Department of Economics Fall 2016

Econ 536
Lecture 11

# Binary Treatment Models, Randomization and Errors in Variables

## 1. Binary Treatment and Randomization

The simplest experimental treatment model is the following

$$y_i = \alpha + \beta D_i + u_i$$

where $D_i$ is 1 if the subject is "treated", and 0 if the subject is a control. In this model the least squares estimator of $\beta$ is,

$$\hat{\beta} = \bar{y}_1 - \bar{y}_0$$

and

$$\hat{\alpha} = \bar{y}_0$$

Why? If, as is common, the response $y_i$ is really a *change* in something after a treatment is completed, then we have instead

$$\Delta y_i = \alpha + \beta D_i + u_i$$

and

$$\hat{\beta} = \overline{\Delta y_1} - \overline{\Delta y_0}$$

This is the beloved diff-in-diff model.

*A Case Study* A classical example is the Lanarkshire milk experiment described by Student (1931). In an effort to improve nutrition for elementary school children in a relatively poor region of Scotland an experiment was undertaken to provide milk in schools. The intention was to *randomly select* between 200-400 kids in each of 67 schools, of which half would get milk each day; the other half would not. Evaluation of the effectiveness of the "treatment" was exactly the diff-in-diff strategy which would be done as a t-test. The response, y, was change in weight.

What went wrong? Teachers decided who got the milk and presumably gave the milk to the poorer, smaller "more deserving" kids. We can check this by noting with randomization the treated and control kids would have the same initial weight but they didn't; treated kids were lighter by approximately 3 months growth, and shorter by 4 months growth in height. Since the initial weighing occurred in February and the final weighing in June, and children were weighed with their clothes on, the real weight response is confounded with the change in the weight of the clothes. Again, if the randomization were done properly this would not be a problem, a source of additional variability of course, but not of bias. As it was, it is a serious bias consideration. *Could this be corrected?* Not really after the fact. Student suggests using a smaller trial with only twins, in a future experiment.

*The Wald Estimator*

In many instances of the treatment-control experiment, there is randomization in what has been called "intention to treat," but often there cannot be any way to force people to accept the treatment. So we have to distinguish *compliance* from *intent to treat*. In the simplest setting this gives rise to a simple form of the IV estimator. Suppose $x_i$ is actual treatment/control as before and $z_i$ is the intent to treat variable, then in our simple original setup we can use the Wald Estimator. The simplest way to obtain the Wald estimator is to consider the model

$$y_i = \alpha + \beta x_i + u_i$$

Suppose $Ez_i u_i = 0$ so we have the moment equations, recalling that $z_i$ is binary,

$$E(y_i|z_i = 1) = \alpha + \beta E(x_i|z_i = 1)$$
$$E(y_i|z_i = 0) = \alpha + \beta E(x_i|z_i = 0)$$

now subtract one from the other to obtain.

$$\beta = \frac{E(y|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$$

so a natural estimator would replace these population quantities by their sample analogues. This is the Ur-iv estimator. Angrist calls it the mother of all IV estimators. In some heuristic sense we "see" the relationship between $y$ and $x$ "through the looking glass" as reflected by the IV $z_i$. When $x_i$ is binary, say $D_i$ to use our prior notation, then $E(D_i|z_i = j) = \Pr(D_i = 1|z_i = j) \equiv \pi_j$ for $j = 0, 1$, so the denominator is the difference in these probabilities. Note that, focusing on the denominator, we might expect that in many situations that the term $E(x_i|z_i = 0)$ would be zero, since subject who aren't "intended to be treated" may find it difficult to *be* treated. On the other hand, $E(x_i|z_i = 1)$ is generally likely to be somewhat less than one, since some of those randomized into the treatment may decide that they don't want to be treated. In the extreme case that the proposed IV $z_i$ doesn't impact the mean of mean of the $x_i$'s, then we have a classical failure of the IV strategy and division by zero.

