Empirical Methods *

MIT 14.771/ Harvard 2390b

The goal of this handout is to present the most common empirical methods used in applied economics. Excellent references for the program evaluation and natural experiment approach are Angrist and Krueger (1999), and Mayer (1999). Angrist and Krueger (1999) contains more material and at a more detailed level than this handout and should be a high priority paper to read for students planning to write a thesis in empirical development, labor of public finance.

1 The evaluation problem

Empirical methods in development economics, labor economics, and public finance, have been developed to try to answer *counterfactual* questions. What would have happened to this person's behavior if she had been subjected to an alternative policy T (e.g. would she work more if marginal taxes were lower, would she earn less if she had not gone to school, would she be more likely to be immunized if there had been an immunization center in village?).

Here is an example that illustrates the fundamental difficulties of program evaluation:
Let us call Y_i^T the average test scores of children in a given school i if the school has textbooks, and Y_i^C the test scores of children in the same school i if the school has no textbooks. We are interested in the difference $Y_i^T - Y_i^C$, which is the effect of having textbooks for school i.
Problem: we will never have a school i both with and without books at the same time. What can we do? We will never know the effect of having textbooks on a school in particular but we may hope to learn the average effect that it will have on schools: $E[Y_i^T - Y_i^C]$.

* Handout by Prof. Esther Duflo

Imagine we have access to data on lots of schools in one region. Some schools have textbooks and others do not. We may think of taking the average in both groups, and the difference between average test scores in schools with textbooks and average test scores in schools without textbooks. This is equal to:

$$D = E[Y_i^T|\text{School has textbooks}] - E[Y_i^C|\text{School has no textbooks}] = E[Y_i^T|T] - E[Y_i^C|C]$$

Subtract and add $E[Y_i^C|T]$, we obtain,

$$D = E[Y_i^T|T] - E[Y_i^C|T] - E[Y_i^C|C] + E[Y_i^C|T] = E[Y_i^T - Y_i^C|T] + E[Y_i^C|T] - E[Y_i^C|C]$$

The first term $E[Y_i^T - Y_i^C|T]$ is the *treatment effect* that we try to isolate (effect of treatment on the treated): on average, in the treatment schools, what difference will the books make?

The difference $E[Y_i^C|T] - E[Y_i^C|C]$ is the selection bias. It tells us that, beside the effect of the textbooks, there may be systematic differences between schools with textbooks and other schools.

Empirical methods try to solve this problem.

2 Randomized evaluations

The ideal set-up to evaluate the effect of a policy X on outcome Y is a randomized experiment. Useful reference is Rosenbaum (1995).

In a randomized experiment, a sample of N individuals is selected from the population (note that this sample may not be random and may be selected according to observables). This sample is then divided *randomly* into two groups: the Treatment group (N_T individuals) and the Control group (N_C individuals). Obviously $N_T + N_C = N$.

The Treatment group is then treated by policy X while the control group is not. Then the outcome Y is observed and compared for both Treatment and Control groups. The effect of policy X is measured in general by the difference in empirical means of Y between Treatments and Controls:

$$\hat{D} = \hat{E}(Y|T) - \hat{E}(Y|C),$$

where $\hat{E}$ denotes the empirical mean.

As Treatment has been randomly assigned, the difference $E[Y_i^C|T] - E[Y_i^C|C]$ is equal to 0 (in the absence of the treatment, schools are the same). Therefore,

$$E[Y_i|T] - E[Y_i|C] = E[Y_i^T - Y_i^C|T] = E[Y_i^T - Y_i^C],$$

the causal parameter of interest.

The regression counterpart to obtain standard errors for $\hat{D}$ is,

$Y_i = \alpha + D \cdot 1(i \in T) + \epsilon_i$

where $1(i \in T)$ is a dummy for being in the Treatment group.

How? The formula for $\hat{D}_{OLS}$ is simple to handle when there is only one independent variable:

$$\hat{D}_{OLS} = \frac{\sum_i 1(i \in T)[Y_i - \bar{Y}]}{\sum_i 1(i \in T)[1(i \in T) - N_T/N]}$$

The denominator is equal to: $Den = \sum_i 1(i \in T)^2 - (N_T/N)\sum_i 1(i \in T) = N_T(1 - N_T/N)$

The numerator is equal to: $Num = \sum_i 1(i \in T)[Y_i - \bar{Y}] = \sum_i 1(i \in T)Y_i - \bar{Y}\sum_i 1(i \in T)$

which implies:

$Num = N_T\hat{E}(Y|T) - N_T[N_T\hat{E}(Y|T) + N_C\hat{E}(Y|C)]/N = N_T(1 - N_T/N)\hat{E}(Y|T) - (N - N_T)\hat{E}(Y|C) = N_T(1 - N_T/N)[\hat{E}(Y|T) - \hat{E}(Y|C)]$.

Taking the ratios of Num and Den, we indeed find that:

$$\hat{D}_{OLS} = \hat{E}(Y|T) - \hat{E}(Y|C).$$

- **Problems of Randomized Experiments**

1. Cost

   (a) Financial costs

   Experiments are very costly and difficult to implement properly in economics. The negative income tax experiments of the late 60s and 70s in the US illustrate most of the issues (see (Pencavel 1986, Ashenfelter and Plant 1990)). As a result they are often either poorly managed, or small, or both (with the corresponding problems we will see below).

   (b) Ethical problems

   It is not possible to run all the experiments we would like to because they might affect substantially the economic or social outcomes of the Treated. Alternatively, NGOs or governments are reluctant to deprive the controls from treatment which they consider potentially valuable. Insisting on the fact that it is a productive use of limited resources may be a good way to go...

2. Threats to internal validity:

   (a) Non response bias:

   People may move off during the experiment. If people who leave have particular characteristics systematically related to the outcome then there is attrition bias. (cf. Hausman and Wise (1979) about attrition in the NIT experiment).

   (b) Mix up of Treatment and Controls:

   Sometimes, maintaining the allocation to control and treatment to be random is almost impossible. Example: (Krueger 2000) evaluation of the Tennessee Star small class size experiment: children were moved to small classes (due to parental pressures, bad behavior,etc..). The actual class is therefore not random even though the initial assignment was random. It is then important to use the *initial assignment* as the treatment, because it is the only variation that was randomly assigned. It can then be used as an instrument for actual class size (cf. below).

3. Threats to external validity

   (a) Limited duration:

   Experiments are in general temporary. People may react differently to a temporary program than to a permanent program.

   (b) Experiment Specificity:

   In general, an experiment is run in a particular geographic area (e.g., the NIT experiments). It is not obvious that the same experiment would have given the same results in another area. Therefore, it is often difficult to generalize the results of an experiment to the total population.

   (c) Hawthrone and John Henry effects:

   Treatment and control may behave differently because they know they are being observed. Therefore the effects may not be generalized to a context where subjects are not observed.

   (d) General Equilibrium effects:

   Extrapolation complicated because of general equilibrium effects: small scale experiments do not generate general equilibrium effects that might be very important when policy is applied to everybody in the population.

