

Ec533: Labour Economics for Research Students

The returns to schooling, ability bias, and regression

Jörn-Steffen Pischke

LSE

October 3, 2009

Ability bias and the returns to schooling

We would like to run the *long regression*

$$y_i = \alpha + \rho S_i + \gamma A_i + e_i$$

where y_i is log earnings, S_i is schooling and A_i is ability. If we don't have a measure of ability we can only run the *short regression*

$$y_i = \alpha_s + \rho_s S_i + e_i^s.$$

What do we get?

The omitted variables bias formula

The relationship between the long and short regression coefficients is given by the *omitted variables bias (OVB)* formula

$$\rho_s = \frac{\text{Cov}(y_i, S_i)}{\text{Var}(S_i)} = \rho + \gamma \delta_{AS}$$

where

$$\delta_{AS} = \frac{\text{Cov}(A_i, S_i)}{\text{Var}(S_i)}$$

is the regression coefficient from a regression of A_i (the omitted variable) on S_i (the included variable).

The OVB formula is a mechanical relationship between two regressions: it holds regardless of the causal interpretation of any of the coefficients.

Griliches (1977) regressions

- The conventional wisdom is $Cov(A_i, S_i) > 0$, so returns to schooling estimates will be biased up.
- Short regression estimates using the NLS

$$y_i = \text{const} + \underset{(0.003)}{0.068} S_i + \text{experience}$$

- Long regression estimates

$$y_i = \text{const} + \underset{(0.003)}{0.059} S_i + \underset{(0.0005)}{0.0028} IQ_i + \text{experience}$$

- The results are consistent with the conventional wisdom.

Classical measurement error

- Measurement error leads to bias. Many economic variables are mismeasured.
- Look at a generic example and start with a simple bivariate regression

$$y_i = \alpha + \beta x_i + e_i.$$

We don't observe x_i but \tilde{x}_i

$$\tilde{x}_i = x_i + w_i$$

where $cov(x_i, w_i) = 0$ and $cov(e_i, w_i) = 0$. This is called the *classical measurement error* model.

Attenuation from classical measurement error

The bivariate regression coefficient we estimate is

$$\begin{aligned}\hat{\beta} &= \frac{\text{cov}(y_i, \tilde{x}_i)}{\text{var}(\tilde{x}_i)} \\ &= \frac{\text{cov}(\alpha + \beta x_i + e_i, x_i + w_i)}{\text{var}(x_i + w_i)} \\ &= \beta \frac{\text{var}(x_i)}{\text{var}(x_i + w_i)} = \beta \lambda.\end{aligned}$$

We see that $\hat{\beta}$ is biased towards zero by an attenuation factor

$$\lambda = \frac{\text{var}(x_i)}{\text{var}(x_i + w_i)}$$

which is the variance in the “signal” divided by the variance in the “signal plus noise.”

Measurement error in the returns to schooling

- Think of y_i as log earnings, and x_i as schooling. Ignore age or experience for the moment.
- Ashenfelter and Krueger (1994) find $\lambda = 0.9$ for schooling.
- This means if the true return to schooling is 0.1, we would expect an estimate of 0.09.

Measurement error with two regressors

- We typically use regression as a tool to control for covariates, (e.g. for ability as in the Griliches example). Hence we are interested in a regression with at least two regressors.
- Consider again the generic case first

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i.$$

and only x_{1i} is subject to classical measurement error, i.e.

$cov(x_{1i}, w_i) = cov(x_{2i}, w_i) = 0$. We will give specific roles to x_{1i} and x_{2i} later.

- Then it can be shown that

$$\begin{aligned}\hat{\beta}_1 &= \beta_1 \lambda' \\ \lambda' &= \frac{\lambda - R_{12}^2}{1 - R_{12}^2}\end{aligned}$$

where λ is the bivariate attenuation factor, and R_{12}^2 is the R^2 from the population regression of \tilde{x}_{1i} on x_{2i} .

Comparing the short and long regression

- The short regression (on just \tilde{x}_{1i}) coefficient is

$$\hat{\beta}_{1,short} = \lambda\beta_1 + \beta_2\delta_{x_2\tilde{x}_1} = \lambda(\beta_1 + \beta_2\delta_{x_2x_1})$$

where the estimate of β_1 is biased both because of attenuation due to measurement error, and because of omitted variables bias (the part $\beta_2\delta_{x_2\tilde{x}_1}$ where $\delta_{x_2\tilde{x}_1}$ is the coefficient from a regression of x_{2i} on \tilde{x}_{1i}).

