Empirical Methods in Applied Economics

Jörn-Steffen Pischke
LSE

October 2007

1 Observational Studies and Regression

1.1 Conditional Randomization Again

When we discussed experiments, we discussed already the possibility that randomization is conditional on a covariate, for example school effects in the STAR experiment. Random assignment conditional on X implies $D|X \perp y_0, y_1$. Since we are only interested in means, mean independence is sufficient, and we can write this condition as $E(y_0|D, X) = E(y_0|X)$. Using a linear model

$$E(y_0|X) = \alpha + \gamma X$$

and given homogeneous treatment effects, $E(y_1|X) = E(y_0|X) + \beta$, so that observed outcomes are described by the regression model

$$y = \alpha + \beta D + \gamma X + \varepsilon$$

Running this regression will recover the treatment effect. This is, of course, the motivation for many of the regressions we run with observational data, i.e. with data which come about because of the natural behavior of agents, rather than because of experimental variation. We are interested in the effect of D on y, and we believe that X affects both y, and selection into treatment. This situation is therefore frequently called selection on observables. If $E(y_0|D, X) = E(y_0|X)$ holds, D is sometimes said to be ignorable.

As an example, think of y as earnings, D is in an indicator for attending the LSE Ph.D. program, and X are undergraduate grades. If admission (and enrollment!) in the LSE Ph.D. program is more likely for those with higher grades but random within grades, then the regression will uncover the causal effect of an LSE Ph.D. on earnings. On the other hand, if admission also depends on recommendations from an undergraduate professor, and the

recommendation carries information for earnings, then controlling simply for grades will not solve the selection problem. Alternatively, say admission depends solely on grades but is otherwise random. However, students once admitted decide about enrollment based on their earnings potential. In this case, the regression controlling for grades would also not recover the causal effect because a control for the student's decision to enroll is missing.

It should be obvious from this discussion that selection bias is nothing other than omitted variables bias in this regression context. However, framing the problem by thinking about an underlying random assignment experiment helps in determining what needs to be controlled in a regression. The variables X need to include *everything* that determines the selection process. Information about how this process happens may come from an economic model (most likely leading to a much more structural approach to identification) or from knowledge of some administrative process. Papers using non-structural identification often rely on the latter, a clear description of some administrative process, as a natural experiment. The key ingredient in a good observational study is to convince the reader that after controlling for X, D is as good as randomly assigned with respect to the counterfactual outcomes.

Of course, in reality, we are often not as fortunate, and the research literature may be groping for good measures of what should be in X. A good example is the literature on the returns to schooling. Suppose that earnings are given by

$$y = \alpha + \beta S + \gamma A + \varepsilon \tag{1}$$

where y is the log of earnings, S is the number of years of schooling, and A is ability. In this semi-log formulation, β measures the return to a year of schooling. Since a measure of A is often not available, researchers typically run the short regression

$$y = a + bS + u$$

The estimate of b is

$$\begin{aligned} \operatorname{p\,lim} \widehat{b} &= \frac{cov(y,S)}{var(S)} = \frac{cov(\alpha + \beta S + \gamma A + \varepsilon, S)}{var(S)} \\ &= \frac{\beta var(S) + \gamma cov(S,A)}{var(S)} \\ &= \beta + \gamma \frac{cov(S,A)}{var(S)} = \beta + \gamma b_{AS} \end{aligned} \tag{2}$$

where b_{AS} denotes the bivariate regression coefficient of A on S. This is the standard omitted variables bias formula. The estimate of β will be biased

up if ability raises earnings ($\gamma > 0$) and the more able get more schooling ($cov(S, A) > 0$). Most people with Ph.D.s like to believe that the latter is the case. However, while this is possible and maybe likely, there is no compelling reason for it. Schooling may well be compensatory rather than complementary for earnings capacity A, so that $cov(S, A) < 0$. For example, Mick Jagger attended the LSE and obtained an MSc degree in economics. He decided not to go on to the Ph.D. because he had a high A, and hence was better off making money by playing music.

Griliches (1977) discusses the model (1) and provides some estimates from the US National Longitudinal Survey (NLS). He estimates the short regression as

$$\begin{array}{rccccc} \ln y = & \text{const.} & + & 0.068 & S & + & \varepsilon \\ & & & (0.03) & & & \end{array}$$

The NLS data contain an IQ test score, which Griliches includes as his measure of ability. He obtains

$$\begin{array}{rccccccc} \ln y = & \text{const.} & + & 0.059 & S & + & 0.0028 & IQ & + & \varepsilon \\ & & & (0.03) & & & (0.0005) & & & \end{array}$$

The estimate of β is indeed lower once the test score is included but not by too much. This result is consistent with $cov(S, A) > 0$, and the covariance being relatively small.