4. Threats to power

   (a) Small samples:

   Because experiments are difficult to administer, samples are often small, which makes it difficult to obtain significant results. It is important to compute power calculation before starting an experiment (what is the sample size required to be able to discriminate from 0 an effect of a given size?). See the command sampsize in stata. But the crucial inputs (mean and variance of the outcomes before treatment) are often missing, so that there is always some guess work involved in planning experiments.

   (b) Experiment design and power of the experiment:

   When the unit of randomization is a group (e.g. a school), we may need to collect data on a very large number of individuals to get significant results, if outcomes are strongly correlated within groups (see below how standard errors are corrected for the grouped structure). This was a difficulty in the Kremer, Glewwe and Moulin (1998) textbook experiments.

3 Controlling for selection bias by controlling for observables

3.1 OLS

OLS is the basic regression design.

3.1.1 Definition

$Y = X\beta + \epsilon$

Suppose we have N observations

Y is the dependent variable $N \times 1$ vector

X are the independent variables $N \times K$ vector (K independent variables). One element of the X may be T, the variable we are interested in. We note $X = (T, x_2, .., x_K)$

ϵ is the error term $N \times 1$ vector

The OLS estimator is: $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\epsilon$

$\hat{\beta}$ is consistent if ϵ and X are *uncorrelated*, that is, $E(X'\epsilon) = 0$.

NB: this is not as strong a requirement as being independent.

Stata OLS command: regress y T x_2.. x_K

where y is the name of the dependent variable, T is the variable of interest and x_2 .. x_K are the names of the $K - 1$ control variables.

3.1.2 Inference

The asymptotic variance, which stata reports, is correct when the variance of the error term is diagonal (this rules out autocorrelation) with identical terms on the diagonal (this rules out heteroscedasticity), that is,

$V(\epsilon) = \sigma_\epsilon^2 I_N$ where I_N is the identity matrix of rank N.

The asymptotic variance of the OLS estimator is given by:

$\text{VAR}(\beta) = \sigma_\epsilon^2 (X'X)^{-1}$

When the error term is non-spherical $V(\epsilon) = \Omega$, the asymptotic variance of the OLS estimator is different from the previous formula and is given by:
VAR $= (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$
There are two important examples of non-spherical disturbances:

1. Heteroskedasticity:

   Ω is diagonal (ϵ_i is uncorrelated with ϵ_j when $i \neq j$) but $Var(\epsilon_i)$ may vary with i.

   Stata command: regress y x1 .. xK, robust

   produces correct standard errors in that case using the White method.

2. Group error structure:

   Example: Survey design in developing countries is often clustered. (cf. Deaton (1997)'s book for more on this). First, clusters (i.e. villages or neighborhoods are randomly selected), then individuals are selected within clusters.

   $Y_{ij} = X_{ij}\beta + \epsilon_{ij}$ where i is the individual and j is the village.

   Assume that there are village common fixed effects:

   $\epsilon_{ij} = \mu_j + \nu_{ij}$ where the ν_{ij} are independent and with constant variance.

   Then the error term matrix Ω is bloc diagonal.

   stata command : regress y x1 .. xK, cluster(village)

   where village is the subgroup indicator, produces standard errors which are corrected both for heterockedasticity and the grouped structure.

3.1.3 Problems with OLS

1. Under-controlling

   The most frequent problem with OLS is that of omitted variable bias. Our coefficient is likely to be biased if we omit relevant control variables. The classic example is that of returns to education. If ability (or other factors affecting future earnings) are correlated with education choice and are not included in the regression, the OLS coefficient is biased.

Suppose our true model is

$$Y_i = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

where T represents our variable of interest (e.g. schooling) and X_3 and X_4 represent other control variables (e.g. ability, family background). However, we do not have information on X_3 and X_4, so we run the "short regression":

$$Y_i = \beta_0^* + \beta_1^* T + \beta_2^* X_2 + \eta$$

Then we know that

$$\beta_1^* = \frac{Cov(Y, \tilde{T})}{Var(\tilde{T})}$$

where $\tilde{T}$ is the residual from the regression of T on X_2 i.e.

$T = \gamma_0 + \gamma_1 X_2 + \tilde{T}$ with $Cov(X_2, \tilde{T}) = 0$

So, the numerator of β_1^* is

$$Cov(Y, \tilde{T}) = Cov(\beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \tilde{T})$$

$$= Cov(\beta_1 T + \beta_3 X_3 + \beta_4 X_4, \tilde{T}) = \beta_1 Var(\tilde{T}) + \beta_3 Cov(X_3, \tilde{T}) + \beta_4 Cov(X_4, \tilde{T})$$

$$\Rightarrow \beta_1^* = \beta_1 + \beta_3 \delta_{31} + \beta_4 \delta_{41}$$

where δ_{31} = coefficient on T when X_3 is regressed on T and X_2, and δ_{41} = coefficient on T when X_4 is regressed on T and X_2. In words:

Short regression coeff. = Long regression coeff. + [coeffs. on omitted variables in long regression] × [coeffs. on omitted variables when regressed on included variables]

This formula is very useful in determining the sign of the omitted variables bias. For instance, in the returns to education example with ability as the omitted variable, we expect that unobserved ability will have a positive impact on wages in the long regression. If we assume that higher ability people choose to get more schooling, then the omitted variables bias is positive, which means that our estimated coeff. on schooling is biased upwards.

2. Over-controlling

Controlling for variables that are *caused* by the variable of interest will also lead to biased coefficient. For example, if wage and ability (as measured by IQ, for example), are both caused by schooling, then controlling for IQ in an OLS regression of wage on education will lead to a downward bias in the OLS coefficient of education (intuitively: the ability variable picks up some of the causal effect of education, namely the increase in wages which is due to the effect of education on ability which itself affects wages).

The relationship between short and long regression coefficients is still given by the omitted variables formula above, only here the short regression is the coefficient we really want, and the long regression is what we mistakenly run. In the case of the schooling example, it thus results in a downward bias.

3. Estimating the extent of omitted variables bias

Computing the formula above explicitly is difficult to do since we typically do not have information on the omitted variables. However, if the true relationship depends on a large number of variables, and the included regressors are a random subset of this set of factors and none of the factors dominates the relationship with wages or schooling, then the relationship between the indices of observables in the schooling and wage equations is the same as the relationship between the unobservables ((Altonji, Elder and Taber 2000)). To get an idea of how much our results might be affected due to unobserved covariates, we can compute how large the omitted variables bias must be to make our results invalid. If our schooling variable takes only 2 values 0 and 1, we can compute the normalized shift in schooling due to observables:

$$\frac{E(X'\beta|S=1)-E(X'\beta|S=0)}{Var(X'\beta)}$$

and ask how large the normalized shift due to unobservables

$$\frac{E(\epsilon|S=1)-E(\epsilon|S=0)}{Var(\epsilon)}$$

would have to be in order to explain away the entire estimate of β_1. If selection on unobservables has to be very large compared to selection on observables in order to attribute all our results to omitted variables bias, we feel more confident about our results.

