Angrist and Krueger (1991): quarter of birth (QOB) instrument

- *School entry age rule*
- *Compulsory schooling law*

Generating “random” variation

School entry age rule

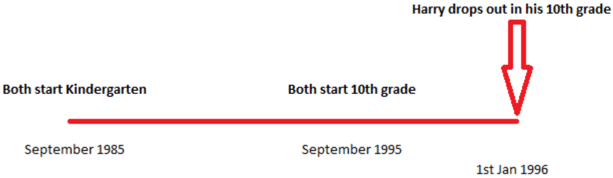
- In most states in the US, children enter Kindergarten in the year they turn 5, whether or not they’ve had a 5th birthday by the time school starts in early September
- Consider two individuals: Harry and Ron
  - Harry’s DOB is 1st January 1980
  - Ron’s DOB is 1st December 1980

Generating “random” variation

Compulsory schooling law

- Some states in the US were allowing to drop out from high school after individuals turned 16, without even finishing the school year
- Suppose both Harry and Ron want to leave school as soon as they are allowed
  - Harry turned 16 ten years later, early in his 10th grade
  - Ron turned 16 eleven years later, after finishing 10th grade and starting 11th grade
- What is the implication of the compulsory schooling law?
  - Ron was forced by accident of birth to complete one more grade than Harry

Generating “random” variation



Harry drops out in his 10th grade

Both start Kindergarten

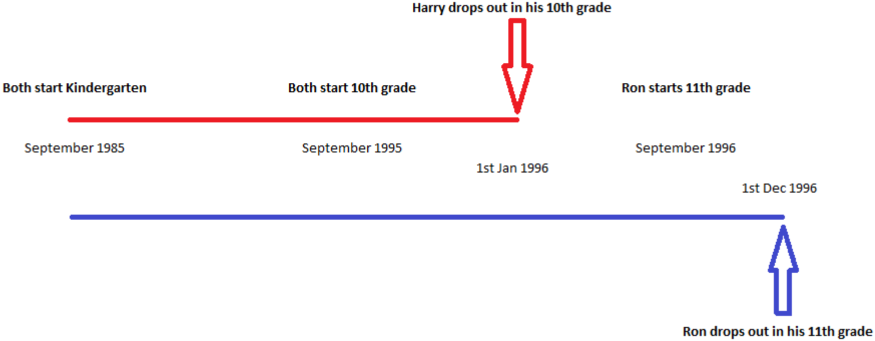
Both start 10th grade

September 1985

September 1995

1st Jan 1996

Generating "random" variation



Generating “random” variation

- School entry age + compulsory schooling laws
  - **Quarter of birth** affects educational attainment
  - Ron acquires 1 additional year of education more than Harry
  - Individuals born in Q4 will have more education than those born in Q1
  - Individuals born in Q4 will have more education than those born in other quarters
  - Use QOB as an IV

Using IV

Data

- 1980 **US Census**
- 329,509 men born 1930-1939
- Observed in their 40s
- Information on:
  - year of birth
  - quarter of birth
  - years of schooling
  - earnings in 1979

Using IV

- The long regression for individual i can be written as

$$Y_i = \alpha^L + \rho^L S_i + \gamma \mathbf{A}_i + e_i^L$$

- Let's define the following Quarter of Birth binary variable

$$Q4_i = \begin{cases} 1 & \text{if the individual is born in Quarter 4} \\ 0 & \text{if the individual is born in Quarter 1, 2 or 3} \end{cases}$$

- Suppose that $Q4_i$ is a **valid instrument**:
  1. $Q4_i$ is as good as randomly assigned $\Rightarrow \mathbb{E}[\mathbf{A}_i|Q4_i] = \mathbb{E}[\mathbf{A}_i]$
  2. $Q4_i$ affects years of schooling S_i: $\mathbb{E}[S_i|Q4_i = 1] \neq \mathbb{E}[S_i|Q4_i = 0]$
  3. $Q4_i$ affects earnings Y_i only through S_i: $Q4_i$ does not enter the long regression

Using IV

- The conditional expectations of Y_i with $Q4_i = 1$ and $Q4_i = 0$

$$\mathbb{E}[Y_i|Q4_i = 1] = \alpha^L + \rho^L \mathbb{E}[S_i|Q4_i = 1] + \gamma \mathbb{E}[\mathbf{A}_i|Q4_i = 1] + \mathbb{E}[e_i^L|Q4_i = 1]$$

$$\mathbb{E}[Y_i|Q4_i = 0] = \alpha^L + \rho^L \mathbb{E}[S_i|Q4_i = 0] + \gamma \mathbb{E}[\mathbf{A}_i|Q4_i = 0] + \mathbb{E}[e_i^L|Q4_i = 0]$$

- Because of (1) and (3)

$$\mathbb{E}[\mathbf{A}_i|Q4_i = 1] = \mathbb{E}[\mathbf{A}_i|Q4_i = 0] = \mathbb{E}[\mathbf{A}_i]$$

$$\mathbb{E}[e_i^L|Q4_i = 1] = \mathbb{E}[e_i^L|Q4_i = 0] = \mathbb{E}[e_i^L] = 0$$

