ALEC BRANDON

EMPIRICAL ANALYSIS

Part I

Empirical Analysis 31000—Azeem Shaikh

Probability preliminaries

Set theory

Sets, they're just collections of things. Some sets have special features that'll be useful to us, but before we define those, it's probably useful to point out the following definitions/features/notation used for sets:

1. The set A can be defined as: $A = \{x \in \mathbb{R} : P(x) \text{ is true } \}$.[1]
2. If $A \subset B$ then if $x \in A$ then $x \in B$.
3. If $A \subset B$ and $B \subset A$ then $A = B$.
4. $A \cup B = \{x : x \in A \text{ or } x \in B\}$.
5. $A \cap B = \{x : x \in A \text{ and } x \in B\}$.
6. $A^c = \{x : x \notin A\}$.
7. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. And $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
8. *DeMorgan's Law*: $(A \cup C)^c = A^c \cap C^c$ and $(A \cap C)^c = A^c \cup C^c$.

OK. Now time for a few definitions that'll be useful for the material in this class:

Def. (Disjoint): A, B are disjoint iff $A \cap B = \emptyset$. E.g., $P\{A \cup B) = P\{A) + P\{B)$ if A, B are disjoint.

Def. (Pairwise Disjoint): $A_1, A_2, \ldots$ are pairwise disjoint iff $A_i \cap A_j = \emptyset \; \forall \; i \neq j$.

Def. (Partition): If $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = \mathbf{S}$ then they form a partition of $\mathbf{S}$.

Probability theory

The basic idea of probability theory is that there is some Sample Space, $\mathbf{S}$, that's the set of all possible outcomes of an experiment.

[1] Where $P(x)$ is whatever you want it to be. E.g., $x \in [0, 1]$.

Then there are Events, A, which are just any collection of possible outcomes of an experiment. Consider an example to see what these things might mean. Say that we're interested in coin flips. Then $\mathbf{S} = \{H, T\}$ and an event would be any subset of possible outcomes of flipping a coin, like, $A = \{H\}$.

More mathematically, an Event is any subset of $\mathbf{S}$, including $\mathbf{S}$. The third ingredient for probability theory is the probability function, P, which is defined by the following axioms:

Figure 1: Andrey Nikolaevich Kolmogorov. Respect.

Def. (Kolmogorov's Axioms): A probability function, P, is a function P that maps from the sample-space, $\mathbf{S}$, to the real numbers and satisfies:

1. $P\{A\} \geq 0$ for all events, A.
2. $P\{S\} = 1$.
3. If events $A_1, A_2, \ldots$ are pairwise disjoint, then $P\{\cup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} P\{A_i\}$.

With those established the following theorems follow:

Thm. (Handful of Probability Theorems): *Presented below without proof. See Chapter 1 of Casella and Berger.*

$P\{\emptyset\} = 0$

$P\{A\} \leq 1$

$P\{A^c\} = 1 - P\{A\}$

$P\{A \cap B^c\} = P\{B\} - P\{A \cap B\}$

$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$

If $A \subset B \implies P\{A\} \leq P\{B\}$

$P\{A\} = \sum_{i=1}^{\infty} P\{A \cup C_i\}$ *for any partition* $C_1, C_2, \ldots$

Boole's Inequality $P\{\cup_{i=1}^{\infty} A_i\} \leq \sum_{i=1}^{\infty} P\{A_i\}$ *for any sets* $A_1, A_2, \ldots$

Bonferonni's Inequality' $P\{A \cap B\} \geq P\{A\} + P\{B\} - 1$.[2]

Inequalities

In the spirit of the inequalities above, another useful inequality is the Cauchy-Schwarz inequality. It doesn't really have anything to do with probability theory, per se, so we'll do a quick section on inequalities here. In that spirit:

[2] Note: This follows directly from Boole's Inequality. All you need is Boole's Inequality, the fact that $P\{A^c\} = 1 - P\{A\}$, and Captain DeMorgan's Law.

Thm. (Cauchy-Schwarz):

$$|a+b|^2 \leq |a|^2 + |b|^2$$

And a closely related inequality is the triangle inequality.

△△△△△△△△△△△△△△△△

△△△△△△△△△△△△△△△△

△△△△△△△△△△△△△△△△

△△△△△△△△△△△△△△△△

△△△△△△△△△△△△△△△△

Thm. (Triangle Inequality):

$$|x+y| \leq |x| + |y| \;\; \textit{and} \;\; |x-y| \geq ||x| - |y||$$

As we will come to see later, inequalities are pretty handy when we're interested in proving something converges to zero.

Conditional probability and independence

GIVEN the name of the subsection, I think we can dispense with a clever intro and just get into it:

Def. (Conditional Probability):

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}$$

Thm. (Bayes's Rule): [3] *Let A_i be a partition of the sample-space and let B be any set. Then:*

$$P\{A_i|B\} = \frac{P\{B|A_i\}P\{A_i\}}{\sum_{j=1}^{\infty} P\{B|A_j\}P\{A_j\}}.$$

Def. (Independence): Two events, A, B, are statistically independent i/f/f $P\{A \cap B\} = P\{A\}P\{B\}$.

Thm.: *If A, B are independent events then:*

A, B^c are independent

A^c, B are independent

A^c, B^c are independent

Def. (Mutual Independence): A collection of events, $A_1, \ldots, A_n$ are called mutually independent i/f/f for any subcollection $A_{i_1}, \ldots, A_{i_k}$ we have:

$$P\left\{\cap_{j=1}^{k} A_{i_j}\right\} = \prod_{j=1}^{k} P\{A_{i_j}\}$$

[3] Of Thomas Bayes fame. We call $P\{A_i|B\}$ the posterior, $P\{A_i\}$ the prior, and $P\{B|A_i\}$ is just the likelihood for the conditional probability.

Random variables and distribution functions

RANDOM variables are crazy useful because we can think about combinations of events as opposed to just events.[4] Anyways, here's the key details:

Def. (Cumulative Distribution Function): The CDF of a random variable, X, $F_X(x)$ is:[5]

$$F_X(x) = P\{X \leq x\}, \forall x$$

Def. (Identically distributed): The random variables X, Y are identically distributed iff for every set A, $P\{X \in A\} = P\{Y \in A\}$. Equivalently we can say that $F_X(x) = F_Y(x)$.

Def. (Probability Mass/Density Function): The PMF of a discrete random variable is:

$$f_X(x) = P\{X = x\} \; \forall x$$

The PDF of a continuous random variable, $f_X(x)$ satisfies:

$$F_X(x) = \int_{-\infty}^{x} f_X(\tilde{x}) d\tilde{x} \; \forall x$$

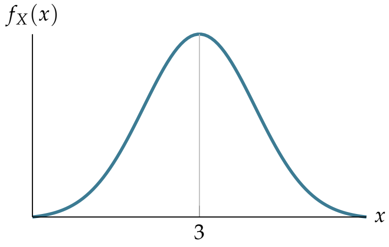

Figure 2: PDF of $X \sim P = \mathcal{N}(3,1)$.

Expectations and moments

EXPECTATIONS are crazy important for econometrics because just about everything is a conditional expectation, so it's worth going over some definitions of this stuff here:

Def. (Expectations): The expected value of a random variable, $g(X)$, denoted by $Eg(X)$ or $E[g(X)]$ is:

$$Eg(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{-\infty}^{\infty} g(x) dF_X, & \text{if } X \text{ is continuous} \\ \sum_x g(x) f_X(x) = \sum_{x \in \mathbf{X}} g(x) P\{X = x\}, & \text{if } X \text{ is discrete} \end{cases}$$

where X is some random variable and $g(\cdot)$ is some function of X. If this seems weird then just think of the random variable, $Y = g(X)$.[6]

These formulas might seem a little overwhelming, but it's just a bunch of notation to describe averages. To see that they're actually quite straightforward, the following theorem will highlights some nice features:

Thm. (Fun with Expectations): *Let X be a random variable and let a, b, c be constants. Then for any functions, $g_1(\cdot)$ and $g_2(\cdot)$ whose expectations exist we have:*

[4] I.e., you don't have to have a seperate piece of notation for every event, but can think about just summing up outcomes of events into one number, X.

[5] Often we'll suppress the X in $F_X(x)$ and just write $F(x)$ when it's clear from the context.

[6] If the sum or integral doesn't exist then we say that $Eg(X)$ doesn't exist. To denote that the expectation exists we generally just write: $Eg(X) < \infty$.

$E[ag_1(X) + bg_2(X) + c] = aEg_1(X) + bEg_2(X) + c$

If $g_1(x) \geq 0$ for all x, then $Eg_1(X) \geq 0$.

If $g_1(x) \geq g_2(x)$ for all x, then $Eg_1(X) \geq Eg_2(X)$.

If $a \leq g_1(x) \leq b$ for all x, then $a \leq Eg_1(X) \leq b$.

The notion of an expectation might seem a bit limited—after all, it's just an average—but it can be expanded to higher orders. As in, the average is the first moment. The variance is the second moment.[7] Here's a definition:

Def. (nth Moment): For each integer n, the n^{th} raw moment of X is:

$$EX^n$$

and the n^{th} central moment of X is:

$$E[(X - EX)^n] = E[(X - \mu_X)^n]$$

The second moment, or variance, is so important that we'll recite a few theorems about it here:

Thm. (Fun with Variance): *If X is a random variable with finite variance, then for any constants a, b:*

$$Var(aX + b) = a^2 Var(X).$$

Also, you can write:

$$Var(X) = E[(X - EX)^2] = EX^2 - (EX)^2.$$

What about sample variances?

Def. (Sample Variance): For $X_1, \ldots, X_n$ in $\mathbb{R}$, the sample variance,[8] denoted, S_n^2 is:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

For higher dimensions, the sample variance-covariance matrix is:

Def. (Sample Variance-Covariance Matrix): For $X_1, \ldots, X_n$ in $\mathbb{R}^k$ the sample variance-covariance matrix, denoted, $\hat{\Sigma}$ is:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$$

[7] The third moment? Who knows. Skew? Kurtosis?

[8] This is only for the univariate case, b/t/w.

Covariances and correlations

LATER we'll derive these results as a consequence of the Delta Method, but for now we'll just state them, as they're useful properties.

Def. (Covariance): Suppose we have draws from a joint distribution, $(X_1, Y_1), \ldots, (X_n, Y_N) \overset{iid}{\sim} P$, on $\mathbb{R}$. Also suppose that $Var(X_i) < \infty, Var(Y_i) < \infty$. Then we can define the covariance of (X_i, Y_i) as:

$$\begin{aligned} Cov[X_i, Y_i] &= E[(X_i - EX_i)(Y_i - EY_i)] \\ &= E[X_i Y_i] - E[X_i]E[Y_i] \end{aligned}$$

Of course, being a rigorous course in these matters, we have to ask our selves: Does $E[X_i Y_i]$ exist?[9]

Thm. (Cauchy-Schwarz with $\mathbb{E}$): *Take U, V as random variables. Then:*

$$E[UV]^2 \leq E[U^2]E[V^2].$$

Covariances are cool and all, but they're lacking in the units department. For an easier to interpret version, we look at correlations:

Def. (Correlation): Borrowing the same setup we used when defining covariances, if we also have that $Var(X_i) > 0$ and $Var(Y_i) > 0$ then the correlation between X_i and Y_i is defined as:

$$\rho_{X,Y}(P) = Corr(X_i, Y_i) = \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i)Var(Y_i)}}$$

One nice trick for correlations is the following theorem:

Thm. (Perfectly Correlated): $|\rho_{X,Y}| \leq 1$ *and with equality* $\iff$ X_i *is a linear function of* Y_i.

How to estimate covariances and correlations, though?[10]

Def. (Sample Covariance):

$$\hat{\sigma}_{X,Y,n} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

Def. (Sample Correlations):

$$\hat{\rho}_{X,Y,n} = \frac{\hat{\sigma}_{X,Y,n}}{S_{X,n}S_{Y,n}}$$

which is consistent if we assume that: $E[X_i^4] < \infty$ and $E[Y_i^4] < \infty$.

[9] Great question, but, stay tuned on that one. We'll prove it in a sidenote later.

[10] We haven't really gotten there yet, but it makes more sense to present here than later, so whatevs.

Large sample theory

Given $X_1, \ldots, X_n \overset{iid}{\sim} P$ which is assumed to satisfy some parametric/non-parametric assumptions.[11] The goal is to learn about some parameter, $\theta(P)$, like the mean ($\mu(P)$) or the variance ($\sigma^2(P)$). By learn, the following three prototypical problems should convey what we mean:

1. Estimate $\theta(P)$. An estimator of $\theta(P)$ is a function, $\hat{\theta}(P)(X_1, \ldots, X_n)$ that provides a best guess for $\theta(P)$.

2. Test the null hypothesis that $\theta(P) = \theta_0$ vs. the alternative hypothesis that $\theta(P) \neq \theta_0$. A test is a function $\phi_n := \phi_n(X_1, \ldots, X_n)$ that takes values in $[0,1]$ (typically just $\{0,1\}$) and it gives the researcher the probability with which to reject the null.

3. Construct a confidence region for $\theta(P)$. A confidence region is a random set $C_n = C_n(X_1, \ldots, X_n)$ s/t $P\{\theta(P) \in C_n\} \approx 1 - \alpha$ for some pre-specified value of $\alpha \in (0,1)$.

And when we work with any of these types of problems we'll be interested in the large-sample behavior because small-sample behavior is sensitive to assumptions about P.

Convergence in probability

Before we get to discussing convergence in probability, it's useful to have a discussion of general forms of convergence, because they'll come up:

Def. (Convergence): A sequence a_n converges to a, sometimes written: $a_n \to a$, if $\forall\ \epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n > N$:

$$|a_n - a| < \epsilon.$$

Thm. (Monotone Convergence): *If $\{a_n\}$ is a monotone sequence*[12] *then $\{a_n\}$ has a limit i/f/f $\{a_n\}$ is bounded.*

[11] iid means independent and identically distributed. By identically distributed we mean:

$$P\{X_i \leq x\} = P\{X_j \leq x\}\ \forall x, \forall i, j$$

and by independent we mean:

$$P\{X_{i_1} \leq x_1, \ldots, X_{i_k} \leq x_k\} = \prod_{1 \leq j \leq k} P\{X_{i_j} \leq x_j\}$$

where $i_1, \ldots, i_k$ are distinct indices.

[12] $a_n \leq a_{n+1}$ or $b_n \geq b_{n+1}$ for all n. Notation note: For the sequences in this footnote we'd write: $a_n \uparrow$ and $b_n \downarrow$. If a_n is a monotone increasing sequence converging to a then we write: $a_n \uparrow a$. If $b_n \downarrow$ is converging to b then we write: $b_n \downarrow b$.

Def. (Pointwise Convergence): We say that $f_n : \mathbb{R} \to \mathbb{R}$ converges pointwise[13] to f i/f/f:

$$\lim_{n \to \infty} f_n(x) = f(x) \; \forall \; x$$

It's important to note, however, that these notions of convergence aren't super useful for probability. For example, if you're flipping a fair coin, there's a non-zero probability that the coin will always land on tails. This is all just to say that we need something better than these forms of convergence.

Def. (Convergence in Probability): A sequence of random vectors, $\{X_n : n \geq 1\}$ converges in probability to another random vector, X, i/f/f:

$$\forall \; (\epsilon > 0, \delta > 0), \exists \; N \in \mathbb{N} \; s.t. \; \forall \; n > N, P\{|X_n - X| \geq \epsilon\} < \delta.$$

Sometimes the δ and N is done away with and the following is used:

$\forall \; \epsilon > 0, P\{|X_n - X| > \epsilon\} \to 0$ as $n \to \infty$.

What do we do if we want to signify that $X_n \xrightarrow{p} +\infty$? Need to show that for all $c > 0$, $P\{X_n > c\} \to 1$. Notation is: $X_n \xrightarrow{p} +\infty$ as $n \to \infty$. For $-\infty$ we say that $X_n \xrightarrow{p} -\infty$ i/f/f for all $c > 0$, $P\{X_n < -c\} \to 1$.

Before we jump into the most important application of convergence in probability, we'll review one important theorem:

Figure 3: Andrey Markov pondering his inequality.

Thm. (Markov's Inequality): *For any random variable, X and any $q > 0, \epsilon > 0$:*

$$P\{|X| > \epsilon\} \leq \frac{E\left[|X|^q\right]}{\epsilon^q}$$

Proof. $P\{|X| > \epsilon\} = E[I\{|X| > \epsilon\}]$ and $I\{|X| > \epsilon\} \leq \frac{|X|^q}{\epsilon^q}$. To see see this last property just consider the two cases where the LHS is 1 and the LHS is 0. Then taking expectations of the inequality yields the theorem. □

Thm. (Weak Law of Large Numbers): *Let $X_1, \ldots, X_n \overset{iid}{\sim} P$ and suppose that $\mu(P), \sigma^2(P)$ exists*[14] *then:*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} E[|X_i|] = \mu(P) \;\; as \; n \to \infty$$

Proof.

$$\begin{aligned} P\{|\bar{X}_n - \mu(P)| > \epsilon\} &\leq \frac{E[|\bar{X}_n - \mu(P)|^2]}{\epsilon^2} \text{ by Markov's Inequality with } q = 2 \\ &= \frac{Var[|\bar{X}_n|]}{\epsilon^2} \\ &= \frac{1}{n} \frac{\sigma^2(P)}{\epsilon^2} \text{ because } Var(|\bar{X}_n|) = \frac{1}{n^2} \sum_{i}^{n} Var(X_i) = \frac{\sigma^2(P)}{n} \\ &\xrightarrow{p} 0 \end{aligned}$$

□

[13] Also known at p'wise.

[14] That is, $E[|X_i|] < \infty$.

