Omitted Variables Bias

Recall the formula for β:

$$\beta = (X'X)^{-1}(X'Y) = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{Cov(x,y)}{Var(x)}$$ (if x and y are de-meaned)

- This is a very useful way of understanding the formula for β. Note that, since the variance is always positive, the sign of β will be the same as the sign of the covariance (you can think of this as basically the correlation) between x and y, since $Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)} * \sqrt{Var(y)}}$.

Suppose that the true relationship between education and wages is as follows:

$w_i = \alpha + \beta ed_i + \gamma Z_i + \nu ability_i + \varepsilon_i$,
where Z is a vector of control variables such as age, race, gender, etc.

To simplify, let's forget about the control variables. So let us call the "true" relationship:

(1) $w_i = \alpha + \beta_1 ed_i + \nu ability_i + \varepsilon_i$

But suppose we cannot measure ability (or we don't know it should be in the regression), so it is left out and we estimate:

(2) $w_i = \alpha + \beta_2 ed_i + \varepsilon_i$

How will that affect our estimate of β? Recall the formula for β:

(3) $\beta = \frac{Cov(w,ed)}{Var(ed)}$
but note that, even though we have assumed the wrong functional form for w, it will be given by the true relationship, so we must plug in the true equation:

$$\beta_2 = \frac{Cov(w,ed)}{Var(ed)} = \frac{Cov(\alpha + \beta_1 ed + \nu ability + \varepsilon, ed)}{Var(ed)}$$
$$= \frac{Cov(\alpha,ed)}{Var(ed)} + \beta_1 \frac{Cov(ed,ed)}{Var(ed)} + \nu \frac{Cov(ability,ed)}{Var(ed)} + \frac{Cov(\varepsilon,ed)}{Var(ed)}$$

Since α is a constant, it doesn't covary with anything, so the first term is zero. According to the assumptions of OLS $Cov(\varepsilon,ed) = 0$. Lastly, note that the cov(ed,ed)=var(ed).

So, canceling terms:

$$(4)\ \beta_2 = \beta_1 + \nu \frac{Cov(ability, ed)}{Var(ed)}$$

This implies that β_2 is equal to the true β, plus a bias term.

What is the expected sign of the bias? Well, recall that ν is the coefficient on ability in the wage regression, so from the same formula we had for β (equation 3):

$$(5)\ \nu = \frac{Cov(ability, w)}{Var(ability)}$$

so ν is positive if ability and wages are positively correlated (since the variance of any variable is always positive).

So, the sign of the bias here depends on the correlation between education and ability. There is some debate about this in the labor economics literature. You might assume that higher ability people get more education, maybe because they are more motivated. However, high ability people have higher wages for each level of education than low ability people, so they can make the same money for *less* education.

Let's assume that *Cov(ability, ed)* > 0. Then, since ν > 0 and *Var(ed)* > 0, the bias is positive and our estimated β_2 is larger than the true β.
What is happening here? Since ability matters for both wages and education, but ability is left out of the equation, β gets the credit for both education and ability. We will erroneously find that education matters more for wages because more educated people have higher ability.

When is there no bias?
This is clear from the omitted variables bias formula. If the omitted variable is uncorrelated with the outcome variable *or with the included independent variable* then the bias is zero. Why will there be no bias if ability is uncorrelated with education, even if ability is correlated with wages and is left out of the equation? Because, even though the regression equation is mis-specified, β will not get the credit for the ability variable because ability and education are uncorrelated.
How does this relate to the problem set?
Recall that the problem set asked about OVB in a regression of changes in infant mortality rates on changes in TSPs *at the county level*. Any omitted variable that would cause bias must be a *county level variable* that is correlated with *changes* in both of these variables.