- Hence

$$\mathbb{E}[Y_i|Q4_i = 1] - \mathbb{E}[Y_i|Q4_i = 0] = \rho^L \left(\mathbb{E}[S_i|Q4_i = 1] - \mathbb{E}[S_i|Q4_i = 0]\right)$$

Using IV

- Finally, the **IV estimand** of ρ^L is given by

$$\frac{\mathbb{E}[Y_i|Q4_i = 1] - \mathbb{E}[Y_i|Q4_i = 0]}{\mathbb{E}[S_i|Q4_i = 1] - \mathbb{E}[S_i|Q4_i = 0]}$$

  where $\mathbb{E}[S_i|Q4_i = 1] - \mathbb{E}[S_i|Q4_i = 0] \neq 0$ because of (2).[5]

- In general the IV estimand is written as

$$\frac{Cov(Y_i, Q4_i)}{Cov(S_i, Q4_i)}$$

[5] Note: The IV estimand with a binary instrument is called the "Wald estimand". Its sample analogue is called the "Wald estimator".

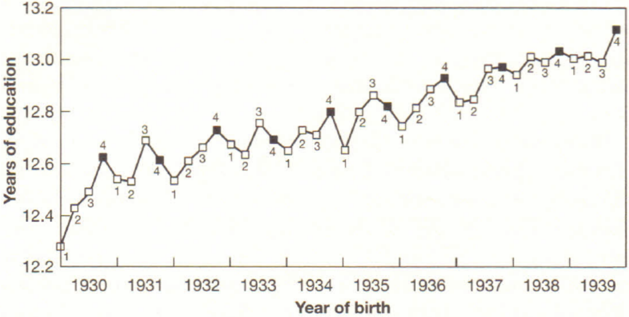
Using IV

IV mechanics

- Two pieces:
  1. **FS** (First Stage): The Effect of $Q4_i$ on S_i
  2. **RF** (Reduced Form): The Effect of $Q4_i$ on Y_i

Using IV

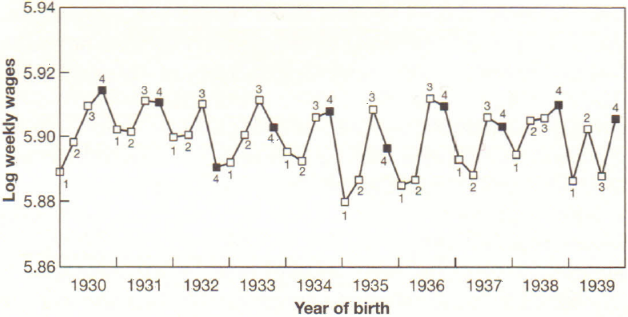
FIGURE 6.1
The quarter of birth first stage



Notes: This figure plots average schooling by quarter of birth for men born in 1930–1939 in the 1980 U.S. Census. Quarters are labeled 1–4, and symbols for the fourth quarter are filled in.

Using IV

FIGURE 6.2
The quarter of birth reduced form



Notes: This figure plots average log weekly wages by quarter of birth for men born in 1930–1939 in the 1980 U.S. Census. Quarters are labeled 1–4, and symbols for the fourth quarter are filled in.

Using IV

Table 1: Average Wages and Schooling by Quarter of Birth and IV

	Q1,Q2,Q3	Q4	Difference
Log weekly wage	5.898	5.905	0.007
			(0.003)
Years of education	12.747	12.839	0.092
			(0.013)
IV estimate of ρ^L			**0.074**
			(0.028)

Note: Standard errors in parentheses. Source: Angrist and Pischke (2015).

Using IV

2SLS mechanics (in words)

- **First stage:** Regression of S on $Q4$
  - Why? We want to use the part of schooling S that is unrelated to ability $\mathbf{A}$
  - If the instrument is valid, that's what we get by running the regression S on $Q4$
  - This part is $\widehat{S}$: the variation in schooling due to variation in QOB
- **Second stage:** Regression of Y on $\widehat{S}$

Using IV

2SLS mechanics (in equations)

- **First stage:**

$$S_i = \delta + \beta Q4_i + v_i$$

$$S_i = \underbrace{\delta + \beta Q4_i}_{\widehat{S}_i} + v_i$$

- **Second stage:**

$$Y_i = \alpha + \rho^L \widehat{S}_i + u_i$$

The 2SLS estimand of ρ^L is

$$\frac{Cov(Y_i, \widehat{S}_i)}{Var(\widehat{S}_i)}$$

Using IV

Why 2SLS?

- The 2SLS estimand is flexible in two different ways
- We can use multiple instruments: 3 quarter of birth dummies.
- We can add controls: 9 year of birth dummies.