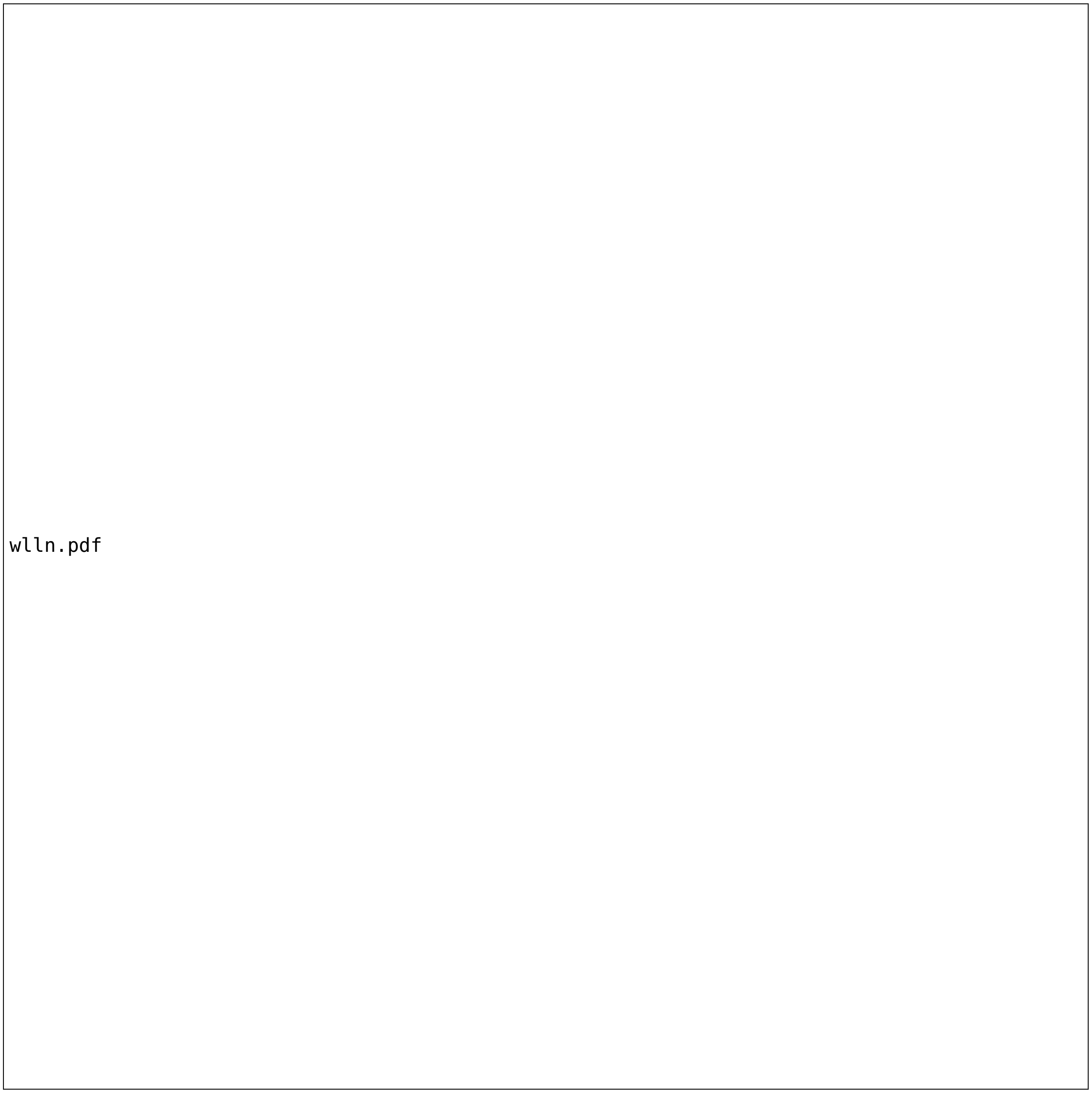

Figure 4: The moving average of $X_i \sim$ Bernoulli(0.3) drawn 1000 times and then replicated 25 times. Seems like the WLLN might be onto something.

We can write any random variable X as: $X = X^+ - X^-$ and $|X| = X^+ + X^-$ where $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$. Then we can call $E[X] =: E[X^+] + E[X^-]$ when both $E[X^+], E[X^-]$ exist.[15] Some terminology now:

$E[X]$ is called the first (raw) moment of X.

$E[X^r]$ is called the r^{th} (raw) moment of X.

$E[(X - E[X])^r]$ is the r^{th} centered moment of X and the second centered moment of X is the variance.

Thm. (Jensen's Inequality): *Let $I \subseteq \mathbb{R}$ be a convex set and $f : I \to \mathbb{R}$ be a convex function*[16] *and let X be a random variable with $P\{X \in I\} = 1$. Then if $E[|X|] < \infty$ and $E[|f(X)|] < \infty$:*[17,18,19]

$$f(E[X]) \leq E[f(X)].$$

Proof. Let $c = E[X]$. Then there are two possibilities for c, either it's in the interior of I or it's not.

Suppose $c \notin Int\{I\}$, then $P\{X = c\} = 1$. Then:

$$f(E[X]) = f(c)$$

and:

$$\begin{aligned} E[f(X)] &= E[f(c)] \text{ because } P\{X = c\} = 1 \\ &= f(c) \text{ because } f(c) \text{ is just a number.} \end{aligned}$$

and then we have that $f(E[X]) = E[f(X)]$ and then the inequality trivially holds.

Now suppose the other case: $c \in Int\{I\}$. Then define:

$$\Delta_{+,h}(c) = \frac{f(c+h) - f(c)}{h}, \ \Delta_{-,h'}(c) = \frac{f(c) - f(c-h')}{h'}.$$

By convexity we have that $\Delta_{-,h}(c) \leq \Delta_{+,h}(c)$.[20] Also, convexity gives us $\Delta_{+,h}(c) \downarrow$ as $h \downarrow 0$. And equivalently $\Delta_{-,h}(c) \uparrow$ as $h \downarrow 0$.[21] Then we have that:

$$-\infty < \Delta_{-,h'}(c) \leq \Delta_{-,h}(c) \leq \Delta_{+,h}(c) \leq \Delta_{+,h'}(c) < \infty$$

for $h' > h$. Then if we define:

$$D_+(c) = \lim_{h \downarrow 0} \Delta_{+,h}(c), \ D_-(c) = \lim_{h \downarrow 0} \Delta_{-,h}(c)$$

we know that both exist and $D_+(c) \geq \Delta_{-,h}(c) > -\infty$ and likewise for $D_-(c)$. Why? Just see what happens to the full inequality with the Δ terms above as $h \downarrow 0$.

[15] Finite.

[16] For any $x_1, x_2 \in I$, $t \in [0,1]$, we say f is a convex function iff: $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$.

[17] If $f(\cdot)$ is concave then the inequality is reversed.

[18] Some non-standard convex functions: $f(x) = |x|$ and $f(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$.

[19] A good mneomonic device to remember the direction is that $Var(X_i) \geq 0$ and:

$$Var(X_i) = E[X_i^2] - E[X_i]^2 \geq 0.$$

[20] Just pick $x_1 = c + h$, $x_2 = c - h$, and $t = 1/2$ and rearrange the definition of convexity.

[21] The notation here means that as h goes from a positive number to 0, it creates a sequence of $\Delta_{+,h}(c)$ that is monotonically decreasing. A proof of this observation follows from fixing $h > 0$ and picking $h' = (1-\theta)h$ where $\theta \in (0,1)$ and then expanding and applying the definition of convexity to $\Delta_{+,h'}(c)$.

Now choose $m \in [D_-(c), D_+(c)]$ and let $L(x) = f(x) + m(x-c)$. If $x = c + h$ then $L(c+h) = f(c) + m(h) \leq f(c) + D_+(c)h \leq f(c) + \Delta_{+,h}(c)h = f(c+h)$ and likewise for $x = c - h$. So $L(x) \leq f(x)\ \forall x\ \in I$ and $L(c) = f(c)$. This is the key step because look what happens next.

Then $E[f(X)] \geq E[L(X)] = L(E[X]) = L(c) = f(c) = f(E[X])$ and we're done! □

Now for a useful fact about the existence of moments:

Thm. (Existence of Moments): *If $E[|X|^k] < \infty$ then $E[|X|^j] < \infty$ for $j < k$.*

Proof. Suppose $E[|X|^k] < \infty$. Then pick $j < k$ and define $f(x) = |x|^{k/j}$ which is a convex function. Then also define $Y = |X|^j$ and $Y_n = \min\{|X|^j, n\}$. The strategy of the proof is going to be to show that $Y_n \uparrow Y$. Then we'll use the Monotone Convergence Theorem to finish up.

Fix n. Then $E[|Y_n|] < \infty$ because if $|X|^j \not< \infty$ then $Y_n = n$. Also note that: $E[|f(Y_n)|] = E[|\min\{|X|^k, n\}|] \leq E[|X|^k] < \infty$, where the last inequality is the assumption of the proof. Then by Jensen's inequality we get:

$$f(E[Y_n]) \leq E[f(Y_n)] \leq E[|X|^k] < \infty$$

and if we apply $f(\cdot)$ we get:

$$|E[Y_n]|^{k/j} \leq E[|X|^k] \implies E[Y_n] \leq \left(E[|X|^k]\right)^{j/k} < \infty$$

Then because $Y_n \uparrow Y$, by the Monotone Convergence Theorem we get that $E[Y_n] \uparrow E[Y]$ this implies that $E[|X|^j] < \infty$. [22] □

It's worth noting that this existence of moments theorem is particularly useful when we're given something like: $E[X^4] < \infty$. Why? Well it tells us that the mean and variance exist and that we can do all the fun stuff that that conclusion gives us.[23]

E.g.: $X_1, \ldots, X_n \overset{iid}{\sim} P$ on $\mathbb{R}$ and suppose $\mu(P)$ exists. Goal is to estimate $\mu(P)$. A natural estimator of $\mu(P)$ is:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is just the sample mean. By the WLLN $\bar{X}_n \xrightarrow{p} \mu(P)$. That is, the sample average is a consistent estimator for $\mu(P)$.

This estimator is an example of the *analog principle* an estimator obtained by replacing the unknown P with an empirical estimator of

[22] Note that a more intuitive proof that (probably) isn't perfect would just apply Jensen's inequality to $E[|X^j|]$:

$$E[|X|^k] = E[f(|X|^j)] \geq f\left(E[|X|^j]\right) = \left(E[|X|^j]\right)^{k/j}$$

which we can pick the first and last terms from and do a little bit of algebra to get:

$$E[|X|^j] \leq \left(E[|X|^k]\right)^{j/k} < \infty.$$

[23] Also note that if $f : I \to \mathbb{R}$ is convex and I is open the proof shows that:

$$f(x) = \sup_{L \in \mathbb{L}} L(x)$$

where

$$\mathbb{L} = \{L \text{ linear s.t. } L(x) \leq f(x) \forall x \in I\}.$$

it, $\hat{P}_n$. In an iid setting, $\hat{P}_n$ is just the empirical distribution and puts equal mass on $X_1, \ldots, X_n$. For example, we can write the empirical CDF as:

$$\hat{F}_n(x) = \frac{1}{n} \sum_i I\{X_i \leq x\}$$

and in the homework we'll show that $\hat{F}_n(x) \xrightarrow{p} F(x)$ for all x. You can actually show further than:

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

which is the Glivenko-Cantelli theorem, a big part of empirical process theory.

Now we've only discussed convergence in probability in terms of scalars. How does it extend to vectors? The next theorem will help us think about that.

Lem. (Marginal and Joint Convergence): *Take the sequence $\{X_n : n \geq 1\}$ and X, two random vectors in $\mathbb{R}^k$. Define X_{n_j}, X_j as the j^{th} component of those vectors. That is, $X = (X_1, \ldots, X_j, \ldots, X_k)$ and equivalently for X_n. Then if $\forall\ 1 \leq j \leq k, X_{n_j} \xrightarrow{p} X_j$ then $X_n \xrightarrow{p} X$.*[24]

Proof. w/t/s that for $\epsilon > 0$, $P\{|X_n - X| > \epsilon\}$ goes to 0.

$$\begin{aligned} P\{|X_n - X| > \epsilon\} &= P\left\{\sum_{1 \leq j \leq k} (X_{n_j} - X_j)^2 > \epsilon^2\right\} \text{ by def. of the metric space} \\ &\leq P\left\{\bigcup_{1 \leq j \leq k} \left\{(X_{n_j} - X_j)^2 > \frac{\epsilon^2}{k}\right\}\right\} \\ &\leq \sum_{1 \leq j \leq k} P\left\{|X_{n_j} - X_j| > \frac{\epsilon}{\sqrt{k}}\right\} \text{ by Boole's Inequality} \\ &\to 0 \end{aligned}$$

Where the second step follows because if $A = \{\sum_{1 \leq j \leq k}(X_{n_j} - X_j)^2 > \epsilon^2\}$ and $B = \left\{\bigcup_{1 \leq j \leq k}(X_{n_j} - X_j)^2 > \frac{\epsilon^2}{k}\right\}$ then $A \subseteq B$.

□

The WLLN is pretty awesome, right? Well. What if it were even better? It would be nice if it could speak to more than just sample averages. The continuous mapping theorem steps up big time here:

Thm. (Continuous Mapping Theorem (CMT1)): *Take the sequence $\{X_n : n \geq 1\}$ and X, two random vectors in $\mathbb{R}^k$. Let the function, $g(\cdot)$, where $g : \mathbb{R}^k \to \mathbb{R}^d$ be a continuous function on a set C*[25] *such that $P\{X \in C\} = 1$ and $X_n \xrightarrow{p} X$. Then:*

$$g(X_n) \xrightarrow{p} g(X).$$

[24] In words, this theorem just says that if each component converges, then the entire vector converges.

[25] $g(\cdot)$ is continuous on a set C means that $g(\cdot)$ is continuous at each point $x \in C$. That is:

$$\forall\ x \in C,\ \exists\ \delta > 0\ s/t\ |x - y| < \delta \implies |g(x) - g(y)| < \epsilon\ .$$

Proof. For $\epsilon > 0$, $P\{|g(X_n) - g(X)| > \epsilon\}$, define the problematic points as:

$$B_\delta = \{x \in \mathbb{R}^k : \exists y \text{ with } |x - y| \leq \delta \text{ \& } |g(x) - g(y)| > \epsilon\}$$

Then if $x \notin B_\delta$ we get:

$$x \notin B_\delta \implies \forall\, y, |x - y| > \delta \text{ or } |g(x) - g(y)| \leq \epsilon$$

and if $X \notin B_\delta$ we get:[26]

$$X \notin B_\delta \implies \forall\, X_n, |X - X_n| > \delta \text{ or } |g(X) - g(X_n)| \leq \epsilon$$

Then:

$$\begin{aligned} P\{|g(X_n) - g(X)| > \epsilon\} &= P\{\{|g(X_n) - g(X)| > \epsilon\} \cap \{X \notin B_\delta\}\} + P\{\{|g(X_n) - g(X)| > \epsilon\} \cap \{X \in B_\delta\}\} \\ &\leq P\{|g(X_n) - g(X)| > \delta\} + P\{X \in B_\delta\} \\ &= \underbrace{P\{|g(X_n) - g(X)| > \delta\}}_{\to 0 \text{ as } n \to \infty} + \underbrace{P\{\{X \in B_\delta\} \cap C\}}_{\to 0 \text{ as } \delta \downarrow 0} \\ &\to 0 \end{aligned}$$

□

E.g. (Applying the CMT1): S_n^2 is a consistent estimator of $\sigma^2(P)$:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right] \quad \text{by expanding } \bar{X}_n \text{ and rearranging.} \\ &= f\left(\frac{n}{n-1}, \frac{1}{n}\sum_i X_i^2, \bar{X}_n\right) \quad \text{with } f(x,y,z) = x(y - z^2) \text{ which is a nice and continuous function.} \\ &\xrightarrow{p} f(1, E[X_i^2], E[X_i]) \\ &= E[X_i^2] - E[X_i]^2 \\ &= \sigma^2(P) \end{aligned}$$

Def. (Convergence in q^{th} moment): For $q > 0$, we say that $\{X_n : n \geq 1\}$ converges in the q^{th} moment to X if:

$$E[|X_n - X|^q] \to 0$$

Convergence in q^{th} moment implies that $X_n \xrightarrow{p} X$.[27,28]

E.g. (Convergence in probability doesn't imply convergence in q^{th} moment): For example, pick $X_n = n$ with probability $\frac{1}{n}$ and 0 otherwise. Check that $X_n \to X$ but that with $q = 1$:[29]

$$E[|X_n - X|] = E[|X_n|] = E[X_n] = 1 \neq 0 = E[X]$$

[26] Seems like we're just moving to random variables in this step.

[27] Proof? Use Markov's inequality.

[28] The converse is false.

[29] Work this out. He told us to check this in class and we know how that sort of thing works!

Convergence in distribution

HAVING a less restrictive sense of convergence than convergence in probability would be useful because at this point there's very little we can say about the limiting distribution of parameters. Sure the WLLN tells us that an analog to a parameter converges in probability to that parameter, but it would be nice if we could put some sort of weight on how certain we are that an estimator jives with a hypothesis we have. Convergence in distribution (in concert with convergence in probability) allows us to do just that.

Def. (Convergence in Distribution): Say a sequence of random vectors $\{X_n : n \geq 1\}$ converges in distribution to another random vector X iff $P\{X_n \leq x\} \to P\{X \leq x\}\ \forall\ x$ at which $P\{X \leq x\}$ is continuous, where the $\leq$ is assessed component by component of the vectors in question.

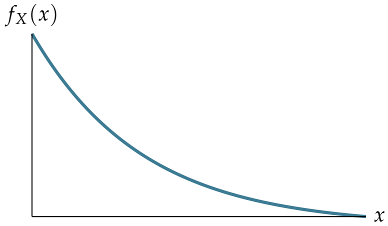

Figure 6: PDF of $X \sim P = Exp\left(\frac{1}{2}\right)$.

One note about "common" distributions like, $\mathcal{N}(0,1)$ or Exp(λ) is that we'll write:

$$X_n \overset{d}{\to} \mathcal{N}(0,1).$$

E.g. (Importance of Continuity Assumption): The definition of convergence in distribution requires convergence only at values of x where $P\{X \leq x\}$ is continuous. For example, take $X_n = \frac{1}{n}$ and $X = 0$. Then the c/d/f of X_n is continuous everywhere except $x = 0$. Then:

$$P\{X_n \leq x\} \to P\{X \leq x\}\ \forall\ x \neq 0$$

but

$$0 = P\{X_n \leq 0\} \not\to P\{X \leq 0\} = 1$$

where $P\{X = 0\}$ is discontinuous because it jumps up to 1 at $x = 0$.

Convergence is distribution is super important, but the definition above isn't that easy to work with. Luckily there's a lemma that establishes some easier ways to work with convergence in distribution.

Lem. (Portmanteau's Lemma): *The following are equivalent:*

1. $X_n \overset{d}{\to} X$
2. $E[f(X_n)] \to E[f(X)]$ *for all continuous, bounded, and real valued functions,* $f(\cdot)$.
3. $E[f(X_n)] \to E[f(X)]$ *for all bounded Lipschitz,*[30] *real valued functions,* $f(\cdot)$.
4. $\liminf_{n\to\infty} E[f(X_n)] \geq E[f(x)]$ *for all non-negative, continuous, real valued functions,* $f(\cdot)$.

[30] $f(\cdot)$ is Lipschitz with constant L iff $|f(x) - f(y)| \leq L|x - y|$.

5. $\liminf_{n\to\infty} P\{X_n \in G\} \geq P\{X \in G\}$ *for all open sets* G.
6. $\limsup_{n\to\infty} P\{X_n \in H\} \leq P\{X \in H\}$ *for all closed sets* H.
7. $P\{X_n \in B\} \to P\{X \in B\}$ *for all Borel sets* B *with* $P\{X \in bdd(B)\} = 0$.

Azeem comments that the first and last are the most useful, even though we'll exclusively (spoiler alert!) work with the first and second or the first and third in these notes.