Returning to the pure randomization model for a moment, there is often, even in well randomized experiments, a temptation to include other covariates in the model, e.g.

$$y_i = \alpha + x_i'\beta + \delta D_i + u_i$$

so $D_i$ is an randomized treatment indicator and $x_i$ denotes a vector of other variables. Now, the randomization implies that

$$x_i \perp D_i$$

and this assumption can be checked. (This is usually done by computing conditional means of the $x$'s with respect to $D$.) What is the advantage of

including the additional covariates? We know that given their orthogonality with $D$ that they shouldn't change our $\hat{\delta}$, so why bother?

The usual answer to this question, exemplified by Gertler (2004) is that including $x_i$'s "improves the power of the estimates". Gertler is analyzing the effect of PROGRESSA the conditional cash transfer program in Mexico. In many respects this program is like the Lanarkshire milk experiment except that cash is distributed directly to households according to a randomized scheme. But children's heights and weights are still the principle measures of program effect. What does "improves the power of the estimates" mean? Presumably, it means "reduces their standard errors". Since $D \perp x$ this has nothing to do with $X'X$, but only with $\hat{\sigma}^2$. Clearly if $x$'s are effective in "explaining" $y$, then their inclusion will reduce $\hat{\sigma}^2$ and thereby reduce the standard error of $\hat{\delta}$. What's not to like about this?

The case against including covariates is laid out nicely in Freedman (2009). He argues that the presumption that the linear specification is a good approximation can be dangerous. Freedman adopts what he calls the Neymann (1923) model. It seems to be a precursor of what is now usually called the Rubin "potential outcomes" model. We have a response variable $y$ and several treatment levels, individual subjects are assigned, in the simplest case, to one and only one of the treatment options. Each individual has a potential outcome associated with each of the treatments, but we only observe one of these, for the treatment that is actually assigned. We would like to estimate the "average" treatment effect for each of the treatments, or alternatively the differential treatment effects, treatment level $i$'s average response minus, say the average response under the control treatment. This is essentially a random coefficient model in which each subject has an individualized response to each of the treatments. The structure is quite distinct from the usual regression model where we tend to automatically assume that treatment effects are constant across subjects and additive. In Freedman's context inclusion of other covariates is potentially dangerous. Depending upon whether we have additivity and balanced design there are possible biases introduced by inclusion of covariates. Generally, with treatment randomization these biases can be show to be asymptotically negligible, but nevertheless they may be significant in particular finite sample settings, and Freedman recommends that the simpler model-free approach to estimating treatment effects be considered as a "more robust" alternative.

## 2. Introduction to Errors in Variables

A simple, yet revealing, estimation problem involves the following measurement error model. Assume we have two measurements of differing reliability from normal distributions having the same mean, i.e.

$$y_i = \mathcal{N}(\mu, \sigma_i^2) \qquad i = 1, 2.$$

When the $\sigma_i$'s are known the problem can be viewed as a *very* simple regression in which we have,

$$y = \mu 1 + u$$

where 1 denotes the 2-vector $(1,1)'$ and $u \sim \mathcal{N}(0,\Omega)$ where

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The *mle* of $\mu$ is the GLS estimator

$$\hat{\mu} = (\sigma_1^{-2} + \sigma_2^{-2})^{-1}(\sigma_1^{-2} y_1 + \sigma_2^{-2} y_2).$$

Substituting this into the likelihood yields the profile likelihood

$$\mathcal{L}(\sigma_1, \sigma_2 | y) = \frac{K}{\sigma_1 \sigma_2} \exp\left\{ \frac{-(y_1 - y_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}.$$

Transforming parameters so that $r = \sigma_1/\sigma_2$ and $d^2 = \sigma_1^2 + \sigma_2^2$ yields

$$\mathcal{L}(r, d | y) = \frac{K}{d^2} \; \frac{r^2 + 1}{r} \exp\left\{ -\frac{1}{2d^2}(y_2 - y_1)^2 \right\}.$$

If we look carefully at this function, we find that $\mathcal{L}$ has a saddlepoint as illustrated in Figure 1. For fixed $r$ corresponding to a ray in $(\sigma_1, \sigma_2)$-space, max $\mathcal{L}$ occurs at $d^2 = (y_2 - y_1)^2$. But for fixed $d$, min $\mathcal{L}$ occurs at $r^2 = 1$ with $\mathcal{L} \to \infty$ as $r^2 \to 0$, or as $r^2 \to \infty$.