- The coefficient from the long regression is

$$\hat{\beta}_{1,long} = \lambda'\beta_1$$

Note that

$$\lambda' < \lambda$$

but

$$\hat{\beta}_{1,short} \leq \hat{\beta}_{1,long}.$$

Comparing the short and long regression

- Notice that it is more difficult to compare the bias from the short regression and the long regression now.
- $\lambda' < \lambda$ implies that the attenuation bias goes up when another regressor is entered which is correlated with \tilde{x}_{1j} .
- There is less attenuation in the short regression but there is also OVB now. Not clear what the net effect is.

- What about the coefficient β_2 ? Even when there is no measurement error in x_{2i} , the estimate of β_2 will be biased:

$$\hat{\beta}_2 = \beta_1 \delta_{x_1 x_2} \frac{1 - \lambda}{1 - R_{12}^2} + \beta_2.$$

- Note that the bias will be larger the larger
 - the measurement error
 - the correlation between x_{1i} and x_{2i}
- The intuition is that
 - β_1 is attenuated, and hence does not reflect the full effect of x_{1i}
 - β_2 will capture part of the effect of x_{1i} through the correlation with x_{2i}

Measurement error in the returns to schooling controlling for ability

We run the regression

$$y_i = \alpha + \rho S_i + \gamma A_i + e_i$$

where S_i is schooling and A_i is ability. Suppose we only have a mismeasured version of schooling, \tilde{S}_i (so S_i takes on the role of x_{1i} before, and A_i takes on the role of x_{2i}). Then the short regression will give

$$\hat{\rho}_{short} = \lambda\rho + \gamma\delta_{A\tilde{S}}$$

and the long regression

$$\hat{\rho}_{long} = \lambda'\rho$$

If ability bias is upwards ($\delta_{A\tilde{S}} > 0$) it is not possible to say a priori which estimate will be closer to ρ .

Putting some numbers on the Griliches example

Pick some numbers for the regression

$$y_i = 0.1S_i + 0.01A_i + e_i$$

and set

$$\begin{aligned}\lambda &= 0.9 \\ \sigma_{\tilde{S}} &= 3, \sigma_A = 15, \sigma_{AS} = 22.5.\end{aligned}$$

Then

$$\delta_{A\tilde{S}} = \frac{\sigma_{AS}}{\sigma_{\tilde{S}}^2} = \frac{22.5}{9} = 2.5$$

and

$$\hat{\rho}_{short} = \lambda\rho + \gamma\delta_{A\tilde{S}} = 0.9 \times 0.1 + 0.01 \times 2.5 = 0.115$$

What about the long regression?

We first need

$$\lambda' = \frac{\lambda - R_{\tilde{S}A}^2}{1 - R_{\tilde{S}A}^2}$$

which is

$$R_{\tilde{S}A}^2 = \left(\frac{\sigma_{AS}}{\sigma_{\tilde{S}}\sigma_A} \right)^2 = \left(\frac{22.5}{45} \right)^2 = 0.25$$
$$\lambda' = \frac{0.9 - 0.25}{1 - 0.25} = 0.867.$$

Then the long regression coefficient is

$$\hat{\rho}_{long} = \lambda' \rho = 0.867 \times 0.1 = 0.087$$

so the short regression coefficient is too large and the long regression coefficient is too small.

Measurement error in ability

Now suppose years of schooling is measured perfectly but we only have mismeasured ability \tilde{A}_i (so S_i takes on the role of x_{2i} before, and A_i takes on the role of x_{1i}). Then

$$\hat{\rho} = \gamma \delta_{AS} \frac{1 - \lambda}{1 - R_{AS}^2} + \rho.$$

If ability bias is upwards ($\delta_{AS} > 0$) then the returns to schooling will be biased up but by less than in the short regression. Controlling for \tilde{A}_i is better than controlling for nothing, but not as good as controlling for true ability A_i .