Convergence in distribution and probability

It isn't immediately apparent how convergence in probability and distribution interact with each other but it's critical for understanding limiting distributions. Conceptually, probability is the stronger condition, but this isn't always true. If this intro seems muddled and confusing it's because the issue isn't that straightforward. Let's just delve into some theorems and lemmas that relate the two concepts:

Lem.: *Let* $\{X_n : n \geq 1\}$ *be a sequence of random vectors and* X *another random vector s/t* $X_n \overset{d}{\to} X$. *Let* Y_n *be a sequence of random vectors s/t* $Y_n - X_n \overset{p}{\to} 0$. *Then* $Y_n \overset{d}{\to} X$.

Proof. By Portmanteau's Lemma it's enough to show that:

$$E[f(Y_n)] \to E[f(X)]$$

for all bounded Lipschitz real-valued $f(\cdot)$ bounded by B. Then our desire is to show that there exists an N s/t $\forall n > N, \epsilon > 0$:

$$|E[f(Y_n)] - E[f(X)]| < \epsilon$$

so starting with our desired statement and adding 0 and applying the $\triangle$ inequality we get:

$$|E[f(Y_n)] - E[f(X))]| \leq \underbrace{|E[f(Y_n)] - E[f(X_n))]|}_{\text{Hmm}} + \underbrace{|E[f(X_n)] - E[f(X))]|}_{\to 0 \text{ b/c } X_n \overset{d}{\to} X \text{ and P.'s Lem.}}$$

and we can play with our Hmmm term to get:

$$\begin{aligned}|E[f(Y_n)] - E[f(X_n))]| &\leq |E[f(Y_n) - f(X_n)]| \\ &\leq E[|f(Y_n) - f(X_n)|] \text{ by the } \triangle \text{ inequality} \\ &= E[|f(Y_n) - f(X_n)|]I\{|Y_n - X_n| \leq \epsilon\} + E[|f(Y_n) - f(X_n)|]I\{|Y_n - X_n| > \epsilon\} \\ &\leq L\epsilon + 2BE[I\{|Y_n - X_n| > \epsilon\}] \text{ b/c } f(\cdot) \text{ is bounded by } B \\ &\leq L\epsilon + 2B\underbrace{P\{|Y_n - X_n| > \epsilon\}}_{\to 0 \text{ by assumption}} \text{ because E of I's are P's} \\ &\to 0 \text{ by choice of small enough } \epsilon\end{aligned}$$

and then we've taken that Hmmmm term to 0 and we've established our desired result. □

Lem.: *Take* $\{X_n : n \geq 1\}$. *If* $X_n \xrightarrow{p} X$ *then* $X_n \xrightarrow{d} X$.

Proof. Let $Y_n = X_n$, $X_n = X$, and $X = X$ in the lemma above. Then we have that $X \xrightarrow{d} X$ and $X_n - X \xrightarrow{p} 0$. Then applying the lemma, we get that $X_n \xrightarrow{d} X$, which is the conclusion of the claim. Boom. □

However, generally the converse is not true.

E.g.: Define X as follows:

$$X = \begin{cases} 1 & \text{, with } p = \frac{1}{2} \\ -1 & \text{, o/w} \end{cases}$$

and let $X_n = -X$. Then dwelling on what the CDF for X looks like, you'll arrive at:[31]

$$P\{X \leq x\} = \begin{cases} 1 & , \; x \geq 1 \\ \frac{1}{2} & , \; x \in [-1, 1) \\ 0 & , x < -1 \end{cases}$$

and if you do the same for X_n you'll find that:

$$P\{X_n \leq x\} = P\{X \leq x\}, \; \forall x.$$

Then we have that $X_n \xrightarrow{d} X$, but:[32]

$$P\{|X_n - X| > 1\} = 1$$

which tells us that $X_n \overset{p}{\not\to} X$.

But there is one important exception:

Lem.: *Take* $\{X_n : n \geq 1\}$ *and c a non-random vector (constant) with* $X_n \xrightarrow{d} c$ *then* $X_n \xrightarrow{p} c$.

Proof. Starting from $X_n \xrightarrow{d} c$ we can apply P.'s Lem. to obtain the fact that:

$$\limsup_{n \to \infty} P\{X_n \in H\} \leq P\{c \in H\}$$

where H is any closed set. Now, recall what we n/t/s:

$$\forall \, \epsilon > 0, \; P\{|X_n - c| > \epsilon\} \to 0$$

which is an open set. But we can easily turn it into a closed set with the following move:

$$P\{|X_n - c| > \epsilon\} \leq P\{|X_n - c| \geq \epsilon\}$$

[31] In words, this just says that X isn't going to less than -1, it's not going to be bigger than 1, and if it's between -1 and 1 then it's a 50/50 proposition.

[32] Where we've fixed $\epsilon = 1$ and shown that there's clearly no freedom to choose a δ that'll satisfy the required definition of convergence in probability.

and from P.'s Lem. we know that for this closed set the following holds:

$$\limsup_{n\to\infty} P\{|X_n - c| \geq \epsilon\} \leq P\{|c - c| \geq \epsilon\} = 0$$

which establishes that we get convergence in probability because the results have held for any $\epsilon > 0$.[33] □

Another important distinction between convergence in probability and convergence in distribution is that marginal convergence does not imply joint convergence in distribution.

E.g. (Marginal and Joint Convergence in Distribution): Take:

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_n \\ \rho_n & 1 \end{bmatrix} \right)$$

with $\rho_n = -1^n$. Then $X_n \xrightarrow{d} \mathcal{N}(0,1)$ and $Y_n \xrightarrow{d} \mathcal{N}(0,1)$ but the joint distribution's covariance terms will never settle down and there is no convergence in distribution.

And of course, there's an exception to this result, which we establish with the next Lemma.

Lem.: *Let $\{(X_n, Y_n) : n \geq 1\}$ be a sequence of random vectors with $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, c a constant, then:*

$$(X_n, Y_n) \xrightarrow{d} (X, c).$$

Proof. From an earlier lemma we have that:

$$Y_n \xrightarrow{p} c$$

and we also have that:

$$|(X_n, Y_n) - (X_n, c)| = |Y_n - c| \xrightarrow{p} 0$$

because there's no distance between X_n and itself.[34] So it's enough to show that $(X_n, c) \xrightarrow{d} (X, c)$.

Then using P.'s Lem., it's enough to show that for all bounded, continuous, and real-valued $f(\cdot)$ that:

$$E[f(X_n, c)] \to E[f(X, c)].$$

But from P.'s Lem. we already have that:

$$E[f(X_n)] \to E[f(X)]$$

which we can use to get the desired result by picking the functions, $f(\cdot, c)$, from those that the condition without c that are continuous, bounded, and real-valued. Then we've satisfied P.'s Lem. for $f(\cdot, c)$, giving us $(X_n, c) \xrightarrow{d} (X, c)$, which gives us the conclusion we wanted. □

For those scoring at home, this proof really goes the extra mile to use the Portmanbro Lemma as many times as possible.

[33] And $\limsup P\{|X_n - c| > \epsilon\} < \delta$ is equivalent to $\lim P\{|X_n - c| > \epsilon\} < \delta$?

[34] This statement is a little jenky because it doesn't note the lack of distance in the first argument. I could imagine writing:

$$|(X_n, Y_n) - (X_n, c)| = |(0, Y_n - c)|$$

instead, but I guess this other approach is more notationally convienient.

Thm. (Continuous Mapping Theorem (CMT2)): *Let $\{X_n : n \geq 1\}$ be a sequence of random vectors on $\mathbb{R}^k$ and X another random vector also on $\mathbb{R}^k$ such that: $X_n \overset{d}{\to} X$. Also, suppose there is a function, $g : \mathbb{R}^k \to \mathbb{R}^d$, which is continuous at each point $x \in C$ and $P\{X \in C\} = 1$. Then:*

$$g(X_n) \overset{d}{\to} g(X).$$

Proof. By P.'s Lem., it's enough to show that for any closed set, H:

$$\limsup_{n\to\infty} P\{g(X_n) \in H\} \leq P\{g(X) \in H\}$$

but we don't really know anything about $g(X)$, so it's useful to relate it to $g(X_n)$. Note:

$$g(X_n) \in H \iff X_n \in g^{-1}(H)$$

where $g^{-1}(H) = \{x \in \mathbb{R}^k : \ g(x) \in H\}$ and to make sure that we're dealing with a closed set, we note that:

$$g^{-1}(H) \subseteq cl(g^{-1}(H)) \subseteq g^{-1}(H) \cup C^c$$

where C is a set of continuity points. Then take $x \in cl(g^{-1}(H))$.[35] Then there exists an $x_n \in g^{-1}(H)$, s/t $x_n \to x$ by the definition of $cl(\cdot)$. If $x \notin C$ then we're done because then $x \in C^c$. If $x \in C$, then by continuity, $g(x_n) \to g(x) \in H$, because H is a closed set and thus contains all of its limit points, so $x \in g^{-1}(H)$.

Then:

$$\begin{aligned}
\limsup_{n\to\infty} P\{g(X_n) \in H\} &= \limsup_{n\to\infty} P\{X_n \in g^{-1}(H)\} \\
&\leq \limsup_{n\to\infty} P\{X_n \in cl(g^{-1}(H))\} \\
&\leq P\{X \in cl(g^{-1}(H))\} \\
&\leq P\{X \in g^{-1}(H) \cup C^c\} \\
&= P\{g(X) \in H)
\end{aligned}$$

and then applying P.'s Lem., we get $g(X_n) \overset{d}{\to} g(X)$. □

A related Lemma to the CMT2 that's probably used way more than it should goes as follows:

Lem. (Slutsky's Lemma): *Let $Y_n \overset{p}{\to} c$, where c is a constant on $\mathbb{R}$, and $X_n \overset{d}{\to} X$, also on $\mathbb{R}$. Then:*

1. $X_n + Y_n \overset{d}{\to} X + c$
2. $X_n Y_n \overset{d}{\to} Xc$
3. $X_n / Y_n \overset{d}{\to} X/c$ *if* $c \neq 0$.

[35] $cl(A) = \{x : \ \exists x_n \in A, x_n \to x\}$.

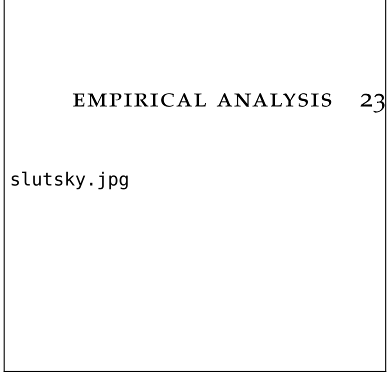

Figure 7: Eugen Slutsky. Gotta respect that hair. Damn.

Proof. $Y_n \xrightarrow{p} c \implies Y_n \xrightarrow{d} c$ so $(X_n, Y_n) \xrightarrow{d} (X, c)$ and apply CMT2 with $g(x,y) = x + y$ or $g(x,y) = xy$, etc. □

Why do we use this bad boy so often? That'll become more apparent once we jump into the next theorem.

When we were working with convergence in probability we had a property that allowed us to relate sample averages to their expectations. Thus far we have no tools that allow us to connect sample moments to a distribution, so you might be like, "Brandon, I've got a question: Who cares about this CMT2 or Slutsky's Lemma because we've got no limiting distributions to mess around with?" That's totally fair, which is why we introduce the following game-changer.

Thm. (Univariate Central Limit Theorem): *Let $X_1, \dots, X_n$ be sequence of random variables that are iid according to P and suppose $\sigma^2(P) < \infty$ then:*

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(P)).$$

Thm. (Cramer-Wold Device): *Let $\{X_n : n \geq 1\}$ be a sequence of random vectors and X a random vector, both in $\mathbb{R}^k$. Then:*

$$X_n \xrightarrow{d} X \iff \forall t \in \mathbb{R}^k \; t'X_n \xrightarrow{d} t'X.$$

That's Wold, not Wald of Transylvanian Abraham Wald Test fame. And Cramer, not Kramer of Seinfeld fame.

Thm. (Multivariate Central Limit Theorem): *Let $\{X_n : n \geq 1\}$ be a sequence of random vectors in $\mathbb{R}^k$ distributed according to P. Furthermore, suppose $\Sigma(P) < \infty$ then:*[36]

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma(P)).$$

Proof. By Cramer-Wold, enough to show:

$$\begin{aligned} t'[\sqrt{n}(\bar{X}_n - \mu(P))] &\xrightarrow{d} t'\mathcal{N}(0, \Sigma(P)) \\ &= \mathcal{N}\left(0, t'\Sigma(P)t\right) \quad \text{which we should check,} \end{aligned}$$

and:

$$t'[\sqrt{n}(\bar{X}_n - \mu(P))] = \sqrt{n}\left(\frac{1}{n}\sum_i^n t'X_i - t'\mu(P)\right) \xrightarrow{d} \mathcal{N}\left(0, t'\Sigma(P)t\right)$$

because $E[t'X_i] = t'\mu(P)$ and $Var[t'X_i] = t'\Sigma(P)t$. □

E.g. (Application of CMT2): Suppose $X_1, \dots, X_n \overset{iid}{\sim} P$ with $\sigma^2(P) < \infty$ then the CLT tells us that:

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(P)).$$

[36] Recall,

$$\Sigma(P) = E[(X_i - \mu(P))(X_i - \mu(P)'].$$

clt.pdf

Figure 8: Central Limit Theorem in action. Each panel plots the density of n observations of $\bar{X}_{100}$ 25 times, where $X_i \sim \text{Bernoulli}(0.3)$. Panel A has $n = 10$. Panel B has $n = 50$ and Panel C has $n = 1000$. Looks pretty normal to me.

Recall that above we proved that $S_n^2 \xrightarrow{p} \sigma^2(P)$. Then using the CMT1 we have that:

$$1/S_n \xrightarrow{p} 1/\sigma(P)$$

which requires us to assume that $\sigma^2(P) > 0$ to give us the continuity requirement. Then by Slutky's Lemma we have:

$$\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} \xrightarrow{d} \frac{1}{\sigma^2(P)} \mathcal{N}(0, \sigma^2(P)) = \mathcal{N}(0,1)$$

where $Z := \mathcal{N}(0,1)$.

Hypothesis testing

Consider testing the null hypothesis, $H_0 : \mu(P) \leq 0$, vs. an alternative hypothesis, $H_A : \mu(P) > 0$ and trying to minimize the following types of error:

Type I error: Rejecting H_0 when it is actually true.

Type II error: Failing to reject H_0 when it is false.

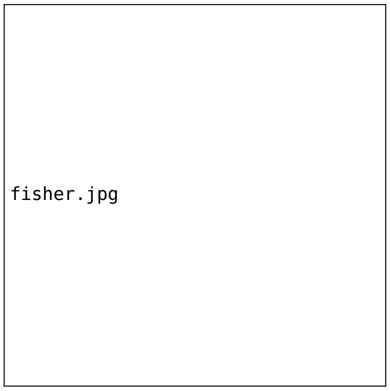

Figure 9: R.A. Fisher looking like a huge nerd at a eugenics conference. Creepy. Anyways, Fisher's credited with the term "test of significance." He also did a bunch of other useful stuff. Real bummer about the eugenics stuff, though.

More formally, a *test* is a function, $\phi_n := \phi_n(X_1, \ldots, X_n)$, that takes values in $[0,1]$ and equals the probability with which we should reject the null hypothesis. Generally we only consider tests, $\phi_n \in \{0,1\}$.

We'll think about *power* as: $E[\phi_n] = E_P[\phi_n]$ when viewed as a function of the distribution, P, this is called the power function of a test and it equals the probability of rejecting the null when the distribution is actually P. For P satisfying the null hypothesis, $E_P[\phi_n]$ is the probability of a Type I error. For P satisfying the alternative hypothesis, $1 - E_P[\phi_n]$ is the probability of a Type II error.

The customary solution to picking ϕ_n is to restrict attention to tests that are *consistent in level*, for P satisfying the null hypothesis:

$$\limsup_{n\to\infty} E_P[\phi_n] \leq \alpha$$

for $\alpha \in (0,1)$ where α is the significance level. Subject to this constraint, the goal of most test statistics is to make the probability of a Type II error "small." In this class, we'll further restrict our attention to tests of the form:[37]

$$\phi_n = I\{T_n > c_n\}$$

where T_n is our test statistic, which will be a function of the data $(X_1, \ldots, X_n)$, s/t large values provide evidence against H_0 and c_n is our critical value, which provides a definition of too large. T_n is assessed under the assumptions of the null hypothesis (e.g., $\mu(P) \leq 0$).

[37] This formulation is convienent because taking expectations of indicator functions yields probabilities. In that vein:

Def. (ϕ_n is Consistent in Level): The hypothesis test $\phi_n = I\{T_n > c_n\}$ is consistent in level i/f/f:

$$\limsup_{n\to\infty} P\{\phi_n = 1\} \leq \alpha$$

under the assumption of H_0.

E.g. (Z-Test): Picking up where we left off with our last example (Application of CMT2), a reasonable test statistic would be:

$$T_n = \frac{\sqrt{n}\bar{X}_n}{S_n}$$

and a reasonable critical value would be:

$$c_n = z_{1-\alpha} = \Phi^{-1}(1-\alpha).$$

Are these consistent in level? Pick any P satisfying the null. Then we w/t/s:

$$\limsup_{n\to\infty} E_P[\phi_n] = \limsup_{n\to\infty} P\{T_n > c_n\} \leq \alpha.$$

Plugging in our choices of test statistic and critical value, we get:

$$\begin{aligned} P\{T_n > c_n\} &= P\left\{\frac{\sqrt{n}\bar{X}_n}{S_n} > z_{1-\alpha}\right\} \\ &= P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} + \underbrace{\frac{\sqrt{n}\mu(P)}{S_n}}_{\leq 0 \text{ for } P \text{ under } H_0} > z_{1-\alpha}\right\} \\ &\leq P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} > z_{1-\alpha}\right\} \end{aligned}$$

Then taking $\limsup$ on both sides we get:

$$\begin{aligned} \limsup_{n\to\infty} P\{T_n > c_n\} &\leq \limsup_{n\to\infty} P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} > z_{1-\alpha}\right\} \\ &= 1 - \liminf_{n\to\infty} P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} \leq z_{1-\alpha}\right\} \\ &= 1 - \Phi(z_{1-\alpha}) \;\; \text{b/c } T_n \xrightarrow{d} Z \;\&\; P\{Z \leq z_{1-\alpha}\} = \Phi(z_{1-\alpha}) \\ &= \alpha \end{aligned}$$

which establishes that the Z-Test is consistent in level.

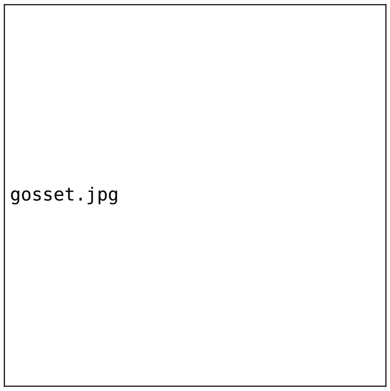

Figure 10: William Sealy Gosset, AKA Student, looking a little schlitzed. E.L. Lehmann notes: The term "studentization" is a misnomer. The idea of replacing $z_{1-\alpha}$ with $t_{n-1,1-\alpha}$ was already used by Laplace. Student's contribution was to work out the *exact* distribution in the one-sample situation.