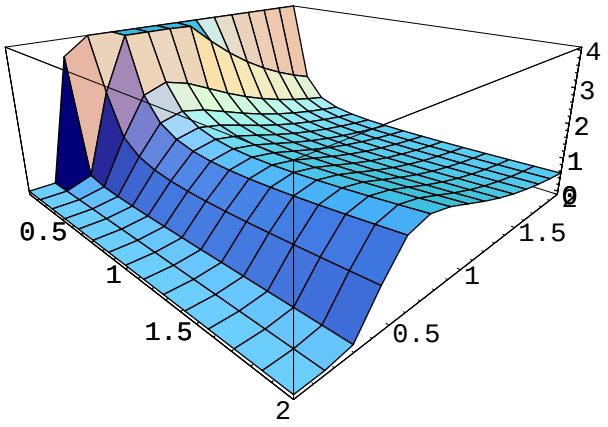


FIGURE 1. Likelihood surface of the measurement error model

What does this say about maximum likelihood estimation in this problem? In effect it says that there is no *mle*, or even more puzzling that the *mle* occurs when either $\sigma_1$ or $\sigma_2$ tends to 0. A better interpretation would be that we require further information about the relative reliability of $y_1$ and $y_2$ before we should be willing to use the *mle*. Note that for any fixed $r$ the problem is entirely conventional and well specified.

What does this have to do with regression? As a next step in this direction consider the following model:

$$\left.\begin{array}{l} y_i = \beta z_i + u_i \\ x_i = z_i + \varepsilon_i \end{array}\right\} \qquad i = 1, \ldots, n$$

Here we have our first encounter with the important class of *latent variable models*. Our interpretation of the model goes as follows: $y_i$ depends on $z_i$, but we don't observe $z_i$ directly, we only observe $x_i$ which is $z_i$ measured with error. The likelihood of $(z, \beta, \sigma_u, \sigma_\varepsilon)$ given the observed $\{y_i, x_i \quad i = 1, \ldots, n\}$ can be written as

$$\begin{aligned} \mathcal{L}(z, \beta, \sigma_u, \sigma_\varepsilon | y, x) &= K\sigma_u^{-n/2} \exp\{-\frac{1}{2\sigma_u^2} \parallel y - \beta z \parallel^2\} \\ &\quad \cdot \sigma_\varepsilon^{-n/2} \exp\{-\frac{1}{2\sigma_\varepsilon^2} \parallel x - z \parallel^2\} \end{aligned}$$

where $\parallel a \parallel^2 = \sum a_i^2$. Note that as in the simple measurement error model, $\mathcal{L} \to \infty$ if either,

$$\text{(i)} \qquad x \to z \text{ so } \sigma_\varepsilon^2 \to 0$$

or

$$\text{(ii)} \qquad y \to \beta z \text{ so } \sigma_u^2 \to 0$$

Corresponding to each of these cases we have an obviously optimal estimator of $\beta$:

$$\begin{aligned} &\text{(i)} \qquad \hat{\beta} = (x'x)^{-1}x'y \\ &\text{(ii)} \qquad \tilde{\beta} = ((y'y)^{-1}y'x)^{-1} \end{aligned}$$

The latter estimator is often called the "reverse regression estimator". The reader should verify that it is natural for case (ii).

If we look carefully at the likelihood, we see that solving for the maximum likelihood estimator of $\beta$, for fixed $\sigma_u, \sigma_\varepsilon$, amounts to minimizing a weighted average of horizontal and vertical distances *squared* to the line

$$y = \beta_0 + \beta_1 z$$

from the points$(x_i, y_i)$. This is illustrated in Figure 2.