3.2 Matching

3.2.1 Matching on observables

Instead of doing a regression, it is possible to use matching methods. Matching is easier to implement when the treatment variable takes only two values. Clearly presented application is (Angrist 1998).

An obvious case is when the treatment effect is random conditional on a set of observable variables X. Example: at Dartmouth, roommates are allocated randomly after conditioning for responses to a set of questions: are you more neat or messy?, do you smoke?, do you listen to loud music? People with the same answers to all of these questions are put in a pile and then randomly allocated to each other and to a room.

What is the effect of the high school score of my roommate on my GPA? ((Sacerdote 2000)). Imagine the treatment variable $T = 1$ if the roommate has a high score in high school. Randomization conditional on observables imply that:

$$E[Y_i^C|X,T] - E[Y_i^C|X,C] = 0$$

So:

$$E[Y_i|X,T] - E[Y_i|X,C] = E[Y_i^T|X,T] - E[Y_i^C|X,T]$$

And therefore:

$$E_X\{E[Y_i^T|X,T] - E[Y_i^C|X,C]\} = E[Y_i^T - Y_i^C|T],$$

Our parameter of interest.

Finally,

$$E_X\{E[Y_i^T|X,T] - E[Y_i^C|X,T]\} = \int \{E[Y_i^T|x,T] - E[Y_i^C|x,C]\}P(X = x|T)dx$$

This means that, if X takes discrete values, we can compare Treatment and Control in all the cells formed by the combination of the Xs (e.g.: neat, smoker, no loud music), and then take a weigthed average over these cells, using as weights the proportion of treated in the cells (this is the sample analog of this expression).

Cells where there are only controls or only treatments are dropped.

Comparing matching and OLS:

- They are the same if the treatment effects are constant

- If treatment effects are different, they will be different, because they apply a different weighting schemes. OLS is efficient under the assumption that the treatment effect is constant, so it weights observation by the conditional variance of the treatment status.

-Matching does not use cells where there are only treatment observations, whereas OLS takes advantage of the linearity assumption to use all the variables: the treatment group and the control groups may be very dissimilar in matching and in OLS (for example, comparing the CPS to the sample of training program participants in the training program mentioned below means that very different people are compared). Matching will throw away all the control observations for which we cannot find at least one treatment observation with the same characteristics.

Important caveat: Sometimes matching on observables might lead to a greater bias than OLS, if matching is not truly random conditional on observables i.e. matching may not eliminate the omitted variables bias due to unobservables. For instance, suppose we match up people on the basis of family background and attribute any resulting difference in wages to differences in education. It is quite possible that people with the same family background have widely varying ability levels, but very similar levels of schooling. In this case, we would obtain a very large estimate of the returns to schooling, due to the omitted variable bias. This might even be larger than the bias in usual OLS, because in the latter case, we have a greater range of schooling levels with probably the same range of ability levels.

3.2.2 Propensity score matching

Exact matching is not practical when X is continuous or contains many variables. A result due to Rosenbaum and Rubin (1984), is that for $p(X)$ equal to the probability that $T = 1$ given X,

$$E[Y_i^C|X,T] - E[Y_i^C|X,C] = 0$$

implies:

$$E[Y_i^C|p(X),T] - E[Y_i^C|p(X),C] = 0.$$

So it is possible to first estimate the propensity score, and then compare observations which have a similar propensity score. It is often easier to estimate non-parametrically or semi-parametrically the propensity score than to directly condition on observables.

Example: (Dehejia and Wahba 1999), revisiting (Lalonde 1986) on the effect of training on earnings, show that the propensity score matching approach leads to results that are close to the experimental evidence, where the regressions approaches failed. In practice, they first estimated a logit model of training participation on covariates and lags of earnings, and then compared treatment and control in each quintile of the estimated propensity scores. They obtained the final estimate by weighting each difference by the proportion of each trainees in the given quintile.

4 Difference-in-differences type estimators

General references: (Campbell 1969, Meyer 1995).

4.1 Simple Differences

As random experiments are very rare, economists have to rely on actual policy changes to identify the effects of policies on outcomes. These are called “natural experiments” because we take advantage of changes that were not made explicitly to measure the effects of policies.

The key issue when analyzing a natural experiment is to divide the data into a control and treatment group.

The most obvious way to do that is to do a simple difference method using data before ($t = 0$) and after the change ($t = 1$):

$Y_{it} = \alpha + \beta \cdot 1(t = 1) + \epsilon_{it}$

The OLS estimate of β is the difference in means $\bar{Y}_1 - \bar{Y}_0$ before and after the change.

Problem: how to distinguish the policy effect from a secular change?

With 2 periods only, this is impossible. The estimate is unbiased only under the very strong assumption that, absent the policy change, there would have been no change in average Y.

With many years of data, it is possible to develop a more convincing estimation methodology. Suppose that years 0,..,T are available and change took place in year t^*.

Put all the year dummies in the regression:

$$Y_{it} = \alpha + \sum_{\tau=1}^{T} \beta_\tau \cdot 1(t = \tau) + \epsilon_{it}$$

Then $\hat{\beta}_\tau = \bar{Y}_\tau - \bar{Y}_0$

Question: is there a rupture in the pattern of $\hat{\beta}_\tau$ around the reform date $\bar{t}$?

Problems: when the reform is gradual, this strategy is not going to work well.

4.2 Difference-in-differences

A way to improve on the simple difference method is to compare outcomes before and after a policy change for a group affected by the change (Treatment Group) to a group not affected by the change (Control Group). Example: Minimum wage increase in New-Jersey but not in Pennsylvania. Compare employment in the fast food industry before and after the change in both states ((Card and Krueger 1992)).

Alternatively: instead of comparing before and after, it is possible to compare a region where a policy is implemented to a region with no such policy. Example: micro-credit, poor households are eligible to borrow from Grameen Bank. Grameen implements the program only in a subset of villages. (Morduch 1998) compares rich households to poor households in villages where Grameen implements the program and other villages.

The DD Estimate is:

$$DD = [\hat{E}(Y_1|T) - \hat{E}(Y_0|T)] - [\hat{E}(Y_1|C) - \hat{E}(Y_0|C)]$$

The idea is to correct the simple difference before and after for the treatment group by substracting the simple difference for the control group.

DD estimates are often cleanly presented in a 2 by 2 box.

The DD-estimate is an unbiased estimate of the effect of the policy change if, absent the policy change, the average change in $Y_1 - Y_0$ would have been the same for treatment and controls. This is the "parallel trend" assumption.

Regression counterpart. Run OLS on,

$$Y_{it} = \alpha + \beta \cdot 1(t = 1) + \gamma \cdot 1(i \in T) + \eta \cdot 1(t = 1) \times 1(i \in T) + \epsilon_{it}$$

The OLS estimate of η is numerically identical to the DD estimate (the proof of this is similar, though somewhat more complicated, than for the simple difference case).

DD estimates are very common in applied work. Whether or not they are convincing depends on the context and on how close are the control and treatment groups. There are a number of simple checks that one should imperatively do to assess the validity of the DD strategy in each particular case.