E.g. (T-Test is Consistent in Level): The T-test is:

$$\begin{aligned} \phi_n &= I\{T_n > c_n\} \\ &= I\left\{\frac{\sqrt{n}\bar{X}_n}{S_n} > t_{n-1,1-\alpha}\right\} \end{aligned}$$

Recall from earlier that our general strategy with this sort of example is to establish the convergence properties of T_n and c_n so that we can use:[38]

$$T_n \xrightarrow{d} T \text{ and } c_n \xrightarrow{d} c \implies P\{T_n \leq c_n\} \to P\{T \leq c\}.$$

[38] Proof? Suppose $T_n \xrightarrow{d} T$ and $c_n \xrightarrow{d} c$. Then how can we show that:

$$P\{T_n \leq c_n\} \to P\{T \leq c\}?$$

First note that $T_n - c_n \xrightarrow{d} T - c$ by CMT2 because both pieces converge in distribution and the function is continuous. Then by the definition of convergence in distribution we have:

$$P\{T_n - c_n \leq x\} \to P\{T - c \leq x\}$$

which if we pick $x = 0$ implies:

$$P\{T_n - c_n \leq 0\} \to P\{T - c \leq 0\}$$

We already know that we can mess around with T_n to get an estimator that converges in distribution to Z, so what can we do about c_n?

$$t_n \overset{d}{=} \frac{Z}{\sqrt{\chi^2_n/n}}$$

and:

$$\frac{1}{n}\chi^2_n = \frac{1}{n}\sum_{i=1}^{n} z_i^2 \xrightarrow{p} \mathbf{E}Z_i = 1$$

so the numerator of t_n converges in distribution to Z and the denominator converges in probability to 1. Then by Slutsky's Lemma we have:

$$t_n \xrightarrow{d} Z$$

and the rest of showing that the T-Test is consistent in level is pretty obvious:

$$\begin{aligned}
\limsup_{n\to\infty}\ P\{T_n > c_n\} &= \limsup_{n\to\infty}\ P\left\{\frac{\sqrt{n}\bar{X}_n}{S_n} > t_{n-1,1-\alpha}\right\} \\
&= \limsup_{n\to\infty}\ P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} + \frac{\sqrt{n}\mu(P)}{S_n} > t_{n-1,1-\alpha}\right\} \\
&\leq \limsup_{n\to\infty}\ P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} > t_{n-1,1-\alpha}\right\} \text{ because under } H_0 \text{ we have: } \mu(P) \leq 0 \\
&= 1 - \liminf_{n\to\infty}\ P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{S_n} \leq z_{1-\alpha}\right\} \\
&= 1 - \Phi(z_{1-\alpha}) \text{ because } T_n \xrightarrow{d} Z, t_{n-1,1-\alpha} \xrightarrow{d} z_{1-\alpha}, \text{ and } P\{Z \leq z_{1-\alpha}\} = \Phi(z_{1-\alpha}) \\
&= \alpha.
\end{aligned}$$

So we're good.

E.g. (Chi-Square Test): Suppose we have $X_1, \dots, X_n$ an iid sequence of random vectors on $\mathbb{R}^k$ distributed according to P. Also, suppose that the variance-covariance matrix, $\Sigma(P) < \infty$. How should we go about constructing T_n and c_n?

Well, from the Multivariate Central Limit Theorem we have:

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma(P)).$$

If $\Sigma(P)$ is non-singular then:

$$n(\bar{X}_n - \mu(P))'\Sigma(P)^{-1}(\bar{X}_n - \mu(P)) \sim \chi^2_k$$

where Z is a multivariate normal with mean zero and variance-covariance matrix $\Sigma(P)$.[39] Then the Continuous Mapping Theorem tells us that:

$$n(\bar{X}_n - \mu(P))'\Sigma(P)^{-1}(\bar{X}_n - \mu(P)) \xrightarrow{d} \chi^2_k$$

[39] This result can be derived using a bit of linear algebra. First take the Cholesky decomposition of $\Sigma(P) = \Lambda\Lambda'$. Then from a property of triangular inversion: $\Sigma^{-1}(P) = \Lambda^{-1'}\Lambda^{-1}$. Then premultiplying our application of the multivariate CLT we get:

$$\begin{aligned}
&\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma(P)) \\
&\implies \Lambda^{-1}\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Lambda^{-1}\Sigma(P)\Lambda^{-1'}) \\
&\implies \Lambda^{-1}\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Lambda^{-1}\Lambda\Lambda'\Lambda^{-1'}) \\
&\implies \Lambda^{-1}\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, I)
\end{aligned}$$

but we can actually do one better than just pretend that we have $\Sigma(P)$. How? Well recall our definition of the sample variance-covariance:

$$\hat{\Sigma}(P) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)'$$

Then we could show (but we won't) that:

$$\hat{\Sigma}(P) \xrightarrow{p} \Sigma(P)$$

and as long as our estimator for the variance-covariance matrix is invertible the CMT2 gives us:

$$n(\bar{X}_n - \mu(P))'\hat{\Sigma}(P)^{-1}(\bar{X}_n - \mu(P)) \xrightarrow{d} \chi^2_k.$$

So now that we've pinned down the distribution of our data, how should we go about testing $H_0 : \mu(P) = 0$ against $H_A : \mu(P) \neq 0$ while satisfying:

$$P\{T_n > c_n) \to \alpha \text{ if } H_0 \text{ is true?}$$

Just choose:

$$T_n = n\bar{X}_n'\hat{\Sigma}(P)^{-1}\bar{X}_n \text{ because under } H_0 \text{ we have: } \mu(P) = 0$$

and:

$$c_n = c_{k,1-\alpha}$$

which is the $(1-\alpha)$th quantile of the χ^2_k distribution.

P-values

SOMETIMES we'll want to say more than just whether our function ϕ_n has returned a 0 or 1 for our choice of α. In particular, we might want to know the α at which our test switches from returning a 0 to returning a 1. This α is called the p-value.[40]

Def. (P-value): The smallest value of α for which we can reject H_0, or the p-value, is:

$$p_n := \inf\{\alpha \in (0,1) \ : \phi_n = 1\}$$

E.g. (p-value for Z-test):

$$\begin{aligned}\hat{p}_n &:= \inf\left\{\alpha \in (0,1) \ : \ \frac{\sqrt{n}\bar{X}_n}{S_n} > \Phi^{-1}(1-\alpha)\right\} \\ &= \inf\left\{\alpha \in (0,1) \ : \ \Phi\left(\frac{\sqrt{n}\bar{X}_n}{S_n}\right) > (1-\alpha)\right\} \\ &= \inf\left\{\alpha \in (0,1) \ : \ \alpha > 1 - \Phi\left(\frac{\sqrt{n}\bar{X}_n}{S_n}\right)\right\} \\ &= 1 - \Phi(T_n)\end{aligned}$$

[40] E.g., α is less than or equal to 0.05.

Delta method

THUS far we have a limited number of tools to deal with changing the dimensionality of our data. The CMT2 helps, but in conjunction with the CLT, it's not always straightforward what the variance-covariance structure looks like for the limiting distribution. This is where the Delta Method comes in handy.

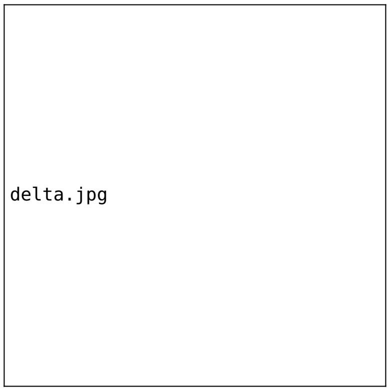

Figure 11: Hold on! I have something for this.

Thm. (Delta Method): *Let $\{X_n : n \geq 1\}$ be a sequence of random vectors on $\mathbb{R}^k$. Let $\{\tau_n\}$ be a sequence of real numbers converging to ∞ and define a vector of constants, c such that:*

$$\tau_n(X_n - c) \xrightarrow{d} X.$$

Define the function, $g : \mathbb{R}^k \to \mathbb{R}^d$ as differentiable at c and denote the (d, k) matrix of partial derivatives of g evaluated at c as: $Dg(c)$. Then:

$$\tau_n(g(X_n) - g(c)) \xrightarrow{d} Dg(c)X.$$

Proof. The place to start is the assumption that $g(\cdot)$ is differentiable at c. Then by Taylor's Theorem we get:

$$g(x) = g(c) + Dg(c)(x - c) + R(x - c)$$

where $R(\cdot)$ is the remainder term and here $R(h) = o(|h|)$.[41] Then:

$$\tau_n(g(X_n) - g(c)) = \underbrace{Dg(c)\underbrace{\tau_n(X_n - c)}_{\xrightarrow{d} X}}_{\xrightarrow{d} Dg(c)X \text{ by CMT}} + \tau_n R(X_n - c)$$

and now we n/t/s:

$$\tau_n R(X_n - c) \xrightarrow{d} 0$$

Then:

$$\tau_n R(X_n - c) = \underbrace{\tau_n |X_n - c|}_{\xrightarrow{d} X \text{ by CMT}} \underbrace{\frac{R(X_n - c)}{|X_n - c|}}_{\xrightarrow{d} 0 \text{ by CMT}} \xrightarrow{d} 0$$

by CMT. □

Aside:

$$X_n - c = \underbrace{\tau_n(X_n - c)}_{\xrightarrow{d} X} \underbrace{\frac{1}{\tau_n}}_{\to 0} \xrightarrow{d} 0.$$

The example below gives a sense of how the notation works for applying the Delta Method.

E.g. (Multivariate Normal): If $X \xrightarrow{d} \mathcal{N}(0, \Sigma)$ then the Delta Method tells us that:

$$\tau_n(g(X_n) - g(c)) \xrightarrow{d} \mathcal{N}(0, Dg(c)\Sigma Dg(c)').$$

[41] $R(h) = o(|h|)$ means that $\frac{R(h)}{|h|} \to 0$ as $|h| \to 0$, with $\frac{R(h)}{|h|} = 0$ if $h = 0$.

Using this we can test $H_0 : \mu(P) = 0$ vs. $H_A : \mu(P) \neq 0$ at level α and construct C_n s/t:

$$P\{\mu(P) \in C_n\} \to 1 - \alpha$$

which highlights the duality b/t hypothesis testing and confidence regions.

One interesting observation about the Delta Method is that the theorem is even valid if $Dg(c) = 0$. In that case $\tau_n(g(X_n) - g(c)) \xrightarrow{d} 0$. That said, it can suck if $Dg(c) = 0$ because then our estimators are converging to a degenerate distribution. Consider the example below and the proposed solution when this crops up:

E.g. (Bernoulli Distribution): Take $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(q)$ where $q \in (0, 1)$. Then the Central Limit Theorem tells us:

$$\sqrt{n}(\bar{X}_n - q) \xrightarrow{d} \mathcal{N}(0, q(1-q))$$

but suppose we're interested in the distribution of the sample variance: $g(\bar{X}_n) - g(q) = \bar{X}_n(1 - \bar{X}_n) - q(1-q)$. Well, we can use the Delta Method:

$$\begin{aligned} \sqrt{n}(g(\bar{X}_n) - g(q)) &\xrightarrow{d} \mathcal{N}(0, Dg(q)g(q)Dg(q)') \\ &\mathcal{N}(0, (1-2q)^2(1-q)q) \;\; \text{because } X_i \text{ is a scalar and } g'(q) = 1 - 2q. \end{aligned}$$

But what if we have that $q = \frac{1}{2}$? Then we'd have:

$$\sqrt{n}(g(\bar{X}_n) - g(q)) \xrightarrow{d} \mathcal{N}(0, 0)$$

which pretty much sucks.

What to do? Take a second-order Taylor Series Expansion of $g(x) - g(q)$:

$$g(x) - g(q) = Dg(q)(x - q) + \frac{D^2 g(q)}{2}(x - q)^2 + R(x - q)$$

where $R(\cdot)$ is the remainder term. We already know that $Dg(\frac{1}{2}) = 0$ and as usual we'll just neglect the remainder term.[42] Setting $x = \bar{X}_n$, $q = \frac{1}{2}$, and pre-multiplying by n for convenience we get:[43]

$$\begin{aligned} n(g(\bar{X}_n) - g(q)) &= -n\,(\bar{X}_n - q)^2 \\ &= -\left(\sqrt{n}(\bar{X}_n - q)\right)^2 \\ &\xrightarrow{d} -\left[\mathcal{N}(0, \frac{1}{4})\right]^2 \quad \text{because } q = \frac{1}{2} \\ &= -\left[\frac{1}{2}\mathcal{N}(0, 1)\right]^2 \\ &= -\frac{1}{4}\chi_1^2 \end{aligned}$$

[42] Recall that $R(h) = o(|h|^2)$, i.e., $\frac{R(h)}{|h|^2} \to 0$ as $h \downarrow 0$ with the fraction equal to 0 when $h = 0$. Just as we showed in our proof of the Delta Method, we get:

$$nR(\bar{X}_n - q) \to 0.$$

[43] So we can apply the CLT.

which tells us that if we pre-multiply by n, we get a full distribution of the estimator for the sample variance of a Bernoulli random variable. Neat and relatively easy trick.

E.g. (Correlations): Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P$ on $\mathbb{R}^2$ with $E[X_i^2] < \infty$ and $E[Y_i^2] < \infty$. Then the covariance of these two terms is just:

$$Cov(X_i, Y_i) = E[X_i Y_i] - E[X_i]E[Y_i]$$

and we know that $E[X_i Y_i]$ exists.[44] And if in addition, $Var(X_i) > 0$ and $Var(Y_i) > 0$ then:

$$Corr(X_i, Y_i) = \rho_{X,Y}(P) = \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i)Var(Y_i)}}.$$

Thm. (Cauchy-Schwarz Inequality): *For any random variables U and V s/t $E[U^2] < \infty$ and $E[V^2] < \infty$, then:*

$$E[UV]^2 \leq E[U^2]E[V^2].$$

Proof. By the same argument we used in the Correlations example, we have that $E[UV]$ exists. Either we're in world where $E[U^2] = 0$ or $E[V^2] = 0$ or we're in a world where both variances are non-zero. If either of the variances are zero then we're done because $0 = 0$.[45]

Now the hairier case assumes non-zero variances for both. Then consider $E[(U - \alpha V)^2] \geq 0$. We can expand and separate to get:

$$E[U^2] - 2\alpha E[UV] + \alpha^2 E[V^2] \geq 0$$

and also note that:

$$E[U^2] - 2\alpha E[UV] + \alpha^2 E[V^2] \geq \arg\min_{\alpha} E[U^2] - 2\alpha E[UV] + \alpha^2 E[V^2] \geq 0$$

so taking the FOC, setting equal to 0, and solving for α gives us:

$$\alpha = \frac{E[UV]}{E[V^2]}$$

which we can plug into the original expression to get:

$$E[U^2]E[V^2] \geq E[UV]^2$$

which was our goal!. □

Lem.: *Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P$ on $\mathbb{R}^2$ with $E[X_i^2] < \infty$ and $E[Y_i^2] < \infty$. Then:*

$$|\rho_{X,Y}(P)| \leq 1$$

and with equality i/f/f there exists a, b s/t $P\{a + bX_i = Y_i\} = 1$.[46]

[44] Why? Well first note that:

$$(|u| - |v|)^2 \geq 0$$

which we can expand to get:

$$|uv| \leq \frac{1}{2}|u|^2 + \frac{1}{2}|v|^2$$

and we can take expectations and pick a convenient choice of u and v to get:

$$E[|X_i Y_i|] \leq \frac{1}{2}E[X_i^2] + \frac{1}{2}E[Y_i^2] < \infty.$$

[45] Does $E[U^2] = 0 \implies E[UV]^2 = 0$?

[46] That is, there is a perfect linear relationship between X_i and Y_i.

Proof. First we'll look at the claim that:

$$|\rho_{X,Y}(P)| \leq 1$$

This follows from expanding $\rho_{X,Y}(P)$ and applying Cauchy-Schwarz by choice of $U = X_i - E[X_i]$ and $V = Y_i - E[Y_i]$.

What about the claim w/r/t equality? Starting from:

$$E[UV]^2 = E[U^2]E[V^2] \iff \exists\ \alpha\ s/t\ P\{U = \alpha V\} = 1$$

we then just have to apply Cauchy-Schwarz with choices of:

$$U = X_i - E[X_i] \text{ and } V = Y_i - E[Y_i].$$

□

Suppose we wanted to estimate correlations or covariances? Well, the natural choice of an estimator for $Cov(X_i, Y_i)$ is:

$$\hat{\sigma}_{X,Y,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

and a natural estimator of $Corr(X_i, Y_i)$ is:

$$\hat{\rho}_{X,Y,n} = \frac{\hat{\sigma}_{X,Y,n}}{S_{Y,n}S_{X,n}}$$

so what can we say about:

$$\sqrt{n}(\hat{\rho}_{X,Y,n} - \rho_{X,Y,n}) \xrightarrow{d} ??$$

Well, we know that these are all smooth functions of sample averages so it'll converge to a normal with mean 0 and a god-awful variance because of the CLT and Delta Method. But the point is that without the Delta Method we'd have a hard time saying much of anything.

Thus far we've spent just about all of our time on sample averages.[47] This example will help us move beyond just that:

E.g. (Beyond Sample Averages): Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P$ on $\mathbb{R}$ and let F denote the CDF. Define:

$$\theta = \inf\{x \in \mathbb{R} :\ F(x) \geq 0.5\}$$

and further suppose that $F(\cdot)$ is differentiable at θ with $F'(\theta) = f(\theta) > 0$. Then the sample median is just:

$$\hat{\theta}_n = \inf\{x \in \mathbb{R} :\ \hat{F}_n(x) \geq 0.5\}$$

where $\hat{F}_n(x)$ is the empirical c/d/f and we want to show:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f^2(\theta)}\right).$$

[47] With the exception of that stuff about medians on the first problem set.