The unweighted sum assuming $\sigma_u = \sigma_\varepsilon$ is just the squared orthogonal distance. As in the simple measurement error problem the case $\sigma_u = \sigma_\varepsilon$ corresponds to a saddle point of the likelihood. This estimator has a long history and is often called orthogonal least squares since it minimizes the sum of distances orthogonal to the fitted line. If we have some reliable way to estimate the relative variance $\sigma_u^2/\sigma_\varepsilon^2$, then we can easily adapt the estimate to this – there is certainly no compelling reason to assume $\sigma_u = \sigma_\varepsilon$ in most applications.

Suppose we ignore the measurement error in $z_i$ and just use $x_i$ in lieu of $z_i$. One often reads in empirical work that an ideal variable $z_i$ is unavailable so the "proxy" variable $x_i$ is used instead. What are the consequences? This

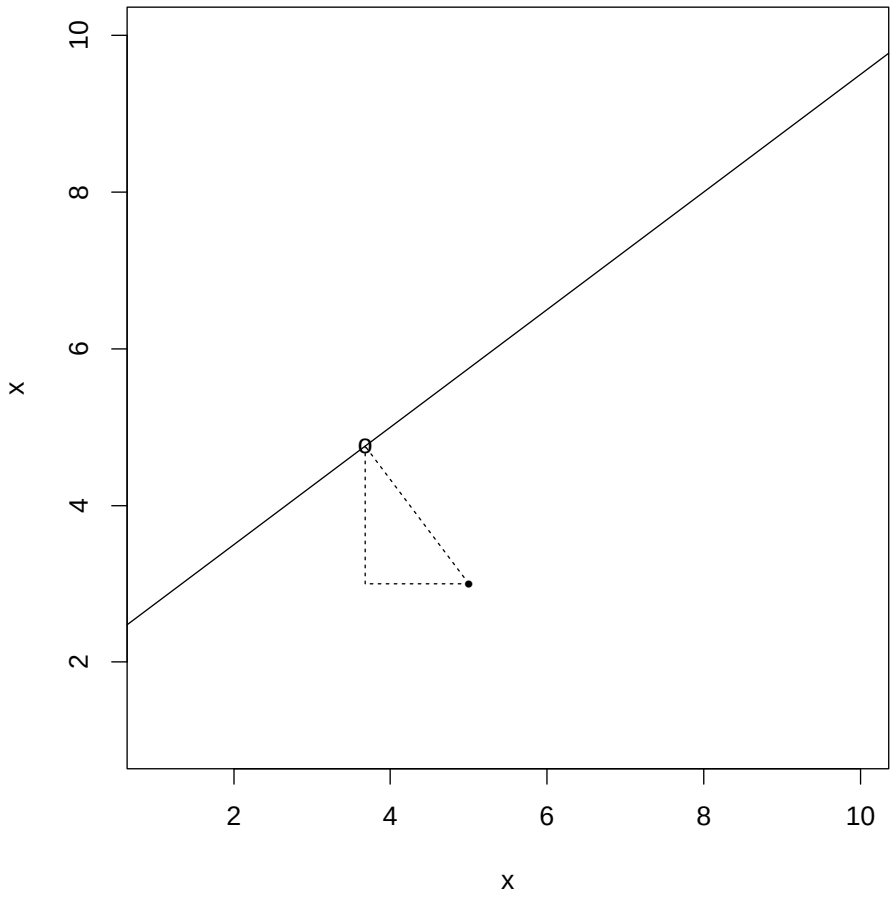


FIGURE 2. Orthogonal regression minimizes sum of the Euclidean distances from the observed points to the fitted line.

is easy to analyze in our simple bivariate model. If we substitute for $z_i$, we obtain,

$$y_i = \beta x_i + u_i - \beta \varepsilon_i$$

so that usual OLS estimator is,

$$\begin{aligned} \hat{\beta} &= (x'x)^{-1}x'y = (x'x)^{-1}x'(\beta x_i + u_i - \beta\varepsilon_i) \\ &= \beta + (x'x)^{-1}x'(u_i - \beta\varepsilon_i) \end{aligned}$$

