

# Supplementary Notes\*

Economics of Education (ECO383)

November 2011

## 1 Outline

This note

1. goes over omitted variables bias (a general framework is presented below);
2. talks in general about the value of experiments (and mention some of their demerits also – see below);
3. discusses some of the details of the K&W (2001) paper, especially focusing on the measurement framework and how this works.

In turn:

## 2 Omitted variables bias

Here, we present a general framework for analyzing the bias due to omitting potentially relevant variables from a linear regression.

Suppose the true underlying model is given by

$$T_i = \alpha + \beta C_i + \gamma X_i + \epsilon_i, \quad (1)$$

where  $T_i$  measures the test score for student  $i$  and  $C_i$  is the associated class size.

In practice, let us assume that the  $X$  variable is omitted, whatever it may be, and a sparser model is estimated, namely:

$$T_i = a + bC_i + \nu_i \quad (2)$$

---

\*Note by Robert McMillan.

A basic requirement for unbiasedness (if you refer back to our simple proof of unbiasedness in the single equation model) is that  $Cov(C, \nu) = 0$ . If this does not hold, then problems can arise. If we estimate equation (2), then  $Cov(C, \nu) \neq 0$ . Can you see why, given (1)?

In the face of omitted variables, we will consider the following possible cases:

## Uncorrelated ('irrelevant') omitted variables

If  $Cov(C, X) = 0$ , then  $C$  cannot 'pick up' any of the influence of the omitted variable, and the estimated parameter  $b$  will be unbiased for  $\beta$ . In this case, the omitted variable is irrelevant to the impact of  $C$ . This is a special case.

## Relevant omitted variables

Four cases can be distinguished here, according to whether

- $Cov(T, X)$  is less than or greater than zero – in short, does more  $X$  raise or lower test scores,  $T$ ?
- $Cov(C, X)$  is less than or greater than zero – is more  $X$  associated with more or less  $C$ ?

To give an example, suppose  $X$  measures parental education. Thus, typically we expect higher  $X$  to be associated with higher  $T$ . In turn, this gives rise to two cases depending on the covariance between  $C$  and  $X$ . If that correlation is positive, then we will have an *upward* bias in our least squares estimate of  $b$  from the restricted model. So we would expect it to be less negative than the true value  $\beta$ , or even positive.

The intuition is as follows: the omitted  $X$  variable is advantageous to producing higher test scores. Due to the positive correlation with  $C$ , higher  $C$  will typically be associated with higher  $X$ , so if  $X$  is omitted (as we assume), then as we raise  $C$ , the full effect of higher  $C$  on  $T$ , which is what the estimated coefficient captures, will reflect both the direct effect of  $C$ , but also the *indirect* effect of higher  $X$ . And recall that higher  $X$  is associated with higher  $T$ .

**Exercise:** Please work out for yourself the bias associated with the remaining three types of case. Specifically, fill out the full 2-by-2 matrix, listing the direction of bias in each case. This will be on the final (in some form or other).

*Actually, here's one of them:  $Cov(T, X) < 0$  and  $Cov(C, X) > 0$  (example of  $X$ : proportion free lunch - disadvantaged). Here, we get a downward bias, as  $C$  has a direct negative effect, but higher  $C$  picks up more  $X$ , which lowers test scores. So the indirect effect via  $X$  is negative, and will 'drag down' the impact of increased  $C$ .*

## Responses

The most obvious solution is to include all relevant variables. When this is not possible, perhaps because data on some relevant factor could not be collected, then we need to think carefully about possible biases. For this purpose, the above framework is useful.

## 3 Experiments

### 3.1 Advantages of experiments

i. Randomized experiments are often thought of as the ‘gold standard’ when we wish to learn about causal effects empirically. Especially attractive in light of the earlier discussion, they allow us to get around omitted variables problems.

Take the causal effects of class size reductions as an example. One can assign some schools (or classes within those schools) to the treatment group, and others to the control group, then compare performance differences.