An alternative interpretation is that the IQ test score is a poor measure of ability, or it is measured with a lot of error so that even the long regression still does not give an unbiased estimate of the true return to schooling. If there is classical measurement error in IQ, and $cov(S, A) > 0$, then the coefficient on IQ will be biased towards zero, and the coefficient on S will be biased upwards. The NLS data contain another test score from the Knowledge of the World of Work test. This score is highly correlated with IQ, so Griliches uses it to instrument the IQ score in order to cure the measurement error problem. The result of the IV regression is

$$\begin{array}{rccccccc} \ln y = & \text{const.} & + & 0.052 & S & + & 0.0051 & IQ & + & \varepsilon \\ & & & (0.04) & & & (0.0009) & & & \end{array}$$

These results are again consistent with the interpretation we had in mind. There seems to be a substantial amount of attenuation bias in the coefficient on the IQ score in the OLS regression. The coefficient almost doubles when instrumented. However, the coefficient on schooling is also affected, and it falls by almost the same amount as it did moving from the short regression to the long regression including IQ. This comes from the fact

that IQ and schooling are positively correlated but the coefficient on IQ is underestimated in the OLS model. Hence, part of the effect of IQ loads on to the schooling coefficient. Including IQ cures some but not all of the omitted variables bias problem. Of course, the question remains, is schooling as good as randomly assigned conditional on the IQ test score?

1.2 Measurement Error in the Treatment

Most survey data are measured with error. We already considered the consequence of measurement error in a conditioning variable. However, the variable measuring the treatment may also be measured with error. You might think that years of schooling are probably measured relatively accurately. Nevertheless, a study of twins by Ashenfelter and Krueger (1994) suggests that the signal to total variance ratio in self-reported schooling is about 0.9. This ratio determines the relative attenuation of a mismeasured regressor in a bivariate classical errors-in-variables model.

We want to think again about comparing the short regression, just including schooling, to the long regression including a measure of ability. In order to see what happens, consider an artificial example. Suppose earnings is given by

$$\ln y = \quad 0.1 \quad S \quad + \quad 0.01 \quad A \quad + \quad \varepsilon$$

Observed schooling $\widetilde{S}$ is measured with error and the attenuation factor

$$\lambda = \frac{\sigma_S^2}{\sigma_{\widetilde{S}}^2} = 0.9$$

as in Ashenfelter and Krueger. Furthermore, assume $\sigma_{\widetilde{S}} = 3, \sigma_A = 15, \sigma_{AS} = 22.5$.

Running the short regression with just $\widetilde{S}$ as a regressor leads to both attenuation because of measurement error, and omitted variables bias. Using (2) we have

$$\begin{aligned} \operatorname{p\,lim} \widehat{b}_{\ln y \widetilde{S}} &= \lambda\beta + \gamma b_{AS} \qquad (3) \\ &= 0.9 \cdot 0.1 + 0.01 \frac{22.5}{9} = 0.115 \end{aligned}$$

The omitted variables bias dominates in this example, and the estimated return to schooling is biased upwards. Now consider adding A to the equation. The regression coefficient on S, $\widehat{b}_{\ln y \widetilde{S}.A}$, from the regression which

conditions on A, is given by

$$\operatorname{p\lim} \widehat{b}_{\ln y \widetilde{S}.A} = \frac{\lambda - R^2_{\widetilde{S}A}}{1 - R^2_{\widetilde{S}A}} \beta$$

where $R^2_{\widetilde{S}A}$ is the R^2 from a regression of $\widetilde{S}$ on A. Compared to (3), the long regression removes the omitted variable bias. However, the attenuation bias from measurement error increases by adding an additional regressor since

$$\frac{\lambda - R^2_{\widetilde{S}A}}{1 - R^2_{\widetilde{S}A}} < \lambda$$

The reason for this is that A and S are correlated. Hence, including A in the regression effectively removes some of the variation in $\widetilde{S}$ while leaving the measurement error unchanged. This reduces the attenuation factor.

