ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

CAUTIONARY NOTES:

IT'S A HARSH WORLD OUT THERE

We have seen in the past two lectures that:

(a) it is almost always valid to run a regression as long as you interpret it correctly, and

(b) randomized experiments are generally the preferred method for estimating causal effects under the potential outcomes framework.

Under what conditions will a linear regression – the bread and butter of econometrics – approximate a randomized experiment? Loosely speaking, we need it to be the case that x_i is "as good as randomly assigned," i.e. x_i needs to be uncorrelated with unobserved factors that determine y_i after controlling for other observable factors. This type of research design is often referred to as "selection on observables." How often does it hold up in practice? Not as often as we would like.

1 LaLonde (1986): The NSW

LaLonde (1986) analyzes a randomized experiment evaluating a job training program, the National Supported Work Demonstration (NSW). The NSW, operated by Manpower Demonstration Research Corporation (MDRC), "admitted into the program AFDC women, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes."[1] (LaLonde 1986, p. 605) While the NSW is shown to increase post-training earnings by \$800-\$900 (1982 dollars), that is not the main focus of the article. Instead, LaLonde uses the experimental estimates as a "benchmark" to test whether typical econometric techniques can reproduce the same

[1] It is unclear from LaLonde's description how the MDRC administrators chose which applicants would enter the experiment. Given that there were only 6,616 trainees distributed between 10 cities, there was presumably a scarcity of slots relative to potential applicants.

results. The short answer is that they cannot.

LaLonde needs a simulated control group in order to conduct his exercise – if he applies any sensible estimator to the experimental data (treated and control groups), he will get reasonable estimates because the treatment is randomly assigned. He therefore constructs a series of simulated control groups using data from the PSID and the CPS (merged with SSA data). This is somewhat unusual in that the treated individuals and the control individuals are drawn from two entirely separate data sets, but it is not unreasonable for his purposes. LaLonde begins the benchmarking exercise by applying a series of differences-in-differences type estimators. The basic model is:

$$(1)\ y_{i,1979} - y_{i,1975} = \delta D_i + (\varepsilon_{i,1979} - \varepsilon_{i,1975})$$

This specification differences out any unobserved individual effects that are constant over time – it is equivalent to including individual fixed effects in the cross-sectional regression. Identification comes from comparing the change in earnings for those that participated in training to the change in earnings for those that did not participate. LaLonde also supplements this model with a regression that, instead of differencing, controls for pre-treatment earnings. This specification is more flexible in that it does not restrict the coefficient on pre-treatment earnings to be one.

$$(2)\ y_{i,1979} = \delta D_i + \beta y_{i,1975} + X_i\gamma + \varepsilon_{i,1979}$$

Table 1 presents estimates from these two specifications. The "Pre-Treatment" column presents differences between the income of treatment and control groups (or simulated control groups) in 1975, before the training program starts. If the treatment is randomly assigned, this difference should be close to zero, and for the true controls it is. LaLonde presents eight simulated control groups – for brevity I present the two control groups per gender that were closest to the experimental sample in terms of pre-treatment income. Table 1 is therefore more favorable to the nonexperimental estimates than LaLonde's equivalent

tables. Nevertheless, we observe large variations between the experimental estimates and the nonexperimental estimates.

The "Diffs-in-Diffs" column presents estimates using the first differences specification presented above (equation 1). The experimental benchmarks are $833 for females and $847 for males. The nonexperimental estimates range from -$1,637 to $3,145, and in only one case does the nonexperimental confidence interval contain the experimental point estimate. The last two columns of Table 1 apply the model presented in equation 2 (first without additional covariates, and then with additional covariates). These models perform slightly better – three of the eight point estimates get reasonably close to the experimental benchmarks (one of them gets quite close). Nevertheless, the pre-treatment differences are actually the worst for the samples that produce the closest results, so there is no reason to believe that an objective econometrician would reliably choose point estimates close to the experimental benchmarks.

Table 1: One-Stage Estimates

Estimator:	Pre-Treatment Differences	Diffs-in-Diffs	Controlling For Previous Earnings	Fully Adjusted
Females				
Controls	-17	833	843	854
	(122)	(323)	(308)	(312)
PSID-3	-77	3,145	3,070	2,919
	(202)	(557)	(531)	(592)
CPS-4	-1,189	2,126	1,222	827
	(249)	(654)	(637)	(814)
Males				
Controls	39	847	897	662
	(383)	(560)	(467)	(506)
PSID-3	455	242	629	397
	(539)	(884)	(757)	(1,103)
CPS-3	337	-1,637	-1,396	1,466
	(343)	(631)	(582)	(984)

Notes: Standard errors in parentheses. Source: LaLonde (1986).

LaLonde then considers the performance of more advanced two-stage estimators. In particular, he applies the Heckman selection correction model from Heckman (1978). The

Heckman selection correction models two equations separately: the participation equation (the first stage, a non-linear probit model) and the earnings equation (the second stage). In this sense it is not unlike two-stage least squares, but there are a couple key differences. First, it uses a "control function" approach to solve the endogeneity problem. Specifically, it uses estimates from the first stage equation as a regressor to control for the expected value of the earnings residual conditional on participation and the determinants of participation. Second, because it specifies the participation (treatment) dummy as a non-linear function of the covariates, it is possible to identify the training coefficient without any instruments (i.e., exclusion restrictions). Nevertheless, LaLonde experiments with several (questionable) instruments to see how well this model performs.