In a regression of the form

$$T_i = \alpha + \beta C_i + \epsilon_i, \quad (3)$$

one can learn about the causal effect of reductions in class size because  $Cov(C, \epsilon) = 0$ , by construction. Further,  $C$  should be uncorrelated with *any* observable variable, at least if the randomization is done correctly. This is because the level of  $C$  is assigned randomly, unrelated to anything else, observable and unobservable.

ii. Experiments allow the researcher to vary  $C$  in a calculated way, to learn about different types of causal impact.

### 3.2 Some problems with Experiments

i. They can be very expensive. To keep costs down, often the scale of the experiment is restricted.

ii. It can be hard to generalize findings from an experimental study to a broader non-experimental setting.

1. Agents being studied can change their behaviour when they know they are being ‘treated.’ For example, teachers may think they have to make class size reductions a success, as this will bring more resources to the teaching profession in future. Thus special incentives operate that one would not expect in a regular setting.

2. General equilibrium effects: when a policy is enacted on a much larger scale than most experiments, this may cause other behaviour to change that leads the direct effects of the policy to be swamped by unforeseen indirect effects. For example, after looking at Tennessee’s Project STAR (Student Teacher Achievement Ratio), California implemented a state-wide class size reduction in 1996-97. When implemented on a statewide basis, there is a problem of where to get the new teachers necessary for smaller classes. Wealthier schools were able to attract better teachers from less fortunate schools, forcing the worst schools to recruit lower quality teachers to access the increased funding associated with the program. As Jepsen and Rivkin (2009) discuss, the positive effects of smaller class sizes on achievement were often counteracted by the decrease in teacher quality.

## 4 K&W (2001)

### 4.1 Randomization

It is customary to check that a randomization looks to have been correctly implemented.

In the case of Project STAR, randomization was conducted within-school for a given grade. Some students were randomly assigned to small classes, while the rest were not (of these, some had teachers’ aides also). So if we look at observable characteristics of students in the ‘small class’ treatment versus regular classes (let’s ignore any subdivision of the latter), their observable characteristics should look the same, on average.

Let us start by considering the equation underlying Table 1. We will use the notation appearing later in the paper.

$$SMALL_{isg} = \alpha_{0g} + \alpha_{1g}X_{is} + FE + \epsilon_{isg}, \quad (4)$$

where the subscripts are as follows:  $i$  is for students,  $s$  is for school, and  $g$  is for grade.

We want to make full sense of this equation, and the interpretation of its estimates, both because they allow us to see how well randomization worked in this case, and also because we will then be able to understand the key equation (1) in the K&W paper. To that end, there are several details that it is important to understand.

### Dummy variables

A dummy variable takes the value 1 if a certain condition is satisfied, and zero otherwise. The dependent variable in (4) takes two values only: 1 if a given student in a given school and grade is in a small class, and zero otherwise.

The  $X_{is}$  variables consist of three observable characteristics: race, gender, and free-lunch status. For example, the race dummy takes value 1 if student  $i$  is white or Asian, and zero otherwise. And the free-lunch dummy is 1 if a student  $i$  receives free or reduced-price lunches and zero otherwise.

Estimation: if a dummy variable serves as a dependent variable, then we can simply regress a 1 or a 0 on explanatory variables. Think how the corresponding scatterplot would look in this case.

When included among regressors, a dummy variable acts as a group intercept shifter. It is analogous to estimating the intercept  $\alpha$  but specifically for members of a relevant subgroup. The size of the estimate of the dummy gives the amount by which the common intercept should shift up or down for the group. Given this, there is no separate coefficient that attaches to the dummy variable: it is analogous to a group-specific intercept shifter.

## Fixed effects

We can extend the previous notion to define dummies for any well-defined subgroup, such as students who are in a given school, students who are in a given grade, and students who are in a given school and grade.

A fixed effect can be estimated in the same manner as a dummy variable (analogous to an intercept), yielding a shifter specific to a relevant sub-group.

Note: when we include fixed effects, they amount to fitting group-specific lines, all sharing the same slope, but allowing for different group-specific intercepts. So the variation that is used to estimate the slope (certainly the key parameter of interest here) comes *within the subgroup*, comparing treated with controls.