- **Checks of DD strategy**

1. Use data for prior periods (say period -1) and redo the DD comparing year 0 and year -1 (assuming there was no policy change between year 0 and year -1). If this placebo DD is non zero, there are good chances that your estimate comparing year 0 and year 1 is biased as well.

   More generally, when many years are available, it is very useful to plot the series of average outcomes for Treatment and Control groups and see whether trends are parallel and whether there is a sudden change just after the reform for the Treatment group.

2. Use an alternative control group C'. If the DD with the alternative control is different from

the DD with the original control C, then the original DD is likely to be biased (cf. (Gruber 1996).

3. Replace Y by another outcome Y' that is not supposed to be affected by the reform. If the DD using Y' is non-zero, then it is likely that the DD for Y is biased as well.

NB: For 1) and 2), it possible to do a DDD strategy. The DDD estimate is the difference between the DD of interest and the placebo DD (that is supposed to be zero).
However, the DDD is of limited interest in general because:

-If the DD placebo is non zero, it will be difficult to convince people that the DDD removes all the bias.

- if the DD placebo is zero, then DD and DDD give the same results but DD is preferable because standard errors are much smaller for DD than for DDD.

(Gruber 1994, Gruber 1996) are neat empirical examples of the use of DD estimators.

Note: The closer are the Treatment and Control groups, the more convincing is the DD approach (note that in the case of a randomized experiment, Treatment and Controls are identical for large sample).
It is often useful to perform simple differences between Treatment and Controls along covariates (such as age, race, income, education, ...) to see whether Treatment and Controls differ systematically.
In the regression framework, it is useful to throw covariates interacted with the time dummy to control for changes in the composition of controls and treatment groups.

- **Common Problems with DD estimates**

- Targeting based on differences

A pre-condition of the validity of the DD assumption is that the program is not implemented based on the pre-existing differences in outcomes. Example:

- "Ashenfelter dip": It was common to compare wage gains among participants and non participants in training programs to evaluate the effect of training on earnings. (Ashenfelter and Card 1985) note that training participants often experience a dip in earnings just before they enter the program (which is presumably why they *did* enter the program in the first place). Since wages have a natural tendency to mean reversion, this leads to an upward bias of the DD estimtate of the program effect.
- In the case of difference-in-differences that combine regional and eligibility variation: Often the regional targeting is based upon the situation of the group of eligible people (e.g. Grameen will locate a bank in the villages where the *poor* are worse off. It is easy to check that this will lead to negative difference-in differences in the absence of the program, if villages differ in terms of distribution of wealth.

- Functional form dependence:

  When average levels of the outcome Y are very different for controls and treatments before the policy change, the magnitude or even sign of the DD effect is very sensitive to the functional form posited.

  Illustration: Suppose you look at the effect of a training program targeted to the young.

  The unemployment level for the young decreases from 30% to 20%.

  The unemployment level for the old decreases from 10% to 5%.

  Because of the dramatic difference in pre-program unemployment levels (30% vs 10%), it is difficult to assess whether the program was effective.

  The DD in levels would be $(30 - 20) - (10 - 5) = 10 - 5 = 5\%$ suggesting a positive effect of training on employment.

  However, if you consider log changes in unemployment, the DD becomes,

  $[\log(30) - \log(20)] - [\log(10) - \log(5)] = \log(1.5) - \log(2) < 0$

  suggesting that training had a negative effect on employment.

- Long-term response versus reliability trade-off:

DD estimates are more reliable when you compare outcomes just before and just after the policy change because the identifying assumption (parallel trends) is more likely to hold over a short time-window. With a long time window, many other things are likely to happen and confound the policy change effect.

However, for policy purposes, it is often more interesting to know the medium or long term effect of a policy change.

In any case, one must be very cautious to extrapolate short-term responses to long-term responses (see literature on labor supply or taxable income elasticities).

- Heterogeneous behavioral responses:

  When both control and treatment groups experience a change but of different size it is still possible to do a DD estimate. However, the DD estimate might be meaningless if the intensity of behavioral responses for Treatments and Controls is different.

  Simple illustration: effect of M_{it} (for example marginal tax rate) on outcome Y_{it} (taxable income). Assume the model is:

  $Y_{it} = \mu_i + \alpha_t + \eta_i M_{it}$

  For treatment individuals, M_{it} increased from 0 to M_T.

  For control individuals, M_{it} increased from 0 to M_C.

$$DD = [Y_1^T - Y_0^T] - [Y_1^C - Y_0^C] = \eta_T \cdot M_T - \eta_C \cdot M_C$$

  If $\eta_C = 2\eta_T$ and $M_T = 2M_C$ then DD is zero even though both η's can be large and positive.

  This issue arises for example in (Feldstein 1995) on taxable income elasticities.

- Inference

  The observations in the control and the treatment group may tend to move together over time. In other words, there may be a common random effect at the time*group level. In this case, the standard error of the estimator should take into account this correlation: we have in effect less information than we think.

  Example: Suppose that the outcome can be described by the equations:

$$y_{it} = \beta_T + \gamma_1 \text{Post} + \alpha_{Tt} + \epsilon_{it},$$

if i belongs to the treatment group.

$$y_{it} = \beta_C + \alpha_{Ct} + \epsilon_{it},$$

if i belongs to the control group, where α_{Tt} and α_{Ct} are random group effects (not necessarily i.i.d).

The variance of the difference in difference estimator should take into account the variance of $\alpha_{Tt} - \alpha_{Ct}$: the variance covariance matrix of the error term is block diagonal. The “standard” OLS variance does not take it into account. With only 2 periods and 1 “treatment”, one “control” group, there is nothing we can do to adjust the standard error of the DD estimator: DD is unbiased, but not consistent. With several periods, we can use the pre-treatment periods to calculate the variance of $\alpha_{Tt} - \alpha_{Ct}$, and adjust the standard error for it.

This problem is described in general terms in Moulton (1986), and for the case of the DD specifically in Lang(2000). Stata offers a correction of the standard error with the command “cluster”. However, this command runs into trouble when the number of clusters is small. Using the formula in Moulton seems to be safer, but it needs to be programmed.

4.3 Fixed Effects

Fixed effects can be seen as a generalization of DD in the case of more than two periods (say S periods) and more than 2 groups (say J groups).

Suppose that group j in year t experiences a given policy T (for example an income tax rate) of intensity T_{jt}. We want to know the effect of T on an outcome Y.

OLS Regression: $Y_{jt} = \alpha + \beta T_{jt} + \epsilon_{jt}$

With no fixed effects, the estimate of β is biased if treatment T_{jt} is correlated with ϵ_{jt} (that is, correlated with the outcome Y_{jt} even if the treatment T_{jt} were all identical across time and groups. This is often the case in practice: for example if T_{jt} is generosity of welfare in state j and year t and Y_{jt} is unemployment, the simple OLS estimate is likely to be biased downward if poorer states with high unemployment levels have less generous benefits.