Using the numbers from our example,

$$R^2_{\widetilde{S}A} = \left[\frac{\sigma_{A\widetilde{S}}}{\sigma_{\widetilde{S}}\sigma_A}\right]^2 = \left[\frac{\sigma_{AS}}{\sigma_{\widetilde{S}}\sigma_A}\right]^2 = \left[\frac{22.5}{3 \cdot 15}\right]^2 = 0.25$$

$$\operatorname{p\lim} \widehat{b}_{\ln y \widetilde{S}.A} = \frac{0.9 - 0.25}{1 - 0.25} 0.1 = \frac{0.65}{0.75} 0.1 = 0.087$$

In the short regression, the positive ability bias was partially offsetting the negative attenuation bias. In the long regression, the ability bias has been removed, while the attenuation bias has increased. The estimated return to schooling is now too small. The bias is about as large as in the short regression, where the two biases partially offset each other. By adding a perfectly valid and necessary control in the presence of measurement error we end up being no better off than before. If we suspect measurement error in S, but are uncertain about the importance of A, it is difficult to know whether the reduction in the regression coefficient in going from the short to the long regression is driven mainly by the measurement error or by omitted variables bias. As a result, it is difficult to judge whether the Griliches estimate of 0.068 or 0.052 is closer to the truth.

1.3 Controlling for Outcomes

A major pitfall in trying to control for selection with observables is the danger to end up controlling for outcomes. This can easily lead to misleading estimates. Consider our returns to schooling example again, where earnings y are determined by

$$y = \alpha + \beta S + \gamma A + \varepsilon$$

and $E(\varepsilon|S, A) = 0$. Suppose that A cannot be measured directly. An ability measure MA is available, and measured ability is related to innate ability by the equation

$$MA = \pi_o + \pi_1 S + A. \tag{4}$$

Schooling increases measured ability. Hence, measured ability is an outcome of the schooling treatment.

Since we are worried about $cov(S, A) \neq 0$, we are running the long regression

$$y = a + bS + cMA + u.$$

What is the estimate $\widehat{b}$ from that regression? Solve (4) for A, and substitute into (1) to get

$$\begin{aligned} y &= \alpha + \beta S + \gamma(MA - \pi_o + \pi_1 S) + \varepsilon \\ &= \alpha - \pi_o + (\beta - \gamma\pi_1)S + \gamma MA + \varepsilon \end{aligned}$$

so that $b = \beta - \gamma\pi_1$, i.e. the long regression with measured ability does not identify the true return to schooling β. By including MA in the regression, which includes the effect of S, part of the effect of S loads on to the MA term. The estimate of β is biased down as a result.

In many applications there are multiple outcomes, which a treatment affects. For example, think of D as participation in a teacher training program in a developing country. You have measures of two relevant outcomes, y_s is a student test score, and y_a is a student's school attendance record. For simplicity, assume that D is randomly assigned. In this case, both the regressions

$$\begin{aligned} y_a &= \alpha_a + \beta_a D + \varepsilon_a \\ y_s &= \alpha_s + \beta_s D + \varepsilon_s \end{aligned}$$

have a causal interpretation. However, notice that attendance of the student is likely to affect student performance, and hence the test score y_s. Hence, it is natural to ask: "What is the direct effect of D on y_s, and what is the indirect effect through y_a?" It is tempting to argue that regression uncovers the partial effects, hence running the regression

$$y_s = \theta_0 + \theta_1 y_a + \theta_2 D + u \tag{5}$$

lets us uncover the direct effect.

There are now three causal effects we are interested in: the effects of D on y_s and y_a, and the effect of y_a on y_s. Our previous discussion puts us in a good position to assess the model (5). For this regression to recover the causal effect of y_a on y_s, it needs to be true that conditional on D, y_a is as good as randomly assigned. Indeed, D is a selection variable, since it affects both y_s and y_a. However, we have not said anything in our discussion so far as to whether y_a may be reasonably thought of as randomly assigned. In the example, student attendance is unlikely to be randomly assigned among the kids whose teacher does not get training, and those who does. In effect, student characteristics like motivation or parental interest in schooling will most likely determine both attendance and performance. In this case, regression (5) will recover neither the causal effect of D, nor of y_a. Although we have $D \perp y_{s0}, y_{s1}$ this does not imply $D|y_a \perp y_{s0}, y_{s1}$. Unconditionally, two individuals with $D = 1$ and $D = 0$ have the same parental background, motivation, etec. But since D tends to raise y_a, once you compare two individuals with the same level of y_a this will no longer be true. Now the $D = 1$ individuals will be the relatively less motivated than the $D = 0$ individuals.

The same happens if there is attrition which depends on D. Then attrition is an outcome that is correlated with both D and counterfactual outcomes. Hence, although we may start with random assignment, we no longer have random assignment in the sample selected by attrition.