## R-squared ( $R^2$ )

This gives a measure of the overall fit of the model. It is equal to  $1 - \frac{RSS}{TSS}$ , and takes on values between 0 and 1. Note that the total sum of squares ( $TSS$ ) is equal to  $\sum_i (y_i - \bar{y})^2$  and the residual sum of squares ( $RSS$ ) is  $\sum_i (y_i - \hat{y}_i)^2$ , where  $y_i$  stands in for  $SMALL_{isg}$  and  $\hat{y}_i$  is its predicted value. The second term is really a measure of unexplained variance (the numerator is unexplained and the denominator is the total variance of the data). Thus, if  $R^2 = 1$ , then all variation is explained by the included variables.

## Tests for joint significance

Note the bottom line of Table 1: it gives a P-value for the joint significance of the explanatory variables, the null hypothesis being that the variables are jointly zero. A P-value close to zero indicates that the probability of obtaining the estimates with the estimated precision is extremely unlikely, in which case we would reject the null (as in column (1)). You will see that once the relevant fixed effects are included in columns (2) and (3), we no longer reject the null, consistent with successful randomization. Do you see why this is? (A P-value that is a number like 0.45 means that we would have an almost 50 percent chance of obtaining the estimates we do if the null hypothesis of zero slope coefficients were true.)

## Table 1 interpretation

Having been over those elements, the purpose of Table 1 is to check the quality of the randomization. One could just do a means comparison across treatment and control groups. But given the nature of the randomization, we need to compare treatment and control groups consisting of students in the same *school* and *grade*. If we do not (see column (1) in Panel A of Table 1), then the observable dummy variables have statistically significantly different estimated values, depending on whether students are in small classes or not. Fortunately, when the relevant fixed effects are added (most economically so in column (3)), then these differences disappear, consistent with the hypothesis of valid randomization.

In Panel B, the evidence suggests that randomization of teachers holds regardless of whether fixed effects are included.

## 4.2 Main estimating equation

Equation (1) in the paper is the main type of estimating equation, taking the form

$$Y_{isg} = \beta_{0g} + \beta_{1g}SMALL_{is} + \beta_{2g}X_{is} + \alpha_{sw} + \epsilon_{isg}, \quad (5)$$

where the dependent variable  $Y_{isg}$  is the percentile score of student  $i$  in grade  $g$  in school  $s$ . The  $X$  variables control for individual student characteristics, as before, and dummy variables are included at the school and entry-wave level, given by  $\alpha_{sw}$ . Why are these fixed effects the right ones? Because randomization is at the school-entry grade level, so we want to compare students in the same school and same entry grade, falling into treatment and control groups respectively. As argued above, the inclusion of these fixed effects will enforce that comparison.

The key coefficient of interest is  $\beta_{1g}$ , which captures the effect of being in a small class on students in a given grade. The equation (5) is estimated on a grade-by-grade basis,

and the results (the estimated coefficients) are plotted in Figure 2 for all students, then for subgroups in Figures 1 and 3. What emerges is that the effects fall by more than half once students progress into untreated grades 4 to 9. This prompts consideration of the longer run effects.

### 4.3 College test taking

The longer-run effects are difficult to analyze, given the data requirements. In this paper, the authors have assembled the requisite data, though an elaborate linkage process. And they also have a coherent estimation strategy, analogous to that used in equation (1) in the paper.

Specifically, suppose we consider a regression of a dummy variable denoting college test taking (versus not) on two different class size dummies,  $X$  control variables, and different sets of fixed effects, analogous to equation (1) in the paper (having switched dependent variable) though no longer on a grade-by-grade basis. The results are given in Table 3.

Aside: the authors estimate something called a logit model, rather than a linear regression, but you can think of the exercise as approximately the same as the sort of dummy variable regression in Table 1.

The results are graphed in Figure 4.

### 4.4 Selection

What is the selection concern? That a non-random subset of students apply to college, and confounding unobservables may be missed as a result. What is the specific concern here? That unobservably weaker students from the treatment group end up taking the test, so the estimates of the long-run treatment effect may be downward-biased.

It turns out that there are strategies for dealing with non-random sorting – rather beyond the scope of this course...