A way to solve this problem is to put time dummies and group dummies in the regression:

$Y_{jt} = \alpha + \gamma_t + \delta_j + \beta T_{jt} + \epsilon_{jt}$

Then identification is obtained out of within group time variation: group specific changes over time. This is a direct extension of DD where there are 2 groups that experience different changes in policy over 2 periods.

Note that changes common to all groups are captured by the time dummies and thus are not a source of variation that identifies β.

The problems and short-coming of Fixed effects are basically the same as DD.

The big advantage relative to DD is that many changes and years can be pooled in a single regression producing more precise and robust results.

However, the disadvantage is that fixed effects is a black-box regression and it is more difficult to check visually trends as can be done with a single change.

Another common criticism of fixed effects is that state policy reforms may respond to trends in outcomes Y (example: increase generosity of welfare benefits when economy is not doing well) and thus produce a spurious correlation even when one controls with time and year dummies.

Fixed effects are valid only if the response is immediate. If full responses take more than 1 period, the fixed effects estimate might be biased because the true model should include lagged variables $T_{j,t-1}$.

5 Instrumental Variable (IV) methodology

5.1 Basics

We know that the OLS regression $Y = X\beta + \epsilon$ is biased when ϵ is correlated with X. A way to get around this issue is to use an instrument Z for X. An instrument Z is set of P variables. P must be equal or larger than K the number of variables in X.

The IV formula is given by

$$\hat{\beta}_{IV} = (X'P_Z X)^{-1}(X'P_Z Y)$$

where $P_Z = Z(Z'Z)^{-1}Z'$

NB: when the number of instruments is exactly equal to the number of independent variables

$(P = K)$, the formula reduces to:

$\hat{\beta}_{IV} = (Z'X)^{-1}(Z'Y)$

$\hat{\beta}_{IV}$ is consistent when Z satisfies two conditions:

1) Z is uncorrelated with ϵ

2) Z is correlated with X ($Z'X$ is of rank K).

Stata command: regress y x1 .. xK (z1 .. zP)

where x1 .. xK is the list of dependent variables and z1 .. zP is the list of instruments.

Note that is general, we are interested by the coefficient on one variable X only (say $x1$) and we are confident that the other controls x2 .. xK are not correlated with ϵ. In that case, x2 .. xK can be used as instruments and we need find only one extra instrument z.

The stata command in that case is: regress y x1 x2 .. xK (z x2 .. xK)

The spirit of OLS is to compare outcomes Y for high X vs low X.

The spirit of IV is to compare outcomes Y for high Z vs low Z. The regression of the outcome Y on the instruments Z is called the *reduced form*.

To understand this clearly, it is useful to consider the case of a single variable X and a single binary instrument Z. For example, X is variable indicating whether you have served in the military during the Vietnam era ((Angrist 1990)), Z is a variable indicating whether you had a high lottery number or a low lottery number in the lottery draft, and Y are your earnings after the war.

In that case, simple computations of the type we did for simple differences shows that:

$$\hat{\beta}_{IV} = \frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{E}(X|Z=1) - \hat{E}(X|Z=0)}$$

That is, the IV estimate is the ratio of the difference of means of the outcome Y for the group $Z = 1$ and the group $Z = 0$ to the difference of means of the variable X for the group $Z = 1$ and the group $Z = 0$. This is the Wald estimator, a very transparent IV estimator.

Good instruments are:

- Strongly correlated with X: $E(X|Z)$ varies a lot with Z. This correlation is checked by the First-stage: regress X on Z.

  $X = Z\gamma + \nu$

  γ has to be non-zero and significant, otherwise the instrument is weak and standard errors for β will be large.

- Uncorrelated with Y beyond the direct effect through X (in other words can be excluded from the equation $Y = X\beta + \epsilon$, that is, is not correlated with ϵ). That cannot be tested and has to be assessed on a case by case basis. When there are more instruments than columns in X, two tests can be used:

  - An overidentification test, which in essence compares all the IV obtained from using different subsets of instruments, and tests whether they are the same.
  - A Hausman test, when you trust an instrument, and comparing the results obtained with only this instrument against the results obtained using the whole set of instruments.

  These tests are useful, but have two problems:

  - They may reject if the treatment effect is heterogenous, and the instruments exploit variation at different parts of the treatment response function (cf. below on the interpretation of IV).
  - Their power is not very strong and they tend to accept too often.

5.2 Where to find instruments?

Instruments do not fall from the sky. Because it is difficult to test the validity of the instruments, you need to be convinced *on a priori grounds* that they are valid. Good instruments are usually generated by real or natural experiments.

Examples:

- Random encouragement designs: These are cases where the *probability* that someone receives a treatment varies randomly across people. The actual treatment status may then result from a choice, and then be endogenous.

  - Vietnam era draft lottery ((Angrist 1990)): a high lottery number makes it more likely that someone is drafted, but he can still dodge the draft if he has a high number, or enroll voluntarily if he has a low number.
  - To test the effect of flu vaccine on flu ((Imbens, K.Hirano, D.Rubin and A.Zhou 2000)). A random encouragement design was done. A letter reminding doctors to propose a flu vaccine to their clients was randomly sent to a set of doctors. The instrument (the letter) is randomly assigned, but not the treatment (flu vaccine).

- Instruments trying to approximate a random encouragement design:

  - Distance to hospital with operating facilities as an instrument for surgery in heart attacks.
  - Distance to school as an instrument for schooling These instruments must be evaluated carefully.

- Policy reforms etc...

  An instrument can be formed by interacting two variables, for example a time and group. We are then using a DD as the first stage of the relationship. In the second stage, we control for the two uninteracted variables.

  For example, consider the school experiment in (Duflo 2000). There are two types of regions (High H and Low L program regions) and two types of cohorts (Young Y and Old O). The program affected mostly the education of young cohorts in the high program regions. Assume that the program affected the wage of the individuals only through its effects on education.

  The difference in differences estimator for the effect of the program on education S is:

$$(E[S|H,Y] - E[S|H,O]) - (E[S|L,Y] - E[S|L,O])$$

  The difference in differences estimator for the effect of the program on wages W is:

$$(E[W|H,Y] - E[W|H,O]) - (E[W|L,Y] - E[W|L,O])$$

The effect of education on wages can be obtained by taking the ratio of the two DD. This is the Wald estimator:

$$\frac{E[W|H,Y]-E[W|H,O])-(E[W|L,Y]-E[W|L,O]}{E[S|H,Y]-E[S|H,O])-(E[S|L,Y]-E[S|L,O]}$$

The corresponding regression would be:

$$W=\alpha+\beta Y+\gamma H+\delta S+\epsilon$$

where H is a dummy equal to 1 in the high program region, Y is a dummy equal to 1 for the young and S is instrumented with the interaction $H\times Y$.