Using IV

2SLS with multiple instruments and controls

- The first stage regression is

$$S_i = \delta + \sum_{K=2}^{4} \beta_K QK_i + \sum_{T=1931}^{1939} \phi_T YEART_i + v_i$$

- The structural equation is

$$Y_i = \alpha + \rho^L S_i + \sum_{T=1931}^{1939} \tau_T YEART_i + u_i$$

Using IV

- After running the first stage we get

$$\widehat{S}_i = \delta + \sum_{K=2}^{4} \beta_K QK_i + \sum_{T=1931}^{1939} \phi_T YEART_i$$

- Assuming that the instrument is valid, $\widehat{S}_i$ is the part of S_i that is uncorrelated with $\mathbf{A}_i$
- Then we can run the second stage

$$Y_i = \alpha + \rho^L \widehat{S}_i + \sum_{T=1931}^{1939} \tau_T YEART_i + u_i$$

Using IV

- The **2SLS estimand** is

$$\frac{Cov(Y_i, \widetilde{\widehat{S}}_i)}{Var(\widetilde{\widehat{S}}_i)}$$

where

$$\widehat{S}_i = \pi_0 + \sum_{T=1931}^{1939} \pi_T YEART_i + \widetilde{\widehat{S}}_i$$

- The 2SLS estimator is the sample analogue of the 2SLS estimand

Using IV

Table 2: Estimates of the Returns to Schooling

	OLS	**2SLS**	OLS	**2SLS**	**2SLS**
Years of education	0.071	0.074	0.071	0.075	0.105
	(0.0004)	(0.028)	(0.0004)	(0.028)	(0.020)
Year of birth controls	No	No	Yes	Yes	Yes
Instruments	None	Q4	None	Q4	Three Q Dummies
First-stage F-statistic	–	48	–	47	33

Note: Standard errors in parentheses. First-stage F-statistic: F-statistic for the joint significance of the instruments in the corresponding first-stage regression.

Source: Angrist and Pischke (2015).

Is QOB a valid instrument?

1. Is it as good as randomly assigned?
2. Is it relevant?
3. Does it satisfy the exclusion restriction?

Is QOB a valid instrument?

- Quarter of birth is related to mother characteristics
  - Buckles & Hungerman (2013), Clarke, Oreffice & Quintana-Domeque (2016): maternal schooling peaks for mothers who give birth in Q2
  - However, the peaks in schooling and earnings in A&K are in Q4
  - If anything, IV estimate is downward biased

Is QOB a relevant instrument?

- Quarter of birth explains years of education
  - The power of QOB instruments has been questioned
  - Rule of thumb for instrument relevance: First stage F-statistic $F > 10$
  - Bias from weak instrument(s) is unlikely to be a problem

Does QOB satisfy the exclusion restriction?

While the exclusion restriction is not testable, we can investigate...

- Does school-starting age matters by itself (age rank in class)?
- A reason other than compulsory schooling behind the effects of quarter of birth on earnings?

Does QOB satisfy the exclusion restriction?

Does school-starting age matters by itself (age rank in class)?

- Harry and Ron start Kindergarten with *different* ages:
  - Harry is 5 years and 8 months
    - Ron is 4 years and 9 months
  - The youngest children in a first-grade class tend to be at a disadvantage, while children who are a litter older than their classmates tend to do better
  - However, in A&K younger entrants do better in schooling and earnings
  - If anything, IV estimate is downward biased

Does QOB satisfy the exclusion restriction?

- A reason other than compulsory schooling behind the effects of quarter of birth on earnings?
- If so, we would expect quarter of birth to be related to earnings for graduates
- **Graduates** are not affected by compulsory schooling laws (*placebo group*)
- If QOB affects the earnings of graduates... violation of the exclusion restriction!
- A&K find no effects of QOB on earnings among graduates

The Return to Schooling for whom?

- Homogeneous causal effects: return to schooling for any individual
- Heterogeneous causal effects?

The Return to Schooling for whom?

- Heterogenous world: **The LATE world**
  - LATE assumptions: (1), (2), (3) + monotonicity
  - Monotonicity: no individuals drop out because born in Q4; no individuals stay in school because born in Q1, Q2 or Q3
  - 2SLS/IV recovers the return to schooling for compliers
  - Compliers: the group of individuals for whom the instrument changes the schooling decision
  - Compliers: those who are on the margin of dropping out
  - Compliers: those who do not drop out because they were born in Q4

The Return to Schooling for whom?

- Benefits?
    - Compulsory schooling laws are effective in compelling some students to attend school
    - Students who are compelled to attend school longer by compulsory schooling laws earn higher wages as a result of their extra schooling
- Costs?
    - Students who are compelled to attend school may interfere with the learning of other students.

Summary

- The CIA can provide plausible identification of causal effects. But the obvious controls (in this context) are hard to observe and include.
- Omitting relevant control variables causes omitted variable bias in the parameter of interest.
- Adding poorly-measured proxies helps OVB but can add attenuation bias.
- Adding controls which are themselves outcomes can add compositional bias.

Summary

- Twin studies rely on siblings as controls; a differencing strategy can help eliminate common genetic factors in pairs but attenuation bias can be made worse.
- Using second (independent) measurements can help by allowing use to estimate the reliability factor and correct.

Summary

- IV is a piece of magic - but it relies on an untestable assumption.

> *"IV is Economics' unwanted gift to Statistics"*
>
> *[source. S Mavroedis]*