Proof. To show this we'll use a slightly stronger CLT than we're used to.[48] Then suppose n is even. Then we know that $\hat{\theta}_n = \frac{n}{2}^{th}$ highest observation of X_i. Then:

$$\begin{aligned}
P\{\sqrt{n}(\hat{\theta}_n - \theta) \leq x\} &= P\left\{\hat{\theta}_n \leq \theta + \frac{x}{\sqrt{n}}\right\} \text{ which is the \# of } X_i > \theta + \frac{x}{\sqrt{n}} \leq \frac{n}{2} - 1 \\
&= P\left\{\sum_{i=1}^{n} Z_i \leq \frac{n}{2} - 1\right\} \text{ when } Z_i = I\left\{X_i > \theta + \frac{x}{\sqrt{n}}\right\} \\
&= P\left\{\frac{\sqrt{n}(\bar{Z}_n - \mu_n)}{\sigma_n} \leq x_n\right\} \text{ where } \mu_n = E[Z_i] = 1 - F\left(\theta + \frac{x}{\sqrt{n}}\right) \to 0.5, \\
&\qquad x_n = \frac{\sqrt{n}(0.5 - 1/n - \mu_n)}{\sigma_n} = \frac{\sqrt{n}(0.5 - \mu_n)}{\sigma_n} - \frac{1/\sqrt{n}}{\sigma_n}, \\
&\qquad \text{and } \sigma_n^2 = \mu_n(1 - \mu_n) \to 0.25 \\
&= P\left\{\frac{\sqrt{n}(\bar{Z}_n - \mu_n)}{\sigma_n} \leq x_n\right\} - \Phi(x_n) + \Phi(x_n) \\
&\leq \underbrace{\frac{C}{\sqrt{n}} \frac{E[(Z_i - \mu_n)^2]}{\sigma_n^3}}_{\to 0} + \underbrace{\Phi(x_n)}_{\text{Hmmm}} \text{ by the B.E. CLT}
\end{aligned}$$

Then what to do with out Hmmmmm term? Well working with our broken-up x_n we get:

$$x_n = \underbrace{\frac{\sqrt{n}(0.5 - \mu_n)}{\sigma_n}}_{\to 2f(\theta)x} + \underbrace{\frac{1/\sqrt{n}}{\sigma_n}}_{\to 0}$$

and:

$$\sqrt{n}(0.5 - \mu_n) = \sqrt{n}(F(\theta + x/\sqrt{n}) - F(\theta)) \to f(\theta)x$$

so:

$$\Phi(x_n) \to \Phi(2f(\theta)x)$$

which is just the CDF of a $\mathcal{N}\left(0, \frac{1}{4f^2(\theta)}\right)$ which gives us the result we wanted because the Hmmmmmmm term from above is converging to our desired distribution and the other term in that inequality is going to 0.[49] □

Confidence intervals

Another popular way of testing a parameter is to construct a confidence interval around it. The general idea is to construct a set C_n for a choice of α such that:

$$P\{\theta(P) \in C_n\} \to 1 - \alpha$$

where $\theta(P)$ is the true value of our parameter of interest. Normally we'll think about constructing large-sample confidence intervals, but there's really no reason to limit our imagination:

∞. Then:

$$\sup_{x \in \mathbb{R}} \left| P\left\{\frac{\sqrt{n}(\bar{X}_n - \mu(P))}{\sigma(P)} \leq x\right\} - \Phi(x)\right| \leq \frac{C}{\sqrt{n}} \frac{E[(X_i - \mu(P))^2]}{\sigma^3(P)}$$

where C is given to you and doesn't depend on P or n.

Azeem writes in class that: $\frac{C}{\sqrt{n}} \frac{E[(Z_i - \mu_n)^2]}{\sigma_n^3} \to \frac{8C}{\sqrt{n}\sigma_n^3} \to 0$. Not clear to me where 8 is coming from as the numerator has a variance that's not 8 and the cubed term in the denominator is 8 and I dunno…

[49] We have not proved the result for n odd, but the proof follows the same logic.

E.g. (Finite Sample Confidence Interval for Bernoulli): Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(q)$ where $q \in (0,1)$ and we want to building a confidence set, $C_n = C_n(X_1, \ldots, X_n)$ such that:

$$P\{\mu(P) \in C_n\} \geq 1 - \alpha$$

for some choice of $\alpha \in (0,1)$.[50] It follows that from the Bernoulli distribution that $\mu(P) = q$.

For finite samples there's really only one tool at our disposal for a task like this, Markov's Inequality:

$$\begin{aligned} P\{|\bar{X}_n - \mu(P)| > \epsilon\} &\leq \frac{Var(\bar{X}_n - q)}{\epsilon^2} \text{ because } \mu(P) = q \\ &= \frac{Var(X_i)}{n\epsilon^2} \\ &= \frac{q(1-q)}{n\epsilon^2} \text{ because of the Bernoulli assumption} \\ &\leq \frac{1}{4n\epsilon^2} \text{ because } q(1-q) \in (0,1) \text{ and } x \leq x\frac{1}{q(1-q)}. \end{aligned}$$

Then we can ϵ such that:[51]

$$\alpha = \frac{1}{n\epsilon^2} \implies \epsilon = \sqrt{\frac{1}{4n\alpha}}$$

and if we plug in this choice of ϵ we get:

$$\begin{aligned} & P\left\{|\bar{X}_n - \mu(P)| > \sqrt{\frac{1}{n\alpha}}\right\} \leq \alpha \\ \implies & 1 - P\left\{|\bar{X}_n - \mu(P)| \leq \sqrt{\frac{1}{n\alpha}}\right\} \leq \alpha \\ \implies & P\left\{|\bar{X}_n - \mu(P)| \leq \sqrt{\frac{1}{n\alpha}}\right\} \geq 1 - \alpha \\ \implies & P\left\{-\sqrt{\frac{1}{n\alpha}} \leq \bar{X}_n - \mu(P) \leq \sqrt{\frac{1}{n\alpha}}\right\} \geq 1 - \alpha \\ \implies & P\left\{-\bar{X}_n - \sqrt{\frac{1}{n\alpha}} \leq -\mu(P) \leq -\bar{X}_n + \sqrt{\frac{1}{n\alpha}}\right\} \geq 1 - \alpha \\ \implies & P\left\{\bar{X}_n - \sqrt{\frac{1}{n\alpha}} \leq \mu(P) \leq \bar{X}_n + \sqrt{\frac{1}{n\alpha}}\right\} \geq 1 - \alpha \\ \implies & P\{\mu(P) \in C_n\} \geq 1 - \alpha \end{aligned}$$

where:

$$C_n = \left(\bar{X}_n - \sqrt{\frac{1}{n\alpha}}, \bar{X}_n + \sqrt{\frac{1}{n\alpha}}\right)$$

which is pretty neat.

[50] With the caveat that we can't just pick $C_n = [0,1]$.

[51] Fix rest of this for: $\alpha = \frac{1}{4n\epsilon^2}$ because $q(1-q) \leq \frac{1}{4}$!!

But what if we're interested in confidence intervals as n gets very large. That is, we want confidence sets that satisfy:

$$P\{\mu(P) \in C_n\} \to 1 - \alpha?$$

Let's consider a relatively simple example:

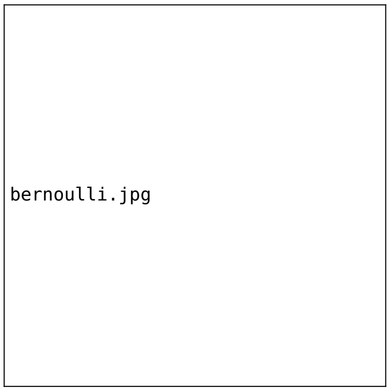

Figure 12: Jakob Bernoulli in all his snooty glory.

E.g. (Bernoulli Confidence Interval): To ape the setup of the finite sample example above: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(q)$ where $q \in (0,1)$ and we want to building a confidence set, $C_n = C_n(X_1, \ldots, X_n)$ such that:

$$P\{\mu(P) \in C_n\} \to 1 - \alpha$$

for some choice of $\alpha \in (0,1)$.

To start, note that:

$$\frac{\sqrt{n}(\bar{X}_n - q)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

which means that:

$$P\left\{\frac{\sqrt{n}(\bar{X}_n - q)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \geq x\right\} \to P\{Z \geq x\} \text{ by the definition of convergence in distribution}$$

and we'll make a convenient choice of x:

$$P\left\{z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}_n - q)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \leq z_{\alpha/2}\right\} = P\left\{\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \leq q \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}\right\}$$
$$\to P\{z_{1-\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1 - \alpha.$$

So our choice of C_n is:

$$C_n(X_1, \ldots, X_n) = \left\{y \in [0,1] : \ \bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \leq y \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}\right\}$$

and it has the desired property of consistency in level.

But we obviously don't observe infinite amounts of data and C_n may behave poorly in finite-samples. In particular, for any n and $\epsilon > 0$ there is a P (that is, a q) s/t $P\{\mu(P) \in C_N\} \leq \epsilon$. Let n, ϵ be given and choose $q = (1 - \epsilon)^{1/n}$. Then $P\{X_1 = 1, \ldots, X_n = 1\} = 1 - \epsilon$ and $X_1 = 1, \ldots, X_n = 1$ implies $C_n = \{1\}$ which implies that $q \notin C_n$. Then $1 - \epsilon \leq P\{q \neq C_n\}$ implies $P\{q \in C_n\} \leq \epsilon$.[52]

How much do we get when we move to asymptotic? Well let's compare the two confidence sets we've derived for $n = 100, \bar{X}_n = 0.5, \alpha = 0.05$:

$$\text{Finite: } C_n = (0.06, 0.94)$$

$$\text{Asymptotic: } C_n = (0.4, 0.6).$$

[52] This observation is very similar to the weak instruments literature.

So it obviously follows that we get a lot of power with our asymptotic.

Thus far we've only thought about confidence intervals for univariate problems. What if we move to higher dimensional problems?

E.g. (Confidence Interval for Multivariate Case): Suppose we have $X_1, \dots, X_n \overset{iid}{\sim} P$ where $X_i \in \mathbb{R}^k$ with $\Sigma(P) < \infty$ and non-singular[53] and we want to construct a confidence region C_n such that:

$$P\{\mu(P) \in C_n\} \to 1 - \alpha.$$

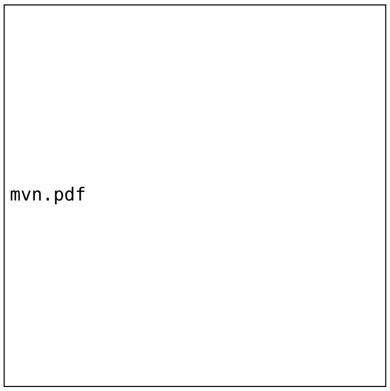

Figure 13: Several vantage points of 1000 i.i.d. draws from $\mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^3$. Sweeeeeet.

Recall that the CLT gives us:

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \Sigma(P))$$

then from non-singularity we get the following fact:

$$\text{For } \; Z \sim \mathcal{N}(0, \Sigma(P)) \implies Z'\Sigma^{-1}(P)Z \sim \chi^2_k.$$

Then by CMT2:

$$n(\bar{X}_n - \mu(P))'\Sigma(P)^{-1}(\bar{X}_n - \mu(P)) \xrightarrow{d} \chi^2_k$$

and we can use the analog principle for $\Sigma(P)$:

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)'$$

and then we can use the further fact that:

$$\hat{\Sigma}_n \xrightarrow{p} \Sigma(P) \implies \hat{\Sigma}_n^{-1} \xrightarrow{p} \Sigma(P)^{-1}$$

because $\Sigma(P)$ is non-singular.

So,

$$n(\bar{X}_n - \mu(P))'\hat{\Sigma}_n^{-1}(\bar{X}_n - \mu(P)) \xrightarrow{d} \chi^2_k$$

and we can use this to test the null, e.g., $H_0 : \mu(P) = 0$ vs. $H_A : \mu(P) \neq 0$ at level α with:

$$T_n = n\bar{X}_n'\hat{\Sigma}_n^{-1}\bar{X}_n$$

and:

$$c_n = c_{k,1-\alpha} = 1 - \alpha \; \text{ quantile of a } \chi^2_k$$

and it's easy to show that $P\{T_n > c_n\} \to 1 - \alpha$. Then:

$$1 - P\{T_n \leq c_n\} \to \alpha \implies P\{T_n \leq c_n\} \to 1 - \alpha.$$

In this spirit define:

$$C_n(X_1, \dots, X_n) = \left\{y \in \mathbb{R}^k : \; n(\bar{X}_n - y)'\Sigma(\hat{P})^{-1}(\bar{X}_n - y) \leq c_{k,1-\alpha}\right\}$$

and it follows from the earlier example that:

$$P\{\mu(P) \in C_n) \to 1 - \alpha.$$

[53] Equivalently it's strictly positive definite.

Tightness

LET's just jump into the definition on this fucker:

Def. (Tight): A sequence of random vectors, $\{X_n : n \geq 1\}$, is tight[54] if for every $\epsilon > 0$ there exists $B > 0$ such that:

$$\inf_n P\{|X_n| \leq B\} \geq 1 - \epsilon.$$

Equivalently, if for every $\epsilon > 0$ and every $n \geq 1$ there exists $B > 0$ such that:[55]

$$P\{|X_i| \leq B\} \geq 1 - \epsilon$$

which just says that the probability that any member of the sequence is less than some constant, B, won't exceed $1 - \epsilon$. Sometimes, though, it's convenient to re-arrange and express tightness with:

$$P\{|X_i| > B\} \leq \epsilon$$

which you can easily derive by taking the complement and multiplying out the negative.

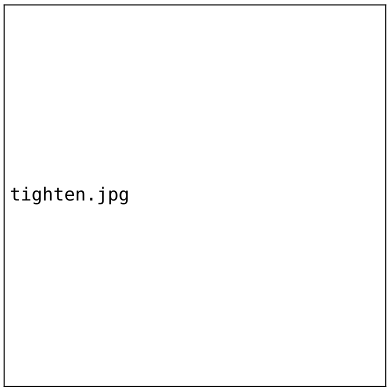

Figure 14: "We didn't say lose weight. . . I might say tighten."

Lem. (Convergence is Distribution $\implies$ Tight): *Suppose* $X_n \overset{d}{\to} X$. *Then* X_n *is tight.*

Proof. Fix ϵ and pick B such that:

$$P\{|X| > B\} \leq \epsilon.$$

Suppose $X_n \overset{d}{\to} X$. Then: $P\{|X_n| > B\} \to P\{|X| > B\}$. Then from the definition of limits, we know there exists some N such that for all $n \geq N$:

$$P\{|X_n| > B\} \leq \epsilon.$$

Then for all $n \in \{1, \ldots, N-1\}$ pick B' such that:

$$B' = \max_{n \in \{1,\ldots,N-1\}} \{|X_n|\} + \delta$$

where $\delta > 0$. Then:

$$P\{|X_n| > B'\} = 0 \text{ for } n \in \{1, \ldots, N-1\}.$$

Pick $B^* = \max\{B, B'\}$. Then for B^* we have that for all n:

$$P\{|X_n| > B^*\} \leq \epsilon.$$

Then X_n is tight. □

In general, the converse is not true, however, there is a sense in which tightness implies convergence in distribution:[56]

[54] Also known as bounded in probability. Recall that a sequence of non-random vectors, x_n, is bounded if $\exists\, M$ s.t.:

$$|x_n| \leq M \;\forall n.$$

[55] The non-random parallel is that if $x_n \to x$ then x_n is bounded.

[56] The Balzano-Weierstrass Theorem is the non-random equivalent, which tells us that if x_n is bounded then $\exists\, n_j$ and x s.t. $x_{n_j} \to x$.

Thm. (Prokhorov's Theorem): *If* $\{X_n : n \geq 1\}$ *is a tight sequence of random vectors, then there exists a subsequence,* n_j *and a random vector* X *such that:*

$$X_{n_j} \xrightarrow{d} X$$

E.g. (Counterexample of Convergence in Distribution and Tightness): Pick $X_n = (-1)^n$ and $X = 1$. Then X_n is tight. Pick $B = 10$ then:

$$P\{|X_n| > B\} = P\{1 > 10\} = 0$$

which holds for any $\epsilon > 0$. But $P\{X_n \leq x\} \not\to P\{X \leq x\}$. Interestingly though, there exists a subsequence, $n_j \in 2\mathbb{N}$, such that the sequence does converge in distribution, which is exactly Prokhorov's claim.

E.g.: Suppose $X_n \xrightarrow{p} 0$. Then X_n is tight. For a proof use the exact same argument in the theorem above.

Figure 15: Yuri Vasilyevich Prokhorov. Known associate of Andrey Nikolaevich Kolmogorov.

A STRONGER version of consistency for an estimator is τ_n consistency.

Def. (τ_n Consistent): Suppose you have a sequence, $\tau_n \to \infty$. If $\tau_n(\hat{\theta}_n - \theta)$ is tight then we say that $\hat{\theta}_n$ is a τ_n consistent estimator of θ.

E.g.: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P$ on $\mathbb{R}$ with $\sigma^2(P) < \infty$ then the CLT tells us that:

$$\sqrt{n}(\bar{X}_n - \mu(P)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(P))$$

which tells us that $\sqrt{n}(\bar{X}_n - \mu(P))$ is tight, which tells us that $\bar{X}_n$ is $\sqrt{n}$-consistent for $\mu(P)$.

Lem. (τ_n consistent $\implies$ consistent): *Proof.* Suppose $\hat{\theta}_n$ is a τ_n consistent estimator for θ. Then:

$$\tau_n(\hat{\theta}_n - \theta)$$

is tight. Which means that for every $\delta > 0$ and every $n \geq 1$ there exists $B > 0$ such that:

$$\begin{aligned} &P\{|\tau_n(\hat{\theta}_n - \theta)| \leq B\} \geq 1 - \delta \\ \implies &P\left\{|\hat{\theta}_n - \theta| \leq \frac{B}{|\tau_n|}\right\} \geq 1 - \delta \text{ because } |xy| = |x||y| \\ \implies &P\left\{|\hat{\theta}_n - \theta| > \frac{B}{|\tau_n|}\right\} < \delta \ \text{ because } P\{A\} = 1 - P\{A^c\}. \end{aligned}$$

Then fix some ϵ. Because B is a fixed number and $\tau_n \to \infty$ we know that for some N, all $n > N$ will have the following feature: $\epsilon > \frac{B}{|\tau_n|}$. Then:

$$\underbrace{\delta > P\left\{|\hat{\theta}_n - \theta| > \frac{B}{|\tau_n|}\right\}}_{\forall\, n} \underbrace{\geq P\left\{|\hat{\theta}_n - \theta| > \epsilon\right\}}_{\forall\, n > N}$$

which gives us consistency! □

Stochastic order notation

If $X_n \xrightarrow{p} 0$ then we can write: $X_n = o_P(1)$. If X_n is tight then we write: $X_n = O_P(1)$. More generally:

The non-random equivalents are: If $x_n \to 0$ then we may write $x_n = o(1)$ and if x_n is bounded then we may write $x_n = O(1)$.

$X_n = o_P(R_n)$ for $X_n = R_n Y_n$ with $Y_n = o_p(1)$.

$X_n = O_P(R_n)$ for $X_n = R_n Y_n$ with $Y_n = O_p(1)$.