5.3 Problems with IV

1. IV can be very biased (much more than OLS).

   Suppose our instrument is not truly exogenous i.e. $Cov(Z,\epsilon)\neq 0$. Consider the example of difference in wages(Y) due to serving in the Vietnam War(X), using the draft lottery number(Z) as an instrument. We know that the OLS estimator $E(Y|X=1)-E(Y|X=0)$ is biased, because serving in the army is likely to be correlated with lots of unobserved characteristics. For the IV Wald estimator, the denominator represents the difference in the probability of serving in the army for people with high and low lottery numbers i.e. this number is less than 1. Suppose in fact the the draft lottery number were not random, then $E(Y|Z=1)-E(Y|Z=0)$ is a biased estimate of the reduced form impact of lottery number on wages. Notice now that even if the bias in the reduced form is of the same order of magnitude as the bias of OLS, the IV estimate as a whole is much *more* biased, because the denominator is less than one.

   If the instrument is strong i.e. a very good predictor of army service, then the denominator is closer to $(1-0)$, and hence this bias due to the violation of the exclusion restriction is less.

2. Even instruments that are randomly assigned can be invalid.

   What is needed is that they don't affect the outcome directly. Examples:

   - Draft Lottery and Military service ((Angrist 1990):

A low number could encourage someone to stay in college to evade the draft, thereby increasing its earning directly.

- Flu vaccine:

  The letter sent to the doctor seems to have convinced them to take other steps to prevent the flu ((Imbens et al. 2000)). It therefore had a direct effect on flu, not due to the shot per se. The IV using the letter as instrument would be an overestimate.

3. How representative is the IV answer?

   Notice that the IV estimator is the ratio of the change in Y due to change in Z to the change in X due to change in Z, and we are assuming that a lower draft lottery number makes army service more likely, not less. We can partition all our sample units into the following categories: those for whom the lottery number makes a difference to the army service decision and those for whom it doesn't (this includes those who would have volunteered anyway, and those who would have avoided the draft irrespective of their lottery number). Then the change in X due to change in Z is non-zero only for the first group (the "compliers") and so the IV estimate represents the impact of army service on wages *only for this group.* This is called the Local Average Treatment Effect i.e. the impact of the treatment (army service) only for the group affected by the instrument(see (Angrist and Imbens 1994, Angrist, Imbens and B.Rubin 1996, Angrist and Krueger 1999) for detailed explanations of this). If we assume that the impact of army service on wages is the same for every individual in the population ("constant treatment effect") then this IV estimate represents a population average. However, if the impact of army service is different for "non-compliers", then we must be careful while extrapolating IV estimates to the whole population.

4. Specification searching and publication bias.

   Papers with T statistic above 2 are more likely to be published. IV have larger standard error than OLS, therefore they also need larger point estimates to be significant. Reported IV will therefore have a natural tendency to be "too high". This is Ashenfelter, Harmon and Oosterbeek (1999) explanation for why IV returns to education tend to be higher than OLS.

5.4 Getting Instruments from Theoretical Models

In many papers, authors write down theoretical models which generate instruments. For example, (Strauss 1986) writes down a model of the effect of food on productivity. Price of food is negatively correlated with food quantity. The model is written such that price of food is also uncorrelated with productivity besides the effect on food intake.

This strategy known as structural model estimation produces a framework that is complete (theoretical model and data application) and estimates that are fully meaningful in the context of the model. However, these estimates are valid only to the extent that the structural model is valid.

6 Regression Discontinuity Design

References

[1] The important reference is (Campbell 1969). See (Angrist and Lavy 1999, Van der Klauw 1996) for convincing applications.

- RDD can be used when the treatment is a discontinuous function of an underlying continuous variable. Examples:
  - Grameen bank eligibility rule: eligible if households owns less then 0.5 hectares.
  - Financial aid at NYU for college studies: step function of an index (grades in highschool, SAT scores, income of parents ...).
  - Maimonides rule for class size in Israel: extra teacher added as soon as the number of pupils in class reaches multiple of 40 students.
- When this rule is followed at least approximately, it means that two people with very close characteristics will be exposed to different treatments.
- Idea of RD: compare outcome for people whose value of the underlying targeting variable is just below and just above the discontinuity.

Formally: Imagine first that treatment rule is based on some number X and that the treatment rule is:
- $T = 1$ if $X \geq \overline{X}$
- $T = 0$ if $X < \overline{X}$

Then with a large sample, you would compute (for some ϵ):

$$E[Y|\overline{X} \leq X < \overline{X} + \epsilon] - E[Y|\overline{X} - \epsilon \leq X < \overline{X} =$$
$$E[Y^T|T, \overline{X} \leq X < \overline{X} + \epsilon] - E[Y^C|C, \overline{X} - \epsilon \leq X < \overline{X}]$$

The assumption is that as the ϵ goes to 0, the difference between the two groups in the absence of the treatment shrinks to 0.

More realistically, the rule increase the *probability* that someone will be treated.
- Not everybody with $X \geq \overline{X}$ will be treated (for example, some people may not have asked for financial aid, even though they would qualify for it.
- Some people with $X < \overline{X}$ will not be treated (for example, some schools with less than 40 students still get a second teacher).

Formally:
- $P(T = 1) = p_1$ if $X >= \overline{X}$
- $P(T = 0) = p_0$ if $X < \overline{X}$, with $p_1 > p_0$.

We can again calculate the difference in outcome between individuals just above and just below $\overline{X}$.

$$E[Y|\overline{X} \leq X < \overline{X} + \epsilon] - E[Y|\overline{X} - \epsilon \leq X < \overline{X}]$$

Under the same assumption as before (that for ϵ small enough, the outcomes in the absence of treatment would be the same in the two groups), we can attribute this difference to the difference in the probability of treatment. But now, there are some treated people and some control people on both sides of $\overline{X}$. To obtain the effect of the treatment, we must "scale up" the difference, by dividing between the difference in the probability of treatment between the two groups.

$$\frac{E[Y|\overline{X} \leq X < \overline{X} + \epsilon] - E[Y|\overline{X} - \epsilon \leq X < \overline{X}]}{E[T|\overline{X} \leq X < \overline{X} + \epsilon] - E[T|\overline{X} - \epsilon \leq X < \overline{X}]}$$

The relationship between this and IV should be clear: This is the Wald estimate (which we derived above), using a dummy for $X \geq \overline{X}$ as instrument for the treatment status. This regression-discontinuity Wald estimator is numerically identical to a non-parametric kernel estimator with a uniform kernel. Under this interpretation, this estimator would be valid even if the IV assumptions were violated. However, it would be asymptotically biased and we would need to use a slightly more complicated non-parametic estimator to reduce the bias ((Hahn, Todd and der Klaauw 2001)).

Researchers have exploited this to construct "IV versions" of the RD estimator:

We start with a model:

$$Y = \alpha T + g(X) + \epsilon,$$

where $g(X)$ is a set of smooth functions of X (polynomials, splines, etc...) (thus controlling for the dependence of Y on X).

A dummy for $X > \overline{X}$ can then be used as instrument for receiving the treatment, in a regular 2SLS strategy

Cautionary remarks:

- It is important to check in the data that there is actually a discontinuity in the probability of being treated at the expected point $\overline{X}$. Example: In the Grameen case, (Morduch 1999) shows that people with more than 0.5 hectares of land are as likely than other people to get credit. The first step should be to regress non-parametrically the treatment variable on the variable X, and check whether the discontinuity is actually present in the data.
- In developing countries, even strong rules are rarely followed to the letter... Fancy means testing procedure are unlikely to generate RD that can be exploited in practice.

