

Economics 696F, Topics in Econometrics

Lecture Note 1: Examples, Identification Concepts, and Large Sample Theory (corrected 1/26/10)

1: Interpreting OLS

Suppose data are $(y_i, x_i), i = 1, \dots, n$, distributed i.i.d. with joint CDF F . Here both y_i and x_i are scalar.

For now, suppose that F is completely unrestricted except for existence of second moments.

Consider the least squares problem:

$$\min_{\alpha, \beta} \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2.$$

Solution is (derive this yourself):

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2};$$

$$\hat{\alpha}_n = \bar{y}_n - \hat{\beta}_n \bar{x}_n;$$

where

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i,$$

provided that $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$.

Note we have *not* assumed that conditional mean of y_i is linear in x_i . This is simply the solution to the least squares problem for any given data set.

What do $(\hat{\alpha}_n, \hat{\beta}_n)$ converge to in large samples?

Rewrite:

$$\hat{\beta}_n = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

So if $E[(x_i - E(x_i))^2] = \text{Var}(x_i) > 0$, then application of the law of large numbers (LLN) and the Continuous Mapping Theorem (reviewed below) gives

$$\hat{\beta}_n \xrightarrow{p} \frac{E[(x_i - \mu_x)(y_i - \mu_y)]}{E[(x_i - \mu_x)^2]} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)},$$

where we have used the notation $\mu_x = E[x_i], \mu_y = E[y_i]$. Then

$$\hat{\alpha}_n \xrightarrow{p} \mu_y - \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \mu_x.$$

Is this an interesting limit??? Consider some different possible cases for F :

- **Linear Conditional Mean:** suppose F is such that

$$E[y_i|x_i] = \alpha_0 + \beta_0 x_i,$$

for some constants α_0, β_0 .

Let $\epsilon_i := y_i - \alpha_0 - \beta_0 x_i$. So by construction $y_i = \alpha_0 + \beta_0 x_i + \epsilon_i$, and

$$E[\epsilon_i|x_i] = E[y_i|x_i] - \alpha_0 - \beta_0 x_i = 0.$$

Then straightforward calculations (which you should verify) give:

$$\frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} = \beta_0,$$

and

$$E[y_i] - \beta_0 E[x_i] = \alpha_0.$$

So in this case, the OLS coefficients converge in probability to the population conditional linear mean coefficients:

$$\hat{\alpha}_n \xrightarrow{p} \alpha_0, \quad \hat{\beta}_n \xrightarrow{p} \beta_0.$$

- **Linear Conditional Mean with Omitted Variable:** suppose that there is another variable w_i , and now the vector (y_i, x_i, w_i) are i.i.d. from some joint distribution. Also suppose that

$$E[y_i|x_i, w_i] = \gamma_0 + \gamma_1 x_i + \gamma_2 w_i,$$

and we are interested in estimating the coefficient γ_1 .

It's straightforward to calculate:

$$\frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} = \gamma_1 + \gamma_2 \frac{\text{Cov}(x_i, w_i)}{\text{Var}(x_i)}.$$

So

$$\hat{\beta}_n \xrightarrow{p} \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \neq \gamma_1 \quad (\text{unless } \text{Cov}(x_i, w_i) = 0).$$

If we only observe (y_i, x_i) , and *not* w_i , then (in large samples) we can learn quantities like $\text{Cov}(x_i, y_i)$ that are associated with the joint distribution of (x_i, y_i) . However, we cannot hope to estimate quantities like $\text{Cov}(x_i, w_i)$ if we don't observe w_i . Then we cannot determine the asymptotic bias

$$\text{plim} \hat{\beta}_n - \gamma_1 = \gamma_2 \frac{\text{Cov}(x_i, w_i)}{\text{Var}(x_i)},$$

without making further assumptions.

- **Best Linear Predictor:**

Suppose we do *not* impose that $E[y_i|x_i]$ is linear. For example it could be that

$$E[y_i|x_i] = g(x_i),$$

for some nonlinear function $g(\cdot)$. Can we still give the OLS coefficients a meaningful interpretation?