Lem. (Calculus of Stochastic Order Notation): *Stochastic order notation has the following rules:*

1. $o_P(1) + o_P(1) = o_P(1)$
2. $o_P(1) + O_P(1) = O_P(1)$
3. $o_P(1)O_P(1) = o_P(1)$
4. $\frac{1}{1+o_P(1)} = O_P(1)$
5. $o_p(O_p(1)) = o_p(1)$

Proof. We'll prove $o_p(1)O_p(1) = o_p(1)$. Let $X_n = o_p(1)$ and $Y_n = O_p(1)$ and we want to show that $X_nY_n = o_p(1)$. That is, we want to show that:

$$\forall \epsilon > 0,\ P\{|X_nY_n| > \epsilon\} \to 0$$

We'll tackle this with a proof by contradiction. Suppose not. Then there exists $\epsilon > 0$ s.t.:

$$P\{|X_nY_n| > \epsilon\} \not\to 0$$

Then by Balzano-Weierstrass gives us that there exists n_j and $\delta > 0$ s.t.:

$$P\{|X_{n_j}Y_{n_j}| > \epsilon\} \to \delta$$

and then by Prokhorov's Theorem there exists n_{j_k} and Y s.t. $Y_{n_{j_k}} \xrightarrow{d} Y$. Then $X_{n_{j_k}} \xrightarrow{p} 0$ because if $X_n \xrightarrow{p} 0$ then any subsequence is doing the same. Then by Slutsky's Lem.:

$$X_{n_{j_k}}Y_{n_{j_k}} \xrightarrow{p} 0$$

but we originally supposed that:

$$P\{|X_{n_{j_k}}Y_{n_{j_k}}| > \epsilon\} \to \delta > 0$$

which establishes our contradiction. $\Rightarrow\Leftarrow$

#dumbledoresarmy

Large sample theory Core questions

E.g. (Summer 2012): Let (X_i, Y_i), $i = 1, \ldots, n$ be a sequence of independent bivariate random vectors such that X_i and Y_i are independent, $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ and $Y_i \sim \mathcal{N}(\mu, \sigma_i^2)$. Suppose further that there exists $\epsilon > 0$ and $B < \infty$ s.t.:

$$\epsilon \leq \sigma_i^2 \leq B \ \forall i.$$

In this question, you are asked to do inference on μ despite the fact that the number of nuisance parameters (i.e., the σ_i^2) is growing with the sample size n. To this end, consider the estimator of μ given by:

$$\hat{\mu}_n = \frac{1}{2n} \sum_{i=1}^{n} (X_i + Y_i).$$

(a) Is $\hat{\mu}_n$ an unbiased estimator of μ? Justify your answer.

Applying the expectations operator yields:

$$\begin{aligned} E[\hat{\mu}_n] &= E\left[\frac{1}{2n} \sum_{i=1}^{n} (X_i + Y_i)\right] \\ &= \frac{1}{2n}\left(\sum_{i=1}^{n} E[X_i] + \sum_{i=1}^{n} E[Y_i]\right) \\ &= \frac{1}{2n}(n\mu + n\mu) \\ &= \mu. \end{aligned}$$

Then by simply invoking the definition of unbiasedness we can conclude that $\hat{\mu}_n$ is unbiased.

(b) Compute $Var(\hat{\mu}_n)$. Is $\hat{\mu}_n$ a consistent estimator of μ? Justify your answer.

$$\begin{aligned} Var(\hat{\mu}_n) &= \frac{1}{4n^2} \sum_{i=1}^{n} Var(X_i + Y_i) \ \text{ b/c ind.} \\ &= \frac{1}{4n^2} \sum_{i=1}^{n} Var(X_i) + Var(Y_i) \ \text{ b/c ind.} \\ &= \frac{1}{4n^2} \sum_{i=1}^{n} 2\sigma_i^2 \\ &= \frac{1}{2n^2} \sum_{i=1}^{n} \sigma_i^2. \end{aligned}$$

But what about consistency?

$$\begin{aligned}
\hat{\mu}_n &= \frac{1}{2n}\sum_{i=1}^{n}(X_i + Y_i) \\
&= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&\xrightarrow{p} \frac{1}{2}(\mu + \mu) \text{ by W.L.L.N. and C.M.T.} \\
&= \mu
\end{aligned}$$

which establishes consistency.

(c) Is $\hat{\sigma}_i^2$ an unbiased estimator of σ_i^2? Justify your answer.

The proposed estimator is:

$$\hat{\sigma}_i^2 = \frac{1}{2}(Y_i - X_i)^2$$

and to see if it's unbiased or not we can simply apply the expectations operator to it:

$$\begin{aligned}
E[\hat{\sigma}_i^2] &= E\left[\frac{1}{2}(Y_i - X_i)^2\right] \\
&= \frac{1}{2}E[Y_i^2 - 2Y_iX_i + X_i^2] \\
&= \frac{1}{2}\left(E[Y_i^2] + E[X_i^2] - 2E[Y_iX_i]\right) \\
&= \frac{1}{2}\left(E[Y_i^2] + E[X_i^2] - 2E[Y_i]E[X_i]\right) \quad \text{b/c of independence} \\
&= \frac{1}{2}\left(E[Y_i^2] + E[X_i^2] - 2\mu^2\right) \\
&= \frac{1}{2}\left(\underbrace{E[Y_i^2] - \mu^2}_{\sigma_i^2} + \underbrace{E[X_i^2] - \mu^2}_{\sigma_i^2}\right) \\
&= \sigma_i^2
\end{aligned}$$

which establishes that our estimator is unbiased.

(d) Compute $Var(\hat{\sigma}_i^2)$. Is $\hat{\sigma}_i^2$ consistent? (Hint: If $Z \sim \mathcal{N}(0, \tau^2)$ then $E[Z^4] = 3\tau^4$.)

To compute the variance of $\hat{\sigma}_i^2$ we'll just apply the formula for variance:

$$\begin{aligned} Var(\hat{\sigma}_i^2) &= E\left[\left(\frac{1}{2}(Y_i - X_i)^2 - \sigma_i^2\right)^2\right] \\ &= E\left[\frac{1}{4}(Y_i - X_i)^4 - (Y_i - X_i)^2\sigma_i^2 + \sigma_i^4\right] \\ &= \frac{1}{4}E[(Y_i - X_i)^4] - \sigma_i^4 \end{aligned}$$

and we know that $E[Y_i - X_i] = 0$ and that $Var(Y_i - X_i) = 2\sigma_i^2$ because the two variables are independent. Then:

$$(Y_i - X_i) \sim \mathcal{N}(0, 2\sigma_i^2)$$

and the hint tells us that:

$$E[(Y_i - X_i)^4] = 3(2\sigma_i^2)^2$$

which we can apply to get:

$$Var(\hat{\sigma}_i^2) = 2\sigma_i^4.$$

Is the estimator consistent?

Conditional expectations

TRADITIONALLY conditional expectations are defined by with something like:

$$E[Y|X=x] = \int y f(y|x) dy$$

where $f(y|x)$ is the conditional density of Y. That's a useful definition but it presupposes the existence of a continuous conditional density function to integrate over.[57] Instead we'll work with an equivalent (but far more general) definition that's tightly connected to linear regressions.

Def. (Conditional Expectation of Y given X): Let Y, X be a random vector with $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$ and we'll assume $E[Y^2] < \infty$. Then define:

$$\mathbb{M} := \left\{ m(X) : \ m : \mathbb{R}^k \to \mathbb{R} \ \text{ and } E[m^2(X)] < \infty \right\}$$

and consider the following minimization problem:

$$\inf_{m(X) \in \mathbb{M}} E[(Y - m(X))^2]$$

then it's possible to show that there exists $m^*(X) \in \mathbb{M}$ s.t.:

$$E[(Y - m^*(X))^2] = \inf_{m(X) \in \mathbb{M}} E[(Y - m(X))^2]$$

and define:

$$E[Y|X] := m^*(X)$$

and we think of $E[Y|X]$ as the "best predictor" of Y given X.

Thm.: *$m^*(X)$ solves the minimization problem i.f.f. for all $m(X) \in \mathbb{M}$:*

$$E[(Y - m^*(X))m(X)] = 0$$

which we can think of as an orthogonality condition.[58] *And if $\tilde{m}(X) \in \mathbb{M}$ also solves the minimization problem, then:*

$$P\{m^*(X) = \tilde{m}(X)\} = 1$$

which is just to say that the they're the same solution in a probabilistic sense.

[57] Also, what a pain in the butt to work with! You aren't going to get closed-form solutions for some of even the best behaved p.d.f.'s.

[58] To see this, call $e := Y - m^*(X)$. Then we just require:

$$E[m(X)e] = 0$$

which is the analog of the familiar orthogonality restriction for OLS.

Proof. There are really two claims in this theorem. The first is an i.f.f. and the second is not. We'll tackle the i.f.f. part first.

Suppose the orthogonality condition holds for all $m(X) \in \mathbb{M}$. Then we want to show that the solution to the minimization problem is $m^*(X)$. Fix $m(X)$. Then starting with the minimization problem, we can add and subtract $m^*(X)$ to get:

$$\begin{aligned} E[(Y-m(X))^2] &= E[(Y-m^*(X)+m^*(X)-m(X))^2] \\ &= E[(Y-m^*(X))^2] + 2E[(Y-m^*(X))\underbrace{(m^*(X)-m(X))}_{\text{Call } \tilde{m}(X)\in\mathbb{M}}] + E[(m^*(X)-m(X))^2] \\ \diamond\diamond\diamond &= E[(Y-m^*(X))^2] + E[(m^*(X)-m(X))^2] \ \text{ b/c orthog. gives us } \ E[(Y-m^*(X))\tilde{m}(X)] = 0 \\ &\geq E[(Y-m^*(X))^2] \ \text{ b/c } (m^*(X)-m(X))^2 > 0 \end{aligned}$$

which holds for any $m(X)$, which establishes that $m^*(X)$ is the minimizing value of $m(X)$ for the minimization problem in the theorem.

Now we do the other direction of the i.f.f. claim. Suppose that $m^*(X)$ solves the minimization problem. We want to find[59] that the orthogonality condition holds. To do this we'll compare $m^*(X)$ with $\check{m}(X) := m^*(X) + \alpha m(X)$ for $\alpha \in \mathbb{R}$. Fix $m(X)$ in $\check{m}(X)$. Then because $m^*(X)$ is the minimizer we get that:

$$E[(Y-m^*(X))^2] \leq E[(Y-\check{m}(X))^2] = E[(Y-m^*(X)-\alpha m(X))^2] \ \text{ for any } \alpha$$

which we can expand and re-arrange to get:

$$2\alpha E[m(X)(Y-m^*(X))] \leq \alpha^2 E[m^2(X)] \ \forall \alpha$$

now recall that our goal is to show that the LHS term is equal to 0. Divide through by 2α and pick $\alpha = 0$. The choice of $\alpha = 0$ makes the inequality an equality.[60] Then we get:

$$E[m(X)(Y-m^*(X))] = 0$$

and then we've attained our goal because this holds for any $m(X)$.

Now we just need to deal with the $\tilde{m}$ part of the theorem. Suppose $\tilde{m}(X)$ solves the minimization problem. Then plug in $m = \tilde{m}$ into $\diamond\diamond\diamond$ above to get:

$$\begin{aligned} E[(\tilde{m}(X)-m^*(X))^2] &= E[(Y-\tilde{m}(X))^2] - E[(Y-m^*(X))^2] \\ &= 0 \ \text{ b/c both } m^*(X) \text{ and } \tilde{m}(X) \text{ minimize.} \end{aligned}$$

And $E[(\tilde{m}(X)-m^*(X))^2] = 0 \implies P\{m^*(X) = \tilde{m}(X)\} = 1.$ □

◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

A LESS restrictive definition of conditional expectations only requires $E[|Y|] < \infty$. In this case we define $E[Y|X]$ to be any $m^*(X)$ with $E[|m^*(X)|] < \infty$ s.t.:

$$E[(Y-m^*(X))I\{X \in \mathcal{B}\}] = 0 \ \forall \text{ (Borel) set } \mathcal{B}$$

[59] wtf

[60] Because:

$$E[(Y-m^*(X))^2] = E[(Y-\check{m}(X))^2]$$

when $\alpha = 0$.

Working off of just this definition we can derive many of the properties we know and love about conditional expectations:

Lem.: *If $Y = f(X)$ then $E[Y|X] = f(X)$.*

Proof. If $Y = f(X)$ then $E[Y|X]$ is any $m^*(X)$ s.t.:

$$E[(Y - m^*(X))I\{X \in \mathcal{B}\}] = 0$$

for any $\mathcal{B}$. Plugging in $Y = f(X)$ we clearly see that the $m^*(X)$ that solves the formula is $f(X)$:

$$E[(f(X) - m^*(X))I\{X \in \mathcal{B}\}] = 0$$

and if $m^*(X) = f(X)$ then we say that: $E[Y|X] = f(X)$. □

Lem.: $E[Z + Y|X] = E[Z|X] + E[Y|X]$.

Proof. For $E[Z + Y|X]$ we have that $m^*(X) = E[Z + Y|X]$ is the solution to:

$$E[(Z + Y - m^*(X))I\{X \in \mathcal{B}\}] = 0$$

for any set $\mathcal{B}$. Then we want to show that another solution is $E[Z|X]$ and $E[Y|X]$. To see why, we'll simply do some algebra:

$$\begin{aligned} E[(Z + Y - E[Z|X] - E[Y|X])I\{X \in \mathcal{B}\}] &= E[(Z - E[Z|X])I\{X \in \mathcal{B}\}] + E[(Y - E[Y|X])I\{X \in \mathcal{B}\}] \\ &= 0 \end{aligned}$$

which establishes that $m^*(X) = E[Z|X] + E[Y|X]$ is also a solution because of the properties for each of those conditional expectations.

□

Lem.: $E[f(X)Y|X] = f(X)E[Y|X]$.

Proof. We'll prove for the case when $f(X) = I\{X \in \tilde{\mathcal{B}}\}$ to get intuition. In this case:

$$E[(f(X)Y - f(X)E[Y|X])I\{X \in \mathcal{B}\}] = E[(Y - E[Y|X])I\{X \in \tilde{\mathcal{B}} \cap \mathcal{B}\}] = 0$$

□

Lem.: *If $P\{Y \geq 0\} = 1$ then $P\{E[Y|X] \geq 0\} = 1$.*

Lem. (Law of Iterated Expectations): *If $\mathcal{B} = \mathbb{R}^k$ we get:*

$$E[Y - E[Y|X]] = 0 \implies E[Y] = E[E[Y|X]].$$

More generally we can write:

$$E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$$

which is also sometimes called the Tower Property.

Lem.: *If Y is independent of X, then:*[61]

$$E[Y|X] = E[Y]$$

Proof.

$$E[(Y - E[Y])I\{X \in \mathcal{B}\}] = E[(Y - E[Y])]P\{X \in \mathcal{B}\} = 0$$

□

We can also extend several results we had for expectations for conditional expectations:

Thm. (Jensen's Inequality—Conditional Expectations): *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Suppose (Y, X) is a random vector on $\mathbb{R} \times \mathbb{R}^k$ such that $E[|Y|] < \infty$ and $E[|\phi(Y)|] < \infty$. Then with probability 1:*

$$E[\phi(Y)|X] \geq \phi(E[Y|X]).$$

Def. (Conditional Variances): Let (X, Y) be a random vector where X takes on values in $\mathbb{R}^k$ and Y takes on values in $\mathbb{R}$. Then is $E[Y^2] < \infty$ we define:[62]

$$Var(Y|X) = E[(Y - E[Y|X])^2|X] = E[Y^2|X] - E[Y|X]^2.$$

[61] This is sometimes called *mean independence*. Independence $\implies$ mean independence $\implies$ uncorrelated but the reverse direction is generally false.

[62] One useful result of this definition is:

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X]).$$

Linear Regression

LET (Y, X, U) be a random vector with: $Y \in \mathbb{R}$, $X \in \mathbb{R}^{k+1}$, and $U \in \mathbb{R}$. We'll assume:

$$X = (X_0, \ldots, X_k)' \text{ with } X_0 = 1$$

and:

$$\beta \in \mathbb{R}^{k+1},\ \beta = (\beta_0, \ldots, \beta_k)'$$

and:

$$Y = X'\beta + U.$$

Then there are three ways to think about interpreting:

1. *Linear Conditional Expectation*: Assume $E[Y|X] = X'\beta$ and define:

   $$U := Y - E[Y|X]$$

   so $Y = X'\beta + U$. Then:

   $$E[U|X] = 0 \implies E[U] = 0,\ E[XU] = 0.$$

   But there is no causal interpretation to β here.[63] That is, if there's a one unit increase in X_j there's nothing that says we should expect a β_j unit increase in Y. What's missing is a model of how Y is determined as a function of X.[64] In the Linear Conditional Expectations approach β is just a convenient way of summarizing $E[Y|X]$.

2. *Best Linear Prediction*[65] *of Y given X*: We don't have to be certain that $E[Y|X]$ is linear, but we can ask for a linear approximation. That is, a function of the form $X'b$ where b is some $b \in \mathbb{R}^{k+1}$ that is "close" to $E[Y|X]$. More precisely:

   $$\min_{b \in \mathbb{R}^{k+1}} E\left[(E[Y|X] - X'b)^2\right].$$

   Equivalently we can work with:

   $$\min_{b \in \mathbb{R}^{k+1}} E\left[(Y - X'b)^2\right].$$

[63] Or marginal effect interpretation here:
$$\frac{\partial E[Y|X]}{\partial X_j}.$$

[64] E.g., ice cream cones don't kill kids in swimming pools.

[65] The non-IO BLP.

Why? Well we'll work through a proof because the equivalence between these two minimization problems is one of the canonical examples of the mechanics of OLS. To do so we'll start with the first expression:

$$
\begin{aligned}
E\left[(E[Y|X]-X'b)^2\right] =& E\left[(\underbrace{E[Y|X]-Y}_{\text{V}}+Y-X'b)^2\right]\\
=& E[V^2]+2E[V(Y-X'b)]+E[(Y-X'b)^2]\\
=& E[V^2]+2E[VY]-2\underbrace{E[VX']}_{=0}b+E[(Y-X'b)^2]\\
=& A+E[(Y-X'b)^2]
\end{aligned}
$$

and if we take a derivative w/r/t b the constant, A, will drop out because it doesn't depend on b. Then we've proved that the two minimization problems are equivalent. Sweet.

We get that $E[VX']=0$ because:

$$
\begin{aligned}
E[VX'] &= E[E[Y|X]X'-YX']\\
&= E[E[YX'|X]]-E[YX']\\
&= E[YX']-E[YX'] \text{ by the law of it. exp.}\\
&= 0 \text{ by mathematics.}
\end{aligned}
$$

What does the minimization problem actually look like? Well, the F.O.C. is:

$$-2E[X(Y-X'b)]=0$$

and we can neglect the -2 and argue that any solution to the choice of b, which we'll call β, has to satisfy:

$$E[X(Y-X'b)]=0.$$

Then in the BLP context we can define the error term as a function of Y, X, and the solution to the minimization problem, β: $U := Y-X'\beta$. Then we can rearrange to get the familiar:

$$Y=X'\beta+U$$

and we can re-express the F.O.C. to get the familiar OLS moment restriction:

$$E[XU]=0.$$

However, just because we've managed to express the BLP problem in terms familiar to, say, estimating treatment effects from an experimental intervention, doesn't mean we have a causal interpretation here. U is just a residual that makes everything work. We need to impose actual structure on U in order to get the interpretation economists' desire.