- Large sample is required, since you will be exploiting only variation coming from individuals around $\overline{X}$.

7 The measurement error problem

7.1 Classical measurement error

Assume that you want to estimate the relationship

$$y_i^* = \beta x_i^* + \epsilon_i,$$

for $i = 1$ to N, where, for example y_i^* is log calories per capita (after having taken out the mean) and x_i^* is log of long run resources per capita.

However, the true y_i^* and the true x_i^* are both unobserved. What you observe are proxies of these measures, i.e. the true variables, measured with error (For example: it is quite difficult to know what people really eat: they eat food out of the home, there is wastage,.... It is also difficult to know people's long run resources. What we observe in a survey is people's current income, which can vary much more).

We model measurement error in the following way. We observe y_i and x_i, which are the true variables, plus some noise.

$$y_i = y_i^* + \nu_i,$$

$$x_i = x_i^* + \upsilon_i,$$

In the "classical" measurement error case, the assumption is that measurement errors are uncorrelated with the truth, and with each other (we will see what happens if we relax this assumption).

So: $E[\nu_i y_i^*] = E[\upsilon_i x_i^*] = E[\nu_i x_i^*] = E[\upsilon_i y_i^*] = E[\upsilon_i \nu_i] = 0$

Obviously, we will also assume that the model is otherwise correctly specified: $E[\epsilon_i x_i^*] = 0$.

Let us rewrite the model in terms of the variable we actually observe:

$$y_i = \beta x_i + (\epsilon_i + \nu_i - \beta \upsilon_i)$$

The source of the problem is that the new error term $w_i = \epsilon_i + \nu_i - \beta \upsilon_i$ is now not uncorrelated with x_i.

To see this, let us express the OLS estimator of β in the observed equation:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} = \frac{\sum_{i=1}^N (x_i^* + \upsilon_i)(\beta x_i^* + \epsilon_i - \nu_i)}{\sum_{i=1}^N (x_i^* + \upsilon_i)(x_i^* + \upsilon_i)}$$

We want to know the probability limit of $\hat{\beta}_{OLS}$ as $N \to \infty$. With our assumptions we obtain:

$$\text{Plim}(\hat{\beta}_{OLS}) = \text{Plim}\left(\beta \frac{\frac{1}{N}\sum_{i=1}^N x_i^{*2}}{\frac{1}{N}\sum_{i=1}^N x_i^{*2} + \upsilon_i^2}\right) = \beta \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_\upsilon^2},$$

where for a random variable x, $\sigma_x^2 = \text{Plim}\left(\frac{\sum_{i=1}^N x_i^2}{N}\right)$ is the variance of x.

Note:

1. There is an attenuation bias: the estimated coefficient is smaller than the true coefficient (the ‘Iron law of econometrics’).
2. The measurement error in y does not lead to attenuation bias, in the uncorrelated case.
3. The larger the variance of the error term relative to the variance of the underlying variable (the ‘signal to noise ratio’), the larger the attenuation bias.

7.2 The problem of measurement error with fixed effects

Imagine you now have the relationship:

$$y^*_{it} = \beta x^*_{it} + \epsilon_{it},$$

with $\epsilon_{it} = \omega_i + \xi_{it}$. You are worried that there is a correlation between ω_i and x_{it} which would lead to a bias in OLS estimation of this equation. If you have two years of data, you might think of taking first differences:

$$y^*_{i2} - y^*_{i1} = \beta(x^*_{i2} - x^*_{i1}) + \xi_{i2} - \xi_{i1}$$

which we rewrite:

$$\Delta y^*_{it} = \beta \Delta x^*_{it} + \Delta \xi_{it},$$

The fixed effect has now disappeared, so we have solved this problem. However, the measurement problem is still here. The probability limit of the first OLS estimate of β in the first difference equation (assuming uncorrelated measurement error) is:

$$\text{Plim}\hat{\beta}_{FD} = \beta \frac{\sigma^2_{\Delta x^*}}{\sigma^2_{\Delta x^*} + \sigma^2_{\Delta \upsilon}}$$

If measurement errors are independent measurement error in each period (an extreme case), then $\sigma^2_{\Delta \upsilon} = 2 * \sigma^2_{\upsilon}$. However, x^* is presumably strongly autocorrelated, so $\sigma^2_{\Delta x^*} < \sigma^2_{x^*}$. Therefore the attenuation bias is stronger in fixed effect. In fact, it can become really large if the underlying variable does not move very much over time but there is measurement error in every period.

7.3 Instrumental variables to solve the measurement error problem

Coming back to the case of a single cross-section, assume that you have another, independent measure of x^*, possibly noisy as well. For example, you ask food expenditure to each spouses in the family (while the other spouse is not in the room).

$$z_i = x_i^* + \mu_i,$$

with $E[\mu_i x_i^*] = E[\mu_i \upsilon_i] = 0$.

The instrumental variable estimator of β is:

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i z_i} = \frac{\sum_{i=1}^N (x_i^* + \upsilon_i)(\beta x_i^* + \epsilon_i + \nu_i)}{\sum_{i=1}^N (x_i^* + \upsilon_i)(x_i^* + \mu_i)}$$

$$\text{Plim}\hat{\beta}_{IV} = \beta$$

So the IV estimator is consistent.

7.4 Non-classical measurement error

This can occur due to various reasons:

1. Measurement error is correlated with the underlying variables X. In this case, there is not necessarily an attenuation bias, and error in the measurement of y^* can also lead to biased estimates. For example, assume that $E[\nu_i x_i^*] = \sigma_{\nu x^*} \neq 0$. For example, there would be a positive correlation between the measurement error in calorie intake and income if calorie intakes tend to be overestimated for high income households (because they waste more) and underestimated for low income households. Then the probability limit of the OLS estimator becomes:

$$\text{plim}(\hat{\beta}_{OLS}) = \beta \frac{\sigma_{x^*}^2 + \sigma_{\nu x^*}}{\sigma_{x^*}^2 + \sigma_{\upsilon}^2},$$

   The bias depend on how large $\sigma_{\nu x^*}$ is.

2. Our regressors are categorical variables e.g. years of schooling (discrete values) or dummy for high-school graduate. In this case, the lowest category cannot under-report and the highest cannot over-report, which means that the distribution of the measurement error is related to the value of the regressor, thus violating the classical assumptions. In the case of only two categories, the OLS estimates are biased downwards, but two-stage IV estimates are biased *upwards.* In the case of only two schooling categories (0 and 1) and two measurements S_1 and S_2 of the true schooling level S^*, we can use S_2 as an instrument for S_1. In this case, Kane et. al. derive the probability limit of the 2SLS estimator to be

   $$\text{plim}(\hat{\beta}_{2SLS}) = \beta \frac{1}{1-(\alpha_1+\alpha_2)}$$

   where $\alpha_1 = Pr(S_1 = 0|S^* = 1), \alpha_2 = Pr(S_1 = 1|S^* = 0)$. Since the denominator is less than 1, the IV estimator is biased upwards.