Assuming that $z_i$ and $\varepsilon_i$ are uncorrelated,

$$\begin{aligned} En^{-1}x'x &= En^{-1}z'z + En^{-1}\sum \varepsilon_u^2 \\ &\equiv \sigma_z^2 + \sigma_\varepsilon^2 \end{aligned}$$

Similarly,

$$\begin{aligned} n^{-1}x'u &= n^{-1}(z+\varepsilon)'u \to 0 \\ n^{-1}x'\varepsilon &= n^{-1}(z+\varepsilon)'\varepsilon \to \sigma_\varepsilon^2 \end{aligned}$$

so

$$\begin{aligned} \hat{\beta} &= \beta - \sigma_\varepsilon^2\beta/(\sigma_z^2+\sigma_\varepsilon^2) \\ &= \beta\sigma_z^2/(\sigma_z^2+\sigma_\varepsilon^2) \end{aligned}$$

And this establishes that $\hat{\beta}$ is biased toward zero. At MIT this result, that in the simple errors in variables model the least squares estimator is biased toward zero, is called the "iron law of econometrics." As we would expect the result also shows that when $\sigma_\varepsilon^2$ is small the bias is small. Note that while it is tempting to extrapolate this result to more general errors in variables settings, this extrapolation has all the dangers of other exercises in extrapolation.

On the other hand, the reverse least squares estimator,

$$\tilde{\beta} = ((y'y)^{-1}y'x)^{-1}$$

can be analyzed in the same way. We have

$$\begin{aligned} En^{-1}y'y &= En^{-1}(\beta^2 z'z + u'u) \\ &= \beta^2\sigma_z^2 + \sigma_u^2 \\ En^{-1}y'x &= En^{-1}(\beta x'x + x'(u-\beta\varepsilon)) \\ &= \beta(\sigma_z^2+\sigma_\varepsilon^2) - \beta\sigma_\varepsilon^2 \end{aligned}$$

so

$$\begin{aligned} \tilde{\beta} &\to (\beta\sigma_z^2/(\beta^2\sigma_z^2+\sigma_u^2))^{-1} \\ &= \beta + \frac{\sigma_u^2}{\beta\sigma_z^2} \end{aligned}$$

which shows $\tilde{\beta}$ is biased *away* from zero.

A common, and very controversial, class of applications of the foregoing ideas involves testing for discrimination in labor markets. In the simplest case we may consider the following model for wages in a Chicago Bank analyzed by Conway and Roberts (*JBES*, 1985, 75-85)

$$y_i = \alpha + x_i\beta + s_i\gamma_+ u_i$$

where

$y_i$ = log wage of employee $i$
$x_i$ = scalar measure of "qualification"
$s_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$

In this model $\gamma$ may be interpreted as the percentage wage premium paid to mean. In the bank example $\hat{\gamma} = .148 \pm .072$ so we might say that women were underpaid by 15% or by $7-28\%$, based on the sample of 274 employees.

However, Conway and Roberts who were hired to defend the bank in court argued that $x_i$ was poorly measured and that $s_i$ was positively correlated with these measurement errors thereby "getting credit" for some of their effect, thereby resulting in an over estimate of $\gamma$. Note that this argument is at odds with the simple errors in variables argument advanced earlier. They suggest the reverse regression

$$x_i = y_i\alpha + s_i\delta + v_i$$

so now (strangely) we purport to explain variability in qualifications by current wage and gender. In this regression $\hat{\delta} = .01 \pm .04$ suggesting that given wages there is no systematic tendency for women to be more highly qualified than men.

## References


Student (1931), The Lanarkshire Milk Experiment, *Biometrika*, 23, 398-406.

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error, *Annals of Stat*, 11, 284-300.

Gertler, M. (2004). Do Conditional Cash Transfers Improve Child Health? *AER* 94, 336-341.

Freedman, D. (2009). On Regression Adjustment in Experiments with Several Treatments, *Annals of Applied Stat.* 2, 176-196.

Angrist J. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*, Princeton U. Press.