Table 2: Two-Stage Estimates

	Females		Males	
	Training	Participation	Training	Participation
Controls	861	284	889	-876
	(318)	(2,385)	(840)	(2,601)
5 Sketchy IVs	1,102	-606	-22	-1,437
	(323)	(480)	(584)	(449)
3 Sketchy IVs	1,256	-823		
	(405)	(410)		
2 Sketchy IVs	1,564	-552	13	-1,484
	(604)	(569)	(584)	(450)
No Instrument	1,747	-526	213	-1,364
	(620)	(568)	(588)	(452)

Notes: Standard errors in parentheses. Source: LaLonde (1986).

The results from the Heckman two-step estimator are reported in Table 2. The Heckman correction allows you to test the exogeneity of the treatment indicator by testing whether the coefficient on the selection correction term in the second stage is significantly different than zero. For brevity I present only results for the samples that displayed the least evidence of selection into the treatment. The two-step estimators perform somewhat better than the one-

step estimators, but the results are still not encouraging. On the positive side, the confidence intervals for all but one of the nonexperimental estimates contain the experimental point estimates. But the standard errors are so large that much of this "encouraging" performance is primarily due to the fact that the confidence intervals are huge. Male nonexperimental estimates are particularly bad, ranging from -$1,333 to $213 (see LaLonde's Table 6 for the full set of results). It seems likely that if additional data were available, the nonexperimental estimates would converge to different values than the experimental estimates.

When LaLonde's paper was published in 1986, it caused significant consternation among applied researchers trying to estimate causal effects. It is probably not an understatement to say that it sparked the pursuit of clean, transparent research designs that continues to this day.

2 Freedman (1991): A Natural Experiment

Freedman (1991) offers a critique of linear regression applications along with an example of an historical natural experiment as an alternative research design. Freedman begins with four possible views of regression, progressing from the most optimistic to the most pessimistic:

> (1) Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong.
>
> (2) Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use.
>
> (3) Regression might work, but hasn't yet.
>
> (4) Regression can't work.

Source: Freedman (1991), p. 292.

Freedman professes that his own view falls between (2) and (3). I'm not sure exactly what (3) entails – the properties of linear regression are pretty well-established, so if it were

going to work, I would think it would have done so by now. But, like Freedman, I agree that "good examples [of causal estimates from regression] are quite hard to find."

In contrast to regression models (and more sophisticated models), Freedman presents the work of John Snow on cholera in the 1850s (that is to say, Snow conducted the work during the 1850s, on cholera at that time). Snow postulated that unsanitary water caused cholera outbreaks (at the time it was believed that cholera arose from poisonous particles in the air). Snow had several pieces of circumstantial evidence to support his position, but in order to prove his hypothesis he observed that water distribution in London gave rise to a natural experiment.

In the area that Snow was studying, two water supply companies, Southwark and Vauxhall Company and Lambeth Company, competed for customers. One company (Lambeth) drew water upstream of the sewage discharge points in the River Thames, while the other (Southwark and Vauxhall) drew water downstream of the discharge points. Both companies had pipes running down virtually every street and alley, and which houses chose which company appeared to be virtually random. Snow wrote, "Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies." In today's terminology, Snow would say that the observable attributes (covariates) were balanced across the two companies. Having convinced himself that the choice of water company was nearly random, he examined the cholera death rate for customers of both companies.

The cholera results, presented in Table 3, are striking. Death rates for the downstream company are over eight times higher than death rates for the upstream company. Given the sample size, and the fact that the customers of both companies are spatially intermixed, it is clear that these results are highly significant despite the absence of standard errors. As Freedman writes (p. 298):

> As a piece of statistical technology, Table [3] is by no means remarkable. But the story it tells is very persuasive. The force of the argument results from the clarity

of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data.

Table 3: Snow's Table IX

	Number of Houses	Deaths from Cholera	Deaths per 10,000 Houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Notes: Source: Freedman (1991).

Freedman's emphasis is that the findings' credibility is due to the persuasive research design in conjunction with an impressive data set, rather than the sophistication of the statistical modeling technique. The implication is that, if you can't get data that have some sort of clean variation in the treatment of interest, then you can't convincingly identify a causal effect, no matter how fancy an estimation technique or theoretical model you apply. My *personal* view is in line with Freedman's, but it is certainly a matter considered open for debate within economics/econometrics. Regardless, as a piece of empirical evidence, Snow's 150-year-old study is clearly more credible than the vast majority of articles published today in economics (or other social sciences).

Freedman gives several examples of unconvincing regression studies – these aren't really worth reading since they are so common. He also stresses the importance of replication in the context of a response to LaLonde (1986) by Heckman and Hotz (pp. 306-307). This is an important point that unfortunately gets little recognition within economics (along this dimension, the medical literature is ahead of us) – our field generally assigns little value to replicative studies unless they produce remarkably different results. The problem is that modern computing power allows the estimation of dozens, if not hundreds, of different models on the same data set. Given that researchers can only be expected to report the results of a few of these models, it is difficult to tell whether a result represents a true underlying relationship, or whether it is simply a statistical artifact of the ability to choose between so many different estimates. The only way to truly test the finding is to verify the result in

different contexts with different data sets. But the incentives are not aligned to encourage researchers to engage in that type of work.

With these caveats in mind, we begin our study of estimation techniques developed to uncover causal effects. Virtually all of these techniques have historical roots that precede the LaLonde (1986) paper by years, if not decades. Nevertheless, their increased popularity is likely in part a reaction to LaLonde's study.

3 Additional References

Heckman, James. "Dummy Endogenous Variables in a Simultaneous Equations System." *Econometrica*, 1978, 46, 931-59.