1.4 Falsification Exercises

How do we determine in practice whether treatment is as good as randomly assigned conditional on some observables? Mostly, this will involve a judgement call. It is instructive to go through an empirical example, in order to see how a strategy of controlling for observables might work. Let me give an example from my own work. Britain changed from a system of selective secondary schooling to a system of comprehensive secondary schooling during the 1970s. In the selective system, children would take a test at age 11 and attend either a more academically oriented grammar school if they did well, or a less academic secondary modern school otherwise. In the comprehensive system, the test was abolished and everyone attends the same school until age 16. We are interested in the effect of selection on student performance. Do students do better in the tracked, selective system or do they perform better in the comprehensive system?

The British reform offers potentially a good testing ground for this question. The reform happened at different speeds in different local education

authorities. By 1969, there were some authorities which were fully comprehensive, some which were still fully selective, and many which were in between. The children in the National Child Development Study (NCDS) were age 11 in 1969. The NCDS has therefore been analyzed extensively to look at this question. Unfortunately, the reform did not happen randomly across authorities. Poorer, left leaning authorities with more poorly performing students tended to convert to the comprehensive system first. Hence, it is necessary to control for these differences in characteristics of the students attending comprehensive schools versus other schools.

Table 1 shows some results from this exercise. It regresses a math test score (scaled to range from 0 to 100) on whether a student attended a comprehensive school. Comprehensive school students scored almost 8 points lower, which is quite a large effect. However, column (2) shows that this is mainly due to the fact that more poorly performing students attend these schools. Controlling for the math test score at age 11 lowers the effect to about 2 points. The data set actually includes various different subject tests, and controlling for four other tests at age 11 does not change the results much. Neither does including a small number of family background controls in column (4). These controls were taken from a previous study by Kerkhoff and coauthors. In fact, the data set has a rich set of controls available. Including about another 50 covariates in column (5) changes the coefficient slightly, but the general magnitude (1.5) is still the same (notice that these extensive controls resulted in a loss of observations because of missing values, so the column (4) and (5) samples are not exactly the same). If it is not completely clear which controls are necessary, including a reasonable set of controls successively is useful. Since the coefficients in all columns from (2) to (5) are somewhat similar, it seems reasonable to conclude that any omitted factors are probably much like the observables already included, and hence would not change the results much more either.

This is not a test of the strategy, however. While the results in table 1 suggest that comprehensive education may lower math scores by 1 to 2 points, this result is actually rather dubious. The data set also contains test scores at age 7. Hence, it allows me to run the same regressions using the age 11 test score as outcome and the age 7 test score as control. This covers the period of primary school when all children are educated together. Although there may be some effects of the selective regime (for example, because children in comprehensive areas will study for the age 11 exam), we would expect any such effects at the younger ages to be much less pronounced. Table 2 displays these results and suggests that the effects are even bigger. No matter how many controls are added, a large comprehensive

school effect remains. This casts doubt on the interpretation of the table 1 effects as causal.

This strategy of looking for data which allow a "falsification exercise" has become quite popular in the applied literature. As in table 2, you want to find some data where you can a priory rule out an effect of the treatment. If you nevertheless find the effect, it casts doubt on the identification strategy. The falsification test really only makes sense if you find an effect of the treatment on the outcome you are interested in. If the effect in the real data is zero, then there is little to be learned from the fact that it might be zero using a different outcome variable, which should not be affected, as well.

Table 1
Math Test Scores at 16
OLS Regressions

Regressor	(1)	(2)	(3)	(4)	(5)
Attends Comprehensive School	-7.74 (0.54)	-2.18 (0.35)	-2.07 (0.34)	-1.80 (0.34)	-1.48 (0.37)
Control Variables:					
Math Test Score at 11	no	yes	yes	yes	yes
Other Test Scores at 11	no	no	yes	yes	yes
Kerkhoff et al. type background controls	no	no	no	yes	yes
Large set of background controls	no	no	no	no	yes
Number of observations	6734	6734	6734	6734	5747

Note: Control variables need to be described somewhere.

Table 2
Math Test Scores at 11
OLS Regressions

Regressor	(1)	(2)	(3)	(4)	(5)
Attends Comprehensive School	-8.39 (0.62)	-6.01 (0.52)	-4.91 (0.46)	-3.96 (0.45)	-3.29 (0.45)
Control Variables:					
Math Test Score at 7	no	yes	yes	yes	yes
Other Test Scores at 7	no	no	yes	yes	yes
Kerkhoff et al. type background controls	no	no	no	yes	yes
Large set of background controls	no	no	no	no	yes
Number of observations	6734	6734	6734	6734	6223

Note: Control variables need to be described somewhere.