3. *Causal Model*: Assume $Y=g(X,U)$ where X is observed determinants of Y and U is observed determinants of Y and $g(\cdot)$ is a model for how Y is determined.[66] Then:

$$\frac{\partial g(X,U)}{\partial X_j} \text{ is the effect of } X_j \text{ on } Y$$

[66] Hopefully $g(\cdot)$ is coming from economics or physics or logic or some other place we like.

and if we assume further that:

$$g(X, U) = X'\beta + U$$

then the partial derivative above is just β_j.

In general, we don't know much about U. For a given problem we generally ask: $E[U] = 0$? $E[XU] = 0$? $E[U|X] = 0$? These are just statements about the relationship between observed and unobserved determinants of Y. But even though we may have $E[U] \neq 0$, we can normalize so that it is. That is, replace β_0 with $\beta_0 + E[U]$ and U with $U - E[U]$.

Linear regression with exogeneity assumption

If there's some reason that know $E[XU] = 0$ then we're necessarily in the third interpretation of linear regression from above.[67] Now let's get down to brass tax and think about how to solve for β. We'll use the following setup as above: (Y, X, U) where $Y \in \mathbb{R}, U \in \mathbb{R}, X \in \mathbb{R}^k$, and:

$$Y = X'\beta + U$$

with $E[XU] = 0, E[XX'] < \infty$, and no perfect collinearity in X.[68] Why do we assume no perfect collinearity in X? The following lemma should make it clear why.

Lem.: *Assume $E[XX'] < \infty$ then $E[XX']$ is invertible i.f.f. there is no perfect collinearity in X.*

Proof. Easy to show the equivalent statement: $E[XX']$ is not invertible i.f.f. there is perfect collinearity in X.

First we'll start from perfect collinearity. Suppose there is perfect collinearity in X. Then $\exists\ c \neq 0$ s.t. $P\{X'c = 0\} = 1$. Then $E[XX']c = E[X(X'c)] = 0$. Then $E[XX']$ isn't invertible because it's equal to 0 and that's a quantity that's notoriously difficult to invert.

To take care of the other direction suppose that $E[XX']$ isn't invertible. Then there's some number $c \neq 0$ s.t. $E[XX']c = 0$. Then if we multiply through by c' we get:

$$c'E[XX']c = E[(X'c)^2] = 0$$

which only holds if $P\{X'c = 0\} = 1$. Then we've established that there's perfect collinearity X and we're all set with this proof. □

Plugging $U = Y - X'\beta$ into $E[XU] = 0$ we get:

$$E[XY] = E[XX']\beta \implies \beta = E[XX']^{-1}E[XY].$$

[67] That's because in Interpretations #1,2, $E[XU] = 0$ was just a result of a functional relationship that the setup imposed, whereas Interpretation #3 requires us to know something about the relationship between X and U that could give us a condition like $E[XU] = 0$ that we ex-ante assume.

[68] Don't know what those words mean? Here's some help:

Def. (Perfect Colinearity): We say there is perfect colinearity (or multicolinearity) in X if:

$$\exists\ c \neq 0, c \in \mathbb{R}^{k+1}\ s.t.\ P\{X'c = 0\} = 1.$$

I.e., we can express one component of X as a linear combination of the others.

If there is perfect collinearity in X, then there are generally multiple solutions to $E[XY] = E[XX']\beta$ but any two solutions, $\tilde{\beta}, \beta$ satisfy:

$$P\{X'\beta = X'\tilde{\beta}\} = 1$$

which doesn't tell us that $\beta = \tilde{\beta}$ but still worth noting.

Solving for sub-vectors of β is a really useful exercise, so we'll go over that now.[69] Write:

$$Y = X_1'\beta_1 + X_2'\beta_2 + U$$

and our expression for β becomes:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} E[X_1X_1'] & E[X_1X_2'] \\ E[X_2X_1'] & E[X_2X_2'] \end{pmatrix}^{-1} \begin{pmatrix} E[X_1Y] \\ E[X_2Y] \end{pmatrix}$$

and our goal is to get an expression for β_1. One approach would be to use the partition matrix inverse formula, but that's messy and not illuminating.[70] Instead we'll define the following notation for two random vectors, A, B:[71]

$$BLP(A|B) = \text{ Best Linear Predictior of } A \text{ given } B.$$

Then call:

$$\tilde{Y} = Y - BLP(Y|X_2) = Y - X_2'\gamma$$
$$\tilde{X}_1 = X_1 - BLP(X_1|X_2) = X_1 - X_2'\delta$$

and consider:

$$\tilde{Y} = \tilde{X}_1'\tilde{\beta}_1 + \tilde{U} \text{ with } E[\tilde{X}_1\tilde{U}] = 0 \text{ as in Interp. \#2}$$

and we claim:

$$\tilde{\beta}_1 = \beta_1.$$

Why? Well let's get our Q.E.D. on.

Proof. Starting with $\tilde{\beta}_1$ we get:

$$\begin{aligned}
\tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1\tilde{Y}] \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}\left(E[\tilde{X}_1Y] - E[\tilde{X}_1BLP(Y|X_2)]\right) \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1Y] \text{ b/c } E[\tilde{X}_1BLP(Y|X_2)] = E[\tilde{X}_1X_2'\gamma] = E[(X_1 - X_2'\delta)X_2']\gamma = E[(X_2U)']\gamma = 0 \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1(X_1'\beta_1 + X_2'\beta_2 + U)] \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1X_1']\beta_1 \text{ b/c } U \text{ is orthogonal to both } X_1, X_2 \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1(\tilde{X}_1 + BLP(X_1|X_2))']\beta_1 \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}\left(E[\tilde{X}_1\tilde{X}_1]\beta_1 + E[\tilde{X}_1BLP(X_1|X_2))']\right)\beta_1 \\
&= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1\tilde{X}_1]\beta_1 \text{ b/c of the same orthogonality condition used above} \\
&= \beta_1.
\end{aligned}$$

which was our claim.[72] □

[69] It's worth noting that nothing is based on sample data right now. Later we'll look at the sample analog of the solution to β and of the solution for sub-vectors of β.

[70] Thank god!

[71] $BLP(A|B)$ is $B'\delta$ where δ is from the minimization of:

$$E[(A - B'\delta)^2]$$

where for $W = A - B'\delta$ the F.O.C. gives us $E[BW] = 0$ for free.

[72] It's worth pointing out that we're basically just using the same orthogonality condition over and over and over again here.

And what's so cool about this result is that it gives meaning to the phrase, β_1 is the effect of X_1 on Y after controlling for X_2.[73,74] Can also apply result with $X_2 = 1$ with $\tilde{X}_1 = X_1 - E[X_1]$ and $\tilde{Y} = Y - E[Y]$ would give us:

$$\beta_1 = E[(X_1 - E[X_1])(X_1 - E[X_1])']^{-1}E[(X_1 - E[X_1])(Y - E[Y])] = Var(X_1)^{-1}Cov(X_1, Y)$$

which is super cool.

Omitted variable bias

SUPPOSE:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

with $E[X_1 U] = E[X_2 U] = E[U] = 0$ (all scalars) and consider:

$$Y^* = \beta_0^* + \beta_1^* X_1 + U^*$$

with the same moment restrictions. In general $\beta_1 \neq \beta_1^*$. Why?

$$\begin{aligned}\beta_1^* &= \frac{Cov(X_1, Y)}{Var(X_1)}\\ &= \frac{Cov(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U)}{Var(X_1)}\\ &= \beta_1 + \underbrace{\beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)}}_{\text{The Bias}}\end{aligned}$$

where The Bias can be pretty substantial if X_1 and X_2 covary. For example, if you're looking at drownings and ice-cream sales and omit weather, the covariance of weather and ice-cream sales is going to be substantial and really mislead your interpretation of β_1^*. Same thing is going on with police and crime in the figure.

If we do the same exercise with vectors then we start with:

$$Y = \beta_0 + X_1'\beta_1 + X_2'\beta_2 + U$$

and consider:

$$Y = \beta_0^* + X_1'\beta_1^* + U^*$$

then:

$$\begin{aligned}\beta_1^* &= Var(X_1)^{-1}Cov(X_1, Y)\\ &= \beta_1 + Var(X_1)^{-1}Cov(X_1, X_2)\beta_2\end{aligned}$$

which is a less straight-forward expression. Also, it's generally what we're thinking about in a regression and can differ from the scalar intuition above.

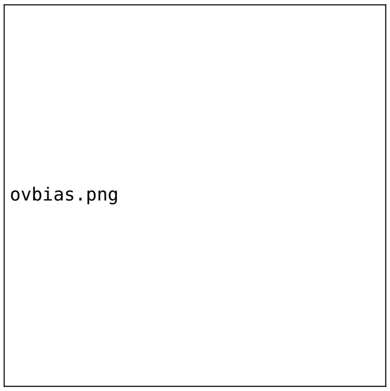

Figure 16: Figure 1 from Levitt (1997). Are police causing more crime to occur?

[73] Residual regression, dawg.

[74] Could have used Y instead of $\tilde{Y}$ and we'd get the same result.

Measurement error

CONSIDER the follow scalar model:

$$Y = \beta_0 + \beta_1 X_1 + U$$

with the standard moment restrictions. Then instead of observing X_1 we actually observe $\hat{X}_1 = X_1 + V$ with $E[V] = 0,\ Cov(X_1, V) = 0$, and $Cov(U, V) = 0$.[75] Then how does β_1^* compare to β_1?

$$Y = \beta_0^* + \beta_1^* \hat{X}_1 + U^*$$

with it's own moment restrictions. Then:

$$\begin{aligned}
\beta_1^* &= \frac{Cov(\hat{X}_1, Y)}{Var(\hat{X}_1)} \\
&= \beta_1 \frac{Cov(\hat{X}_1, X_1)}{Var(\hat{X}_1)} + \frac{Cov(\hat{X}_1, U)}{Var(\hat{X}_1)} \\
&= \beta_1 \frac{Cov(\hat{X}_1, X_1)}{Var(\hat{X}_1)} \quad \text{b/c } Cov(\hat{X}_1, U) = Cov(X_1, U) + Cov(V, U) = 0 \\
&= \beta_1 \underbrace{\frac{Var(X_1)}{Var(X_1) + Var(V)}}_{\text{Signal-to-noise ratio}}
\end{aligned}$$

where the signal-to-noise ratio will be less than 1 and attenuate the observed value of β_1, β_1^*.

More generally if we move out of the scalar case with:

$$Y = \beta_0 + X_1' \beta_1 + U$$

and:

$$Y = \beta_0^* + \hat{X}_1' \beta_1^* + U^*$$

where:

$$\hat{X}_1 = X_1 + V.$$

Then:

$$\beta_1^* = (\underbrace{Var(X_1) + Var(V)}_{\text{Nightmare!}})^{-1} Var(X_1) \beta_1.$$

But we aren't always screwed if we're in search of intuition because we can use our residual regression approach from above. I.e., think of:

$$Y = \beta_0 + \beta_1 X_1 + X_2' \beta_2 + U$$

with the usual assumptions but with $\hat{X}_1 = X_1 + V$ and $E[V] = 0,\ Cov(X_1, V) = 0,\ Cov(U, V) = 0,$ and $Cov(X_2, V) = 0$ and we run the following regressions as in Interp. #2:

$$Y = \beta_0^* + \beta_1^* \hat{X}_1 + X_2' \beta_2^* + U^*$$

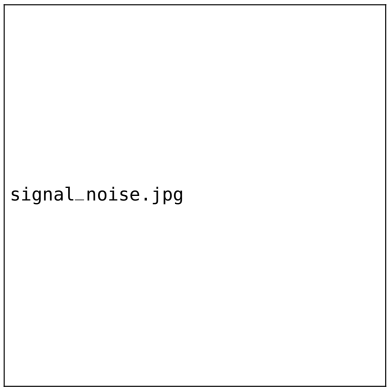

Figure 17: Nate Silver possibly misinterpreting what a signal-to-noise ratio musses up.

[75] The classical error-in variables model. Classic!

and define:

$$\begin{aligned}\tilde{\hat{X}}_1 &= \hat{X}_1 - BLP(\hat{X}_1|X_2)\\ &= X_1 + V + BLP(X_1|X_2)\\ &= \tilde{X}_1 + V \ \text{ where } \tilde{X}_1 = X_1 - BLP(X_1|X_2)\end{aligned}$$

and we can then think of our estimator of interest as:[76]

$$\beta_1^* = \frac{Cov(\tilde{\hat{X}}_1, Y)}{Var(\tilde{\hat{X}}_1)} = \beta_1 \underbrace{\frac{Var(\tilde{X}_1)}{Var(\tilde{X}_1) + Var(V)}}_{\text{Signal-to-noise ratio again}} .$$

This is the classic result of attenuation bias for β_1 in the presence of measurement error.[77]

Estimating β

Suppose (Y, X, U) satisfies:

$$Y = X'\beta + U$$

with:

$$E[XU] = 0 \text{ and } E[XX'] < \infty$$

with no perfect collinearity. This implies that:

$$\beta = E[XX']^{-1}E[XY].$$

Suppose we also have $(Y_1, X_1), \ldots, (Y_n, X_n)$, an i.i.d. sample from distribution of (Y, X). Then an analog estimator for β would be:

$$\hat{\beta}_n := \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

which we call the ordinary least squares (OLS) estimator.[78] Some popular terminology for OLS is:

$\hat{Y}_i := X_i'\hat{\beta}_n =: i^{th}$ fitted value.

$\hat{U}_i := Y_i - \hat{Y}_i = Y_i - X_i'\hat{\beta}_n =: i^{th}$ residual.

And two useful properties of these values are:

1. We can move between fitted values and the observed outcome with the following relationship:

$$Y_i = \hat{Y}_i + \hat{U}_i.$$

2. The sum of the product of the residuals and covariates is 0:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \hat{U}_i = 0.$$

[76] Where:

$$Var(\tilde{\hat{X}}_1) = Var(\tilde{X}_1 + V) = Var(\tilde{X}_1) + Var(V)$$

because we've assumed they're uncorrelated and:

$$\begin{aligned}Cov(\tilde{\hat{X}}_1, Y) &= Cov(\tilde{X}_1, X_1)\beta_1\\ &= Cov(\tilde{X}_1, \tilde{X}_1 + BLP(X_1|X_2))\beta_1\\ &= Var(\tilde{X}_1)\beta_1.\end{aligned}$$

[77] Bias on β_2 is hard to analyze. Typically people will say that everything gets attenuated, but that's not neccesarily true.

[78] Why do we call it that? Because $\hat{\beta}_n$ minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2$$

and the first order condition of this minimization problem gives us a $\hat{\beta}_n$ that satisfies:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}_n) = 0$$

which we can write as:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

where the LHS will be invertible with high probability for large n.

Projection interpretation of OLS

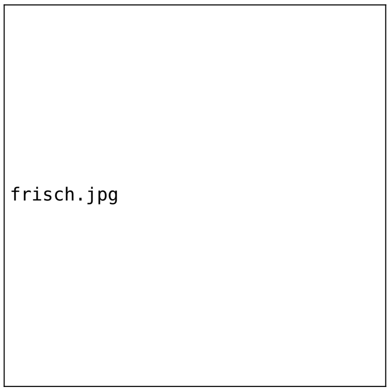

Figure 18: Ragnar Frisch of Frisch-Waugh-Lowell and Frisch labor supply elasticity fame.

FOR $(Y_1, X_1), \ldots, (Y_n, X_n)$, an i.i.d. sample from distribution of (Y, X) with $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$, define:

$$\begin{aligned}
\mathbb{Y} &= (Y_1, \ldots, Y_n)' \\
\mathbb{X} &= (X_1, \ldots, X_n)' \\
\mathbb{U} &= (U_1, \ldots, U_n)' \\
\hat{\mathbb{Y}} &= (\hat{Y}_1, \ldots, \hat{Y}_n)' = \mathbb{X}\hat{\beta}_n \\
\hat{\mathbb{U}} &= (\hat{U}_1, \ldots, \hat{U}_n)' = \mathbb{Y} - \hat{\mathbb{Y}}
\end{aligned}$$

Then in this notation:

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

and $\hat{\beta}_n$ solves:

$$\min_b |\mathbb{Y} - \mathbb{X}b|^2$$

where $\mathbb{X}b$ is a vector in $col(\mathbb{X})$.[79] One nice interpretation of the solution to this representation is that $\mathbb{X}\hat{\beta}_n$ is the vector in the column space of $\mathbb{X}$ that's orthogonal to the project of $\mathbb{Y}$ onto the column space of $\mathbb{X}$. That is, $\mathbb{X}\hat{\beta}_n$ is the vector in the column space that's closest to $\mathbb{Y}$. We can think of this mathematically by:

$$\mathbb{X}\hat{\beta}_n = \underbrace{\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}}_{:=\mathbb{P}}\mathbb{Y} = \mathbb{P}\mathbb{Y}$$

where $\mathbb{P}$ is the matrix that projects an n dimensional vector onto the space $col(\mathbb{X})$. The projection matrix is nicely well-behaved[80] and we can use that behavior to develop some tools to think about the sample/projection equivalent of solving for sub-vectors of β. Call:

$$\mathbb{M} := \mathbb{I} - \mathbb{P}$$

the "residual maker" matrix which is also a projection matrix and is orthogonal to $col(\mathbb{X})$. Then with just a little bit of algebra we get:[81]

$$\mathbb{M}\mathbb{Y} = \hat{\mathbb{U}}.$$

THEN to estimate subvectors of β we try to estimate:

$$Y = X_1'\beta_1 + X_2'\beta_2 + U$$

where:

$$(Y_1, X_{1,1}, X_{2,1}), \ldots, (Y_n, X_{1,n}, \ldots, X_{2,n})$$

[79] I.e., all vectors $\mathbb{X}b$.

[80] That is, $\mathbb{P}$ is symmetric and $\mathbb{P}^2 = \mathbb{P}$.