Recall that $\hat{\alpha}_n, \hat{\beta}_n$ solve:

$$\min_{\alpha, \beta} \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2.$$

Equivalently, they solve

$$\min_{\alpha, \beta} \frac{1}{n} \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2.$$

Consider the “large sample” version of this problem:

$$\min_{\alpha, \beta} E [(y_i - \alpha - \beta x_i)^2].$$

In the population version of the problem, the solution α^*, β^* can be interpreted as the coefficients of the *best linear predictor*, the linear function of x_i that gives the smallest expected squared prediction error. They can be calculated explicitly as:

$$\begin{aligned} \beta^* &= \text{Cov}(x_i, y_i) / \text{Var}(x_i) \\ \alpha^* &= E[y_i] - \beta^* E[x_i]. \end{aligned}$$

(See Goldberger for this derivation.)

- **Aside:** what if $\text{Var}(x_i) = 0$? Then

$$P(x_i = \mu_x) = 1,$$

that is, x_i has a degenerate distribution at μ_x . What can be learned in this case?

2: Auction Example

Consider an auction for a single item. There are J bidders, and bidders $j = 1, \dots, J$ have a valuation v_j drawn independently from a common distribution F_v . We interpret v_j as the amount bidder j is willing to pay for the item.

Bidders observe their own valuation v_j , but not the valuations of other bidders. J and F_v are common knowledge. Each bidder decides on an amount to bid, b_j . If bidder j wins the good, she receives a payoff of $v_j - p$, where p is the price paid for the good.

Suppose that the auction is a second-price sealed-bid auction. That is, bidders submit bids simultaneously; the highest bidder wins the item and pays the second highest bid.

In this setting, a strategy for a given bidder j is a mapping $v_j \mapsto b_j$. It is well known that the strategy of bidding your own valuation ($b_j = v_j$) is a weakly dominant strategy. (See Krishna, *Auction Theory*.) We will assume that all bidders follow this strategy.

Empirical analysis: suppose we observe bids from a large number of auctions for identical items. Our goal is to use the bid data to learn the valuation distribution F_v .

More formally, assume that auctions $i = 1, \dots, n$ all have the same number $J \geq 2$ bidders. Let the valuation of bidder j in auction i be v_{ij} , and assume

$$v_{ij} \stackrel{\text{iid}}{\sim} F_V, \quad i = 1, \dots, n, j = 1, \dots, J$$

So valuations are i.i.d. within and across auctions.

If we observe all the bids, then the data are

$$\{b_{ij} : i = 1, \dots, n, j = 1, \dots, J\}.$$

Under the assumption that bidders play the weakly dominant strategy of making bids equal to valuations, each $b_{ij} = v_{ij}$. Then there is a natural way to estimate F_V . For any number $v \in \mathbb{R}_+$,

$$F_V(v) = Pr(v_{ij} \leq v) = E[1(v_{ij} \leq v)].$$

So a natural estimator is

$$\hat{F}_n(v) = \frac{1}{nJ} \sum_{i=1}^n \sum_{j=1}^J 1(b_{ij} \leq v).$$

By the LLN, for any v , $\hat{F}_n(v) \xrightarrow{P} F_V(v)$. We can in fact show the stronger result:

$$\sup_v |\hat{F}_n(v) - F_V(v)| \xrightarrow{P} 0,$$

which says that the entire function $\hat{F}_n(\cdot)$ converges to $F_V(\cdot)$ in probability. So with a large number of auctions, it is possible to learn $F_V(\cdot)$.

Now suppose that we do not observe all bids, but only the transaction price p_i for each auction. Since this is a second price auction, p_i will equal the second highest bid, which we denote $b_{i,(J-1)}$. By the assumption that bidders bid their valuations, this will equal the second highest valuation $v_{i,(J-1)}$. Now our data is

$$\{p_1, \dots, p_n\} = \{v_{1,(J-1)}, \dots, v_{n,(J-1)}\}.$$

Under this observational scheme, can we estimate F_V consistently?

Yes! Consider the empirical CDF of price,

$$\hat{G}_n(v) = \frac{1}{n} \sum_{i=1}^n 1(p_i \leq v).$$

By the same reasoning as above,

$$\hat{G}_n(v) \xrightarrow{P} G(v),$$

where $G(v)$ is the population distribution of p_i , or equivalently, the population distribution of the second highest draw of J i.i.d. draws from F_V . It turns out that there is a one-to-one mapping between G and F_V . So if we can consistently estimate G , then there is an implied F_V that must have generated it. See Athey and Haile (2002) for more details.