Instrumental variables solve the measurement error problem

- Suppose you have an instrument z_i , correlated with the signal x_{1i} and uncorrelated with the error w_i .
- In the bivariate regression you get

$$\hat{\beta}_{IV} = \frac{\text{cov}(y_i, z_i)}{\text{cov}(\tilde{x}_i, z_i)} = \frac{\text{cov}(\alpha + \beta x_i + e_i, z_i)}{\text{cov}(x_i, z_i)} = \frac{\beta \text{cov}(x_i, z_i)}{\text{cov}(x_i, z_i)} = \beta$$

- In the multivariate regression you get for similar reasons:

$$\begin{aligned}\hat{\beta}_{1,IV} &= \beta_1 \\ \hat{\beta}_{2,IV} &= \beta_2\end{aligned}$$

(notice that only x_{1i} is instrumented)

More Griliches (1977) regressions

- Recall the Griliches long regression estimates

$$y_i = \text{const} + \underset{(0.003)}{0.059} S_i + \underset{(0.0005)}{0.0028} \text{IQ}_i + \text{experience}$$

- Instrumenting IQ with results from the Knowledge of the World of Work test he gets

$$y_i = \text{const} + \underset{(0.004)}{0.052} S_i + \underset{(0.0009)}{0.0051} \text{IQ}_i + \text{experience}$$

- This is still consistent with two stories:
 - there is upward ability bias in the bivariate return to schooling and measurement error in IQ but not schooling (in this case the estimate of 0.052 is correct)
 - there is measurement error in both schooling and ability (in this case the true coefficient could be anything).

What to control for?

In the quest for identifying causal effects, which variables belong on the right hand side of a regression equation?

- Yes: Variables determining the treatment and correlated with the outcome (e.g. ability).
 - in general these variables will be fixed characteristics or pre-determined by the time of treatment (e.g. schooling)
- Yes: Variables uncorrelated with the treatment but correlated with the outcome
 - these variables may help reducing standard errors
- No: Variables which are outcomes of the treatment itself. These are *bad controls*.

Some researchers regressing earnings on schooling (and experience) include controls for occupation. Does this make sense?

- Clearly we can think of schooling affecting the access to higher level occupations, e.g. you need a Ph.D. to become a college professor. This gives rise to a two equation system

$$\begin{aligned}y_i &= \alpha + \rho S_i + \gamma O_i + e_i \\ O_i &= \lambda_0 + \lambda_1 S_i + u_i\end{aligned}$$

You could think about these as a simultaneous equations system. Occupation O_i is an endogenous variable. As a result, you could not necessarily estimate the first equation by OLS.

- Occupation is a *bad control*.

Example 1 of bad control

S is schooling ($S = 0$ and $S = 1$)

O is occupation (B = blue collar, W = white collar)

W is wage

O and W are potential outcomes indexed against schooling: O_0 is occupation with low schooling, O_1 is occupation with high schooling, etc.

| | <i>occupation</i> | | <i>wage</i> | | <i>Observed data</i> | |
|--------|-------------------|-------|-------------|-------|----------------------|----------|
| | O_0 | O_1 | W_0 | W_1 | $S = 0$ | $S = 1$ |
| Type 1 | B | B | 600 | 600 | (B, 625) | (B, 600) |
| Type 2 | B | W | 650 | 700 | | (W, 700) |
| Type 3 | W | W | 700 | 700 | (W, 700) | |

Example 2 of bad control

| | <i>occupation</i> | | <i>wage</i> | | <i>Observed data</i> | |
|--------|-------------------|-------|-------------|-------|----------------------|----------|
| | O_0 | O_1 | W_0 | W_1 | $S = 0$ | $S = 1$ |
| Type 1 | B | B | 600 | 625 | (B, 625) | (B, 625) |
| Type 2 | B | W | 650 | 700 | (B, 625) | (W, 725) |
| Type 3 | W | W | 725 | 750 | (W, 725) | (W, 725) |

Sometimes we control for a variable in the best of intentions. Suppose our regression of schooling on earnings

$$y_i = \alpha + \rho S_i + \gamma A_i + e_i$$

has a causal interpretation conditional on ability.

Instead of ability we only have a test score taken at age 18, call it A_{1i} for late ability. The problem is that schooling will already have influenced the late ability (some students will have dropped out by 18). Suppose

$$A_{1i} = \pi_0 + \pi_1 S_i + \pi_2 A_i$$

i.e. that age 18 test scores are influenced by both schooling and true ability.

What do we get with proxy control?

Substituting for A_i in our regression above we get

$$y_i = \left(\alpha - \gamma \frac{\pi_0}{\pi_2} \right) + \left(\rho - \gamma \frac{\pi_1}{\pi_2} \right) S_i + \frac{\gamma}{\pi_2} A_{li} + e_i.$$

If

$$\begin{aligned} \rho &> 0 \\ \gamma &> 0 \\ \pi_1 &> 0, \pi_2 > 0 \end{aligned}$$

then

$$\left(\rho - \gamma \frac{\pi_1}{\pi_2} \right) < \rho$$

and we will estimate a return to schooling that is too small.