   In the general case of multiple categories, we cannot even determine the direction of bias. Different ways of solving these problems include estimating the extent of measurement error by using a validation data set (Pishke 1995), or putting restrictions on the form of the measurement error (Card 1996) or estimating the extent of measurement error from the data by using the presence of two measures of the regressor (Kane et.al. 1999).

3. Suppose people do not report the mismeasured x and y, but instead are aware of the possibility of mismeasurement and report their best estimate of x^* and y^*, based upon the observed x and y. For instance, if people are asked how much food they buy in a month and they report a monthly figure based on last week's consumption. Or people are asked their income levels, and they report their best estimate of it, which may not include some components like interest income or capital gains. The best estimate $\tilde{x}$ is $E[x^*|x]$, which for our linear model would be a linear combination of the observed x and μ_x, the unconditional mean of x (and similarly for y). Crucially, the measurement error between the reported value and the true value $(\tilde{x} - x^*)$ would now be uncorrelated with the reported $\tilde{x}$ (property of conditional expectation), so that measurement error in the regressor *does not* lead to a downward bias in OLS. Further, using IV in such a situation would result in an upward-biased estimate.

   On the other hand, the reported $\tilde{y}$ is $E[y^*|y] = \lambda y + (1-\lambda)\mu_y \Rightarrow Cov(\tilde{y}, x^*) = \lambda Cov(y, x^*) = \lambda Cov(y^*, x^*) \Rightarrow$ our OLS estimates are biased downwards when there is measurement error in y, but not biased if there is measurement error in x. This is the reverse of the results with classical measurement error! Hyslop and Imbens (2000) also consider the case when the respondents report their best estimate $\tilde{x}$ taking into account both the observed y and the observed x, and find that measurement error can even lead to *upward* biases in OLS.

References

Altonji, Joseph, Todd E. Elder, and Christopher R. Taber (2000) ‘Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools.’ NBER Working Paper No. W7831

Angrist, Joshua D. (1990) ‘Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security Administrative records.’ *American Economic Review* 80(3), 313–336

____ (1998) ‘Estimating the labor market impact of voluntary military service using social security data on military applicants.’ *Econometrica* 66(2), 249–88

Angrist, Joshua D., and Alan B. Krueger (1999) ‘Empirical strategies in labor economics.’ Forthcoming Handbook of Labor Economics

Angrist, Joshua D., and Guido Imbens (1994) ‘Identification and estimation of local average treatment effects.’ *Econometrica* 62(2), 467–475

Angrist, Joshua D., and Victor Lavy (1999) ‘Using Maimonides’ rule to estimate the effect of class size on scholastic achievement.’ *Quarterly Journal of Economics* 114(2), 533–575

Angrist, Joshua D., Guido W. Imbens, and Donald B.Rubin (1996) ‘Identification of causal effects using instrumental variables.’ *Journal of the American Statistical Association* 91(434), 444–455

Ashenfelter, Orley, and David Card (1985) ‘Using the longitudinal structure of earnings to estimate the effect of training programs.’ *Review of Economics and Statistics* 67(4), 648–60

Ashenfelter, Orley, and Mark W. Plant (1990) ‘Nonparametric estimates of the labor-supply effects of negative income tax programs.’ *Journal of Labor Economics* 8(1), 396–415

Ashenfelter, Orley, Colm P. Harmon, and Hessel Oosterbeek (1999) ‘A review of estimates of the schooling/earnings relationship.’ *Labour Economics* 6(4), 453–470

Campbell, Donald T. (1969) ‘Reforms as experiments.’ *American Psychologist* 24, 407–429

Card, David, and Alan Krueger (1992) ‘Does school quality matter? Returns to education and the characteristics of public schools in the United States.’ *Journal of Political Economy* 100(1), 1–40

Deaton, Angus (1997) *The Analysis of Household Surveys* (World Bank, International Bank for Reconstruction and Development)

Dehejia, Rajeev H, and Sadek Wahba (1999) 'Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs.' *Journal of the American Statistical Association* 94(448), 1053–62

Duflo, Esther (2000) 'Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment.' Working Paper 7860, National Bureau of Economic Research, August

Feldstein, Martin (1995) 'The effect of marginal tax rates on taxable income: A panel study of the 1986 tax reform act.' *Journal of Political Economy* 103(3), 551–72

Gruber, Jonathon (1994) 'The incidence of mandated maternity benefits.' *American Economic Review* 84(3), 622–641

—— (1996) 'Cash welfare as a consumption smoothing mechanism for single mothers.' Working Paper 5738, National Bureau of Economic Research

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001) 'Identification and estimation of treatment effects with a regression-discontinuity design.' *Econometrica* 69(1), 201–209

Hausman, Jerry A, and David A Wise (1979) 'Attrition bias in experimental and panel data: The gary income maintenance experiment.' *Econometrica* 47(2), 455–73

Imbens, Guido, K.Hirano, D.Rubin, and A.Zhou (2000) 'Estimating the effect of flu shots in a randomized encouragement design.' *Biostatistics* 1(1), 69–88

Kremer, Michael, Paul Glewwe, and Sylvie Moulin (1998) 'Textbooks and test scores: Evidence from a prospective evaluation in Kenya.' Mimeo, Harvard

Krueger, Alan (2000) 'Experimental estimates of education production functions.' forthcoming, Quarterly Journal of Economics

Lalonde, Robert J. (1986) 'Evaluating the econometric evaluations of training programs using experimental data.' *American Economic Review* 76(4), 602–620

Mayer, Susan E. (1999) ‘How did the increase in economic inequality between 1970 and 1990 affect poor american children’s educational attainment?’ Mimeo

Meyer, Bruce D. (1995) ‘Natural and quasi-experiments in economics.’ *Journal of Business and Economic Statistics* 13(2), 151–161

Morduch, Jonathan (1998) ‘Does microfinance really help the poor? new evidence from flagship programs in bangladesh.’ Mimeo

——— (1999) ‘The microfinance promise.’ Forthcoming,*Journal of Economic Literature*

Pencavel, John (1986) ‘Labor supply of men.’ In *Handbook of Labor Economics,* ed. Orley Ashenfelter and Richard Layard (Elsevier Science)

Rosenbaum, Paul, and Donald B. Rubin (1984) ‘Estimating the effects caused by treatments: Comment [on the nature and discovery of structure].’ *Journal of the American Statistical Association* 79(385), 26–28

Rosenbaum, Paul R. (1995) ‘Observational studies.’ In ‘Series in Statistics’ (New York: Heidelberg and London: Springer)

Sacerdote, Bruce (2000) ‘Peer effects with random assignment: Results for Dartmouth roommates.’ Working Paper 7469, National Bureau of Economic Research

Strauss, John (1986) ‘Does better nutrition raise farm productivity?’ *Journal of Political Economy* 9(2), 297–320

Van der Klauw, Wilbert (1996) ‘A regression-discontinuity evaluation of the effect of financial aid offers on college enrollment.’ Mimeo, New York University, Department of Economics