3: Identification Concepts

The law of large numbers and its various extensions suggest that with large samples, we can learn the distribution of the data. However, this may not be enough to learn the quantities of interest. This is the fundamental question of identification. Here we will begin to develop some notation and terminology for identification.

We want to learn about a *distribution of interest* $F \in \mathcal{F}$. We may be interested in the whole distribution, or in some feature of the distribution, which we will call a *parameter of interest* $\theta(F)$. Here $\theta(\cdot)$ is a known functional of F . (It could be the entire distribution, i.e. $\theta(F) = F$.)

We have an *observable distribution* G , and there is a known mapping $F \mapsto G$. Let \mathcal{G} be the set of possible observable distributions. The idea is that G is the thing we can learn from data. If it is possible to determine $\theta(F)$ from G , we say that θ is *identified*.

Usually, the distribution of interest F involves some variables that are not directly observable. (For example, we do not directly observe the actual valuations that bidders possess in the auction example.) The mapping from F to G indicates how the observable variables depend on the full set of variables.

Example: consider the Linear Conditional Mean with Omitted Variable case above. Then F is the joint distribution of (y_i, x_i, w_i) , and $\theta(F)$ is the parameter γ_1 . The set of possible distributions of interest \mathcal{F} is the set of all distributions that have $E[y_i|x_i, w_i]$ linear, bounded second moments, and $\text{Var}(x_i) > 0$.

The observable distribution G is the (marginal) joint distribution of (y_i, x_i) . Obviously, any F implies a G . However, knowledge of G does not allow us to determine $\gamma_1 = \theta(F)$. For example, consider F and \tilde{F} that have the same marginal distribution for (y_i, x_i) but different dependence between (y_i, x_i) and w_i . Then we may have $\theta(F) \neq \theta(\tilde{F})$, but both F and \tilde{F} lead to the same observable distribution G . [In the exercise, you will derive a specific expression for $\theta(F)$ to show to formally.] \square

Often, the distribution of interest F is a *structural* or *causal* model, which is meant to capture the underlying causal or economic mechanisms determining observed behavior. Then we talk about $\theta(F)$ as being a structural or causal parameter. At this level of generality, there is no distinction between specifications of F that are causal/structural and those that are not. But in practice, when we speak of causal or structural parameters, we imply that they are invariant to policy or treatment interventions that we might contemplate.

Auction Example: Here the underlying distribution F is the joint distribution of (v_1, \dots, v_J) , which by assumption has each component independent and distributed according to F_v . The parameter of interest is the valuation distribution F_v . If we only observe selling prices, then the observed distribution G is the distribution of prices implied by F_v . In particular, if bidders bid their valuation, then the distribution of prices is the distribution of the second highest out of J draws from F_v .

The valuation distribution F_v could be considered a structural parameter. For example, we might consider changing auction from a second price auction to a first price auction. If we applied this policy change to the same population, we would expect that individuals would still draw their valuations from the same distribution F_v , although they may well use a different strategy for bidding. \square

We also talk about a model being *testable*. Informally, this means that there are possible observable distributions that cannot be generated by any $F \in \mathcal{F}$. Then if we “observe” (estimate) such a G , this would suggest that our original assumption that the distribution of interest lies in \mathcal{F} is not correct. Clearly, testability depends on the choice of \mathcal{F} and our assumption about the form of the mapping $F \mapsto G$.

4: Some Large Sample Results

In identification analysis, we suppose that we know the observable distribution F and ask whether it allows us to recover the parameters of interest. This abstracts from the real empirical situation, where we have only a finite sample of data from F . However, if the data set is large, then the LLN and its extensions suggest that we can learn F reasonably well.

Of course, after verifying that some parameter of interest is identified, we would then want to construct an estimator of it, and study its properties. Often, the identification analysis will suggest how to construct the estimator. The following standard concepts results are good to keep in mind. (Drawn from van der Vaart, Ch. 2., Wooldridge, Ch.3, and Severini Ch. 11, 12.1-12.5, 13.2.)

Random Variables: we will use the term *random variable* to refer to random vectors in \mathbb{R}^k . The *distribution function* of a random variable X is the function

$$F(x) = P(X \leq x).$$

Convergence Concepts

Convergence of Deterministic Sequences: a sequence of nonrandom numbers $X_i: i = 1, 2, \dots$ converges to a limit a if, for any $\epsilon > 0$, there exists n_ϵ such if $n > n_\epsilon$ then $|X_n - a| < \epsilon$. We write $X_n \rightarrow a$ as $n \rightarrow \infty$.

