

9.2 Using Proxy Variables for Unobserved Explanatory Variables

1. Consider the following model for wages

$$\log(wage) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{ability} + u$$

Our problem is that we know *ability* affects wages, but have no way to know how. Non-inclusion of ability biases our estimate on education and experience – which is what we care about.

2. **Using proxy variables:** A proxy variable is something that is related to the unobserved variable that we would like to control for in our analysis. For instance, let us assume that *IQ* scores are a proxy for the unobservable *ability*.
3. Thinking about ability and IQ: we need them to have some relationship and we do not need ability to be completely explained by IQ. This implies:

$$\text{ability} = \delta_0 + \delta_1 \text{IQ} + \delta_2 \text{somethingelse} + v$$

4. The requirement for a proxy variable for ability is that $\delta_1 \neq 0$. Then we use IQ instead of ability in our original regression model to obtain a **plug-in solution to the omitted variable problem**. However, there are two conditions under which this solution could lead to consistent estimators of β_1 and β_2 .
- (a) The error term should be uncorrelated with *educ*, *exper*, *ability* and *IQ*, i.e. the omitted variable as well as the proxy variable.
 - (b) The error term v should be uncorrelated with *educ*, *exper* and *IQ*

$$E(\text{abil}|\text{educ}, \text{exper}, \text{IQ}) = E(\text{abil}|IQ) = \delta_0 + \delta_1 IQ$$

5. **Note:** you could do the algebra on a general model with x_o as omitted variable and x_p as proxy variable. You can substitute predicted x_o i.e. $\hat{x}_o = \hat{\delta}_0 + \hat{\delta}_1 x_p$ into the main model and obtain the modified Gauss Markov assumptions in the augmented model.

6. Consider the returns to education i.e. the $\hat{\beta}$ on *educ* without proxying for ability. The estimated returns to education is 6.5% (example 9.3 and computer exercise). When JW includes IQ scores in this model, then the returns fall to 5.4%. This tells us that by correcting for ability bias, one could find the true returns to education.
7. **Multicollinearity vs Appropriate proxy variable:** It is possible that there could be potential multicollinearity between IQ and education. This could make our estimates imprecise.
 - (a) But including IQ leads to lower error variance and more precise estimates.
 - (b) Also, including IQ at the cost of some multi-collinearity might be required to get a more unbaised estimator.

9.2.1 Using Lagged Dependent Variables as Proxy Variables

1. **Idea:** If we have no proxy variables which we could think of, we could alternatively use lagged dependent variables with the idea that it would proxy a part of the unobservable omitted variable.
2. **Example:** consider the model on city crime rates:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{t-1} + u$$

where *unem* is the unemployment rate in the city, *expend* is the expenditure on law and enforcement and *crime_{t-1}* is the previous year's per capita crime rate.

The idea of *crime_{t-1}* is to capture this idea that cities which have higher crime rate historically would have spent more on law and enforcement hence leading to lower crime. Thus factors unobserved by us that affect current *crime* could be correlated with *expend* or *unem*. By including *crime_{t-1}*, we take part of the stuff from *u* to the equation.

9.2.2 A Different Slant on Multiple Regression

1. Tired of thinking about ideal relationship between y and x_1, x_2, \dots, x_k ?
2. Instead: deal with y on a subset of x_1, x_2, \dots, x_k which are observable and try to focus on getting a consistent and unbiased estimate conditional on the observables.
3. Does this solve the bias induced by leaving out the unobservable?
4. No, it does not. But if it is the next best thing possible, then go with what you can with the data you have rather than thinking about what you ought to have!

My take: This is a practical empirical solution, but it does not help us completely. I would still run behind the understanding the direction of bias caused by unobservables.

9.2.3 Potential Outcomes and Proxy Variables

The conclusions in this section come from ATE dealt with in section 7.6.

1. Consider a binary treatment indicator w and $y(0)$ and $y(1)$ denote without treatment and with treatment predicted values.
2. In a scenario when we have unobservables $v(0)$ and $v(1)$, we could proxy it with $\mathbf{x} = (x_1, x_2, \dots, x_k)$.
3. This implies that conditional on the \mathbf{x} , our program w is independent and tells us about its effect on the dependent variable of interest.
4. This takes the following form:

$$y(0) = \mu_0 + v(0) \quad y(1) = \mu_1 + v(1)$$

$$E[v(0)|w, \mathbf{x}] = E[v(0)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta}_0$$

$$E[v(1)|w, \mathbf{x}] = E[v(1)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta}_1$$