Convergence in Probability: a sequence of random variables $\{X_i\}$ converges in probability to X if, for all $\epsilon > 0$,

$$P(|X_i - X| > \epsilon) \rightarrow 0.$$

This is denoted by $X_i \xrightarrow{p} X$, and we also write that $X = \text{plim } X_i$.

Usually, the limit X will be a constant, but random limits are allowed in the definition. When X_i is a vector and its probability limit is a constant, the definition above is equivalent to convergence in probability of each of the elements of X_i .

Convergence in Distribution: a sequence of random variables $\{X_i\}$ is said to converge in distribution to a random variable X if

$$P(X_i \leq x) \rightarrow P(X \leq x)$$

at every point x at which the limit distribution function $P(X \leq x)$ is continuous. This will be denoted by $X_i \xrightarrow{d} X$, or sometimes $X_i \rightsquigarrow X$.

Basic Asymptotic Results:

Weak Law of Large Numbers (WLLN): Let X_1, X_2, \dots be a sequence of i.i.d. random variables such that $E(\|X_1\|) < \infty$. Then

$$\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} E(X_1).$$

Multivariate Central Limit Theorem (CLT): Let X_1, X_2, \dots be i.i.d. random vectors in \mathbb{R}^k with mean $\mu = EX_1$ and covariance matrix $\Sigma = E(X_1 - \mu)(X_1 - \mu)'$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) = \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

Continuous Mapping Theorem (CMT): Let $g(\cdot)$ be a function from \mathbb{R}^k to \mathbb{R}^m , and suppose g is continuous at every point in a set C s.t. $P(X \in C) = 1$. Then

(i) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$;

(ii) If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Note that matrix addition and multiplication are continuous functions.

There are various useful facts about convergence in probability and convergence in distribution:

Result:

(i) Convergence in probability implies convergence in distribution:

$$X_i \xrightarrow{p} X \Rightarrow X_i \xrightarrow{d} X.$$

(ii) Convergence in distribution to a constant, implies convergence in probability:

$$X_i \xrightarrow{d} a \text{ (a constant)} \Rightarrow X_i \xrightarrow{p} a.$$

(iii) If $X_n \xrightarrow{d} X$ and $\|X_n - Y_n\| \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$.

(iv) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ where a is a constant, then the vector $(X_n, Y_n) \xrightarrow{d} (X, c)$.

(v) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$. (This is not true for convergence in distribution.)

A useful corollary of the previous result:

Slutsky's Lemma: Let X_n, X , and Y_n be random vectors or matrices. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a constant, then

(i) $X_n + Y_n \xrightarrow{d} X + c$;

(ii) $Y_n X_n \xrightarrow{d} cX$;

(iii) $Y_n^{-1}X_n \xrightarrow{d} c^{-1}X$, provided c is invertible.

Delta Method: Let X_n be a sequence of d -dimensional random vectors such that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \Sigma),$$

where Σ is positive definite and finite. Let g denote a continuously differentiable function from \mathbb{R}^d into \mathbb{R}^k , and let $G(x) = \partial g / \partial x$ denote the $k \times d$ matrix of partial derivatives. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, G(\mu)\Sigma G(\mu)').$$

Convergence of Functions: many of the convergence results for scalar and vector quantities extend to functions. For example the law of large numbers extends to estimating the entire CDF.

Glivenko-Cantelli Theorem: Suppose X_1, X_2, \dots are i.i.d. random variables with distribution function F on the real line. Let $\hat{F}_n(\cdot)$ be the empirical CDF defined for any $x \in \mathbb{R}$ as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x).$$

Then

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{p} 0.$$

There are analogs of the continuous mapping theorem and the delta method as well. See van der Vaart, Chs. 18-19.

First Exercise, due Feb 8 at start of class:

1. Fill in all the missing details of calculations in Sections 1 and 2 of this note.
2. For the Linear Conditional Mean with Omitted Variable example, express γ_1 as a functional of the joint distribution of (y_i, x_i, w_i) . Based on this, show that γ_1 is *not* identified given only knowledge of the distribution of (y_i, x_i) .
3. Drawing on the notation in Section 3, give a formal (mathematical) definition of identification and of testability. Be as precise as possible. You may need to expand on the notation and definitions in Section 3 to do so.