[81] If you're curious, the algebra is just:

$$\mathbb{M}\mathbb{Y} = \mathbb{Y} - \mathbb{P}\mathbb{Y} = \mathbb{Y} - \mathbb{X}\hat{\beta}_n.$$

are i.i.d. and we want to derive an expression for $\hat{\beta}_{1,n}$ by itself. Then define:

$$\mathbb{P}_1 = \text{ projection matrix onto } col(\mathbb{X}_1)$$

giving us: $\mathbb{M}_1 = \mathbb{I} - \mathbb{P}_1$. Define $\mathbb{M}_2, \mathbb{P}_2$ equivalently. Then if we note that:

$$\mathbb{Y} = \mathbb{X}_1'\hat{\beta}_{1,n} + \mathbb{X}_2'\hat{\beta}_{2,n} + \hat{\mathbb{U}}$$

we can pre-multiply by $\mathbb{M}_2$ to get:

$$\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \underbrace{\mathbb{M}_2\mathbb{X}_2\hat{\beta}_{2,n}}_{=0} + \underbrace{\mathbb{M}_2\hat{\mathbb{U}}}_{=\hat{\mathbb{U}}}$$

and if we pre-multiply again by $(\mathbb{M}_2\mathbb{X}_1)'$ then we get:

$$(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{Y} = (\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \underbrace{(\mathbb{M}_2\mathbb{X}_1)'\hat{\mathbb{U}}}_{=\mathbb{X}_1'\hat{\mathbb{U}}=0}$$

and $(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{X}_1$ is invertible if $\mathbb{X}'\mathbb{X}$ is invertible. Then we get the following nice expression for $\hat{\beta}_{1,n}$:

$$\hat{\beta}_{1,n} = \left[(\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{X}_1)\right]^{-1}\left[(\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{Y})\right]$$

which is the Frisch-Waugh-Lowell estimator and we can think of $(\mathbb{M}_2\mathbb{X}_1)'$ as the residuals for a regression of X_1 onto X_2.

Measures of fit

A POPULAR number to report in the estimation of a regression is R^2, which is merely a measure of how well the model you estimate fits the data. We write:

$$R^2 := \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where:[82]

$$ESS := \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2$$
$$TSS := \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$
$$SSR := \sum_{i=1}^{n}\hat{U}_i^2.$$

Lem.: *$TSS = ESS + SSR$.*

[82] ESS stands for explained sum of squares. TSS stands for total sum of squares and SSR is an acronym for sum of squared residuals. As you can see in the definitions, they're all exactly what they sound like they'd be.

Proof. To prove this, we'll just bust out some algebra:

$$\begin{aligned}
TSS &= \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2 \\
&= \sum_{i=1}^{n}(\hat{U}_i + \hat{Y}_i - \bar{Y}_n)^2 \\
&= \sum_{i=1}^{n}\hat{U}_i^2 + 2\sum_{i=1}^{n}\hat{U}_i(\hat{Y}_i - \bar{Y}_n) + \sum_{i=1}^{\mathcal{N}}(\hat{Y}_i - \bar{Y}_n)^2 \\
&= SSR + ESS + 2\sum_{i=1}^{n}\hat{U}_i(\hat{Y}_i - \bar{Y}_n) \\
&= SSR + ESS + 2\left(\sum_{i=1}^{n}\hat{U}_i\hat{Y}_i - \sum_{i=1}^{n}\hat{U}_i\bar{Y}_n\right) \\
&= SSR + ESS + 2\underbrace{\left(\sum_{i=1}^{n}\hat{U}_iX_i'\right)}_{=0}\hat{\beta}_n - 2\left(\sum_{i=1}^{n}\hat{U}_i\right)\bar{Y}_n \\
&= SSR + ESS - 2\bar{Y}_n\sum_{i=1}^{n}\hat{U}_i \\
&= SSR + ESS \;\; \text{b/c} \;\; \sum_{i=1}^{n}\hat{U}_i = 0 \;\; \text{if constant in } X.
\end{aligned}$$

□

This also establishes that:

1. $0 \leq R^2 \leq 1$
2. $R^2 = 1 \implies \hat{U}_i = 0 \; \forall i$
3. $R^2 = 0 \implies \hat{Y}_i = \bar{Y}_n \; \forall i$

Also R^2 is monotone in regressions.[83] Also, we can represent R^2 as:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\frac{1}{n}\sum_i^n \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{\mathcal{N}}(Y_i - \bar{Y}_n)^2}$$

which we can view as the sample analog of:

$$1 - \frac{Var(U)}{Var(Y)}.$$

The sense in which R^2 is monotone in regressors suggests a degree of freedom adjustment:

$$\bar{R}^2 := 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

[83] It always goes up if you add a regressor. *Azeem suggests we try to prove this.*

which establishes that $\bar{R}^2 \leq R^2 \leq 1$ but we could also get $\bar{R}^2 < 0$. Also $\bar{R}^2$ is not monotone in regressors. We call $\bar{R}^2$ the adjusted R^2. Depending on what you're going for, R^2 or $\bar{R}^2$ isn't that useful. A low or high R^2 doesn't tell us anything about the validity of a causal interpretation (Interp. 3).

Properties of OLS estimator

We'll now go back to the scalar setup with the following assumptions:

$$Y = X'\beta + U, \; E[XU] = 0, \; E[XX'] < \infty$$

and no perfect colinearity in X with $(Y_1, X_1), \ldots, (Y_n, X_n)$ is an i.i.d. sample from distribution of (Y, X).

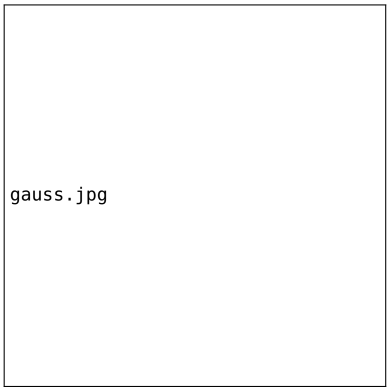

Figure 19: I held back as long as possible. Without further adieu, ladies and gentleman, the main attraction: Carl Friedrich Gauss! ::crowd noises::

Bias

A desirable property of an estimator is unbiasedness.[84] If we assume further that $E[U|X] = 0$ then $E[\hat{\beta}_n] = \beta$. Why?

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{U}$$

then if we take expectations:

$$E[\hat{\beta}_n|X_1, \ldots, X_n] = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' \underbrace{E[\mathbb{U}|X_1, \ldots, X_n]}_{=0}$$

giving us unbiasedness. But unbiasedness doesn't come for free. We need the conditional expectation of the error term to be 0.[85]

Efficiency

Another desirable property of an estimator would be that it's the most efficient. The following theorem helps formalize that thinking for OLS.

Thm. (Gauss-Markov Theorem): *This is another finite sample property. If we assume $E[U|X] = 0$ and homoskedasticity*[86] *and we restrict attention to estimators of β of the form $\mathbb{A}'\mathbb{Y}$ for:*

$$\mathbb{A} := \mathbb{A}(X_1, \ldots, X_n)$$

the class of conditionally linear functions of Y on $X_1, \ldots, X_n$, and:

$$E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \beta$$

which are just two properties we might demand in an estimator.[87] *Then*

[84] A small sample property.

[85] $E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \mathbb{A}'\mathbb{X}\beta + \mathbb{A}'E[\mathbb{U}|X_1, \ldots, X_n] = \mathbb{A}'\mathbb{X}\beta$. The unbiasedness condition, then, is equivalent to requiring:

$$\mathbb{A}'\mathbb{X} = \mathbb{I}.$$

[86] I.e., $Var(U|X) = \sigma^2$.

[87] Clearly OLS satisfies these conditions.

among this class, the "best" estimator is OLS. By best we mean that:

$$Var(\mathbb{A}'\mathbb{Y}|X_1,\dots,X_n)$$

is minimized. I.e., partial order given by $B \leq \tilde{B}$ if $\tilde{B} - B$ is positive semi-definite.[88]

Proof. OLS choice of $\mathbb{A}$ gives us:

$$Var(\mathbb{A}'\mathbb{Y}|X_1,\dots,X_n) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$$

by plugging in the OLS $\mathbb{A}$. Then we need to show for any $\mathbb{A}$ s.t. $\mathbb{A}'\mathbb{X} = \mathbb{I}$ that:

$$\mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1}$$

is positive semi-definite. Define:

$$\mathbb{C} := \mathbb{A} - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}$$

and then:

$$\begin{aligned} \mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1} &= [\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}]'[\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}] - (\mathbb{X}'\mathbb{X})^{-1} \\ &= [\mathbb{C}' + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'][\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}] - (\mathbb{X}'\mathbb{X})^{-1} \\ &= \mathbb{C}'\mathbb{C} + \mathbb{C}\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{C} \\ &= \mathbb{C}'\mathbb{C} + \mathbb{C}' + \mathbb{X}'\mathbb{C} \end{aligned}$$

where the last step establishes positive semi-definiteness. Note: We're using that implication of unbiasedness in these steps. We might not be obsessed with unbiasedness. Still, it's small sample, dawg. Deal with it. □

Consistency

Our OLS moment restriction, $E[XU] = 0$, tells us that $E[XY]$ exists because:

$$E[XY] = E[XX']\beta + E[XU]$$

and those subcomponents all exist. Then:

$$\hat{\beta}_n = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i Xi'\right)^{-1}}_{\xrightarrow{p} E[XX']^{-1}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)}_{\xrightarrow{p} E[XY]}$$

which we can apply CMT1 to in order to get:

$$\hat{\beta}_n \xrightarrow{p} E[XX']^{-1}E[XY] = \beta$$

which gives us consistency.

[88] $Var(\mathbb{A}'\mathbb{Y}|X_1,\dots,X_n) = Var(\mathbb{A}'\mathbb{U}|X_1,\dots,X_n) = \mathbb{A}'Var(\mathbb{U}|X_1,\dots,X_n)\mathbb{A} = \sigma^2\mathbb{A}'\mathbb{A}$ where the last step follows from homoskedasticity assumption.

Limiting distribution of $\hat{\beta}_n$

If we assume further that $Var(XU)$ exists, then:

$$Var(XU) = E[(XU - E[XU])(XU - E[XU])'] = E[XX'U^2]$$

and then:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

where:

$$\Omega = E[XX']^{-1} Var(XU) E[XX']^{-1}.$$

Why? Let's bust out our proof shoes:

Proof. Just working with $\hat{\beta}_n$ we get:

$$\begin{aligned} \hat{\beta}_n &= \left(\frac{1}{n} \sum_{i=1}^{n} X_i Xi' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right) \\ &= \beta + \left(\frac{1}{n} \sum_{i=1}^{n} X_i Xi' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i U_i \right) \end{aligned}$$

and then:

$$\sqrt{n}(\hat{\beta}_n - \beta) = \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} X_i Xi' \right)^{-1}}_{\xrightarrow{p} E[XX']^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i U_i \right)}_{\xrightarrow{d} \mathcal{N}(0, Var(XU))}$$

which we can apply Slutsky's Lem. to in order to get:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} E[XX']^{-1} \mathcal{N}(0, Var(XU)) = \mathcal{N}(0, \Omega).$$

□

Consistent estimation of Ω

Here we'll work with the same assumptions that we had for deriving the limiting distribution. First we'll assume homoskedasticity.[89] Then:

$$Var(XU) = E[XX'U^2] = E[XX'E[U^2|X]] = \sigma^2 E[XX']$$

allows us to write:

$$\Omega = E[XX']^{-1} \sigma^2$$

which we can estimate with:

$$\hat{\Omega} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_i^2 \right) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \hat{\sigma}_n^2.$$

[89] I.e., $E[U|X] = 0$ and $Var(U|X) = \sigma^2$.

But is this bad-boy consistent? Let's look at it piece-by-piece:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 = \underbrace{\frac{1}{n}\sum_{i=1}^n U_i^2}_{\xrightarrow{p} E[U^2]} + \underbrace{\frac{1}{n}\sum_{i=1}^n (\hat{U}_i^2 - U_i^2)}_{\text{Hmm}}$$

where $E[U^2] = E[U^2|X] = \sigma^2$. What about that Hmmm term, though? Well note:

$$\hat{U}_i = Y_i - X_i'\hat{\beta}_n = Y_i - X_i'\beta - X_i'(\hat{\beta}_n - \beta) = U_i - X_i'(\hat{\beta}_n - \beta)$$

which tells us that:

$$\hat{U}_i^2 - U_i^2 = -2U_iX_i'(\hat{\beta}_n - \beta) + (X_i'(\hat{\beta}_n - \beta))^2$$

so our Hmmmm term then becomes:

$$\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 = \underbrace{-2\frac{1}{n}\sum_{i=1}^n U_iX_i'(\hat{\beta}_n - \beta)}_{\text{A}} + \underbrace{\frac{1}{n}\sum_{i=1}^{\mathcal{N}} (X_i'(\hat{\beta}_n - \beta))^2}_{\text{B}}$$

and:

$$\text{A} = -2\left[\frac{1}{n}\sum_{i=1}^n U_iX_i'\right](\hat{\beta}_n - \beta) = -2\left[\frac{1}{n}\sum_{i=1}^n U_iX_i'\right]o_p(1)$$

and:

$$\text{B} \leq \frac{1}{n}\sum_{i=1}^n |X_i'(\hat{\beta}_n - \beta)|^2 \leq \frac{1}{n}\sum_{i=1}^n |X_i|^2|\hat{\beta}_n - \beta|^2 \xrightarrow{p} 0$$

and then we've established that our estimator is consistent because all of Hmmmmmm is going to 0 in probability.

Homoskedasticity is more or less a dumb assumption, though. We'll scrap that now and do some real work to show consistency. Now Ω becomes:

$$\Omega = E[XX']^{-1}Var(XU)E[XX']^{-1}$$

and to estimate $Var(XU) = E[XX'U^2]$ we'll use:

$$\frac{1}{n}\sum_{i=1}^n X_iX_i'\hat{U}_i^2 = \underbrace{\frac{1}{n}\sum_{i=1}^n X_iX_i'U_i^2}_{\xrightarrow{p} E[XX'U^2]} + \underbrace{\frac{1}{n}\sum_{i=1}^{\mathcal{N}} (\hat{U}_i^2 - U_i^2)}_{\text{Ugh}}$$

and look at the $(j,l)^{th}$ element of Ugh. I.e.,

$$|\frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,l}(\hat{U}_i^2 - U_i^2)| \leq \frac{1}{n}\sum_{i=1}^n |X_{i,j}X_{i,l}|\ |\hat{U}_i^2 - U_i^2| \leq \underbrace{\frac{1}{n}\sum_{i=1}^n |X_{i,j}X_{i,l}|}_{\xrightarrow{p} E[|X_jX_l|]} \underbrace{\max_{1\leq m\leq n} |\hat{U}_m^2 - U_m^2|}_{o_p(1)?}$$

where we know that $E[|X_j X_l|]$ is finite by a component-by-component application of $E[XX']'s$ finiteness. But we still need to show that:

$$\max_{1\leq m\leq n} |\hat{U}_m^2 - U_m^2| \xrightarrow{p} 0.$$

We'll start with a Lemma that'll help us deal with that claim.

Lem.: *Let $Z_1, \dots, Z_n$ be an i.i.d. sequence of random variables s.t. $E[|Z_i|^r] < \infty$ then:*

$$n^{-1/r} \max_{1\leq i\leq n} |Z_i| \xrightarrow{p} 0$$

or, equivalently:

$$\max_{1\leq i\leq n} |Z_i| = o_p\left(n^{1/r}\right).$$

Proof. Let $\epsilon > 0$ be given. Then note:

$$\begin{aligned}
P\left\{n^{-1/r}\max_{1\leq i\leq n}|Z_i| > \epsilon\right\} &= P\left\{\max_{1\leq i\leq n}|Z_i| > \epsilon n^{1/r}\right\} \\
&= P\left\{\cup_{1\leq i\leq n}\{|Z_i|^r > \epsilon^r n\}\right\} \\
&\leq \sum_{1\leq i\leq n} P\{|Z_i|^r > \epsilon^r n\} \text{ by Boole/Bonferonni} \\
&\leq \sum_{i=1}^n \frac{1}{\epsilon^r n} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \text{ by Markov's Ineq.+} \\
&= \frac{1}{\epsilon^r} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \\
&\to 0 \text{ by Dom. Conv. Thm. and assumption that } E[|Z_i|^r] < \infty.
\end{aligned}$$

□

Thm. (Markov's Inequality+): *For any random variable W:*

$$P\{|W| > \epsilon\} \leq \frac{1}{\epsilon} E[|W| I\{|W| > \epsilon\}].$$

Proof. If $I\{|W| > \epsilon\} \leq \frac{|W|}{\epsilon} I\{|W| > \epsilon\}$ then I can take expectations and get the result of Markov's Inequality+. □

So returning to our Ughhh term, we can apply the Lem. to get that:

$$\max_{1\leq m\leq n} |\hat{U}_m^2 - U_m^2| \leq 2\underbrace{\sqrt{n}|\hat{\beta}_n - \beta|}_{=O_p(1)} \underbrace{n^{-1/2}\max_{1\leq m\leq n}|U_m||X_m|}_{=o_p(1)} + \underbrace{n|\hat{\beta}_n - \beta|^2}_{=O_p(1)} \underbrace{\frac{1}{n}\max_{1\leq m\leq n}|X_m|^2}_{=o_p(1)}$$

because $|\hat{U}_m^2 - U_m^2| \leq 2|U_m||X_m||\hat{\beta}_n - \beta| + |X_m|^2|\hat{\beta}_n - \beta|^2$.[90] Also, worth noting that the first $o_p(1)$ term follows if $E[|UX|^2] < \infty$ which follows from $E[XX'U^2] < \infty$ and the second $o_p(1)$ term follows if $E[|X|^2] < \infty$ which follows from $E[XX'] < \infty$.

Then we've derived our desired result and proven that our estimator for Ω, $\hat{\Omega}_n$ is consistent. Sweeet.

Inference

SUPPOSE we're interested in testing a single linear restriction:

$$H_0 : r'\beta = c \text{ vs. } H_A : r'\beta \neq c$$

[90] This follow from the observation that: $\hat{U}_m^2 = U_m^2 - 2U_{i,m}X'_{i,m}(\hat{\beta}_n - \beta) + (X_{i,m}(\hat{\beta}_n - \beta))^2$.

where r is a $(k+1)$ by 1 vector that's not equal to 0. Then we know that:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

and that $\hat{\Omega}_n \xrightarrow{p} \Omega$, so:

$$r'[\sqrt{n}(\hat{\beta}_n - \beta)] = \sqrt{n}(r'\hat{\beta}_n - r'\beta) \xrightarrow{d} r'\mathcal{N}(0, \Omega) = \mathcal{N}(0, r'\Omega r)$$

and that:

$$r'\hat{\Omega}_n r \xrightarrow{p} r'\Omega r > 0$$

which all implies that:

$$\frac{\sqrt{n}(r'\hat{\beta}_n - r'\beta)}{\sqrt{r'\hat{\Omega}_n r}} \xrightarrow{d} \mathcal{N}(0,1)$$

so a natural test statistic is $|T_n|$ where:

$$T_n = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\Omega}_n r}}$$

with critical value:

$$z_{1-\alpha/2} = \Phi(1 - \alpha/2).$$

Then when the null is true:

$$P\{|T_n| > z_{1-\alpha/2}\} = 1 - P\{-z_{1-\alpha/2} \leq T_n \leq z_{1-\alpha/2}\} = 1 - P\{-z_{\alpha/2} \leq T_n \leq z_{1-\alpha/2}\} \to 1 - (1-\alpha) = \alpha$$

where we can swap between $-z_{1-\alpha/2} = z_{\alpha/2}$ because the normal distribution is symmetric. So we've got a test that's consistent in level, but how should we think about p-values? Well we reject at level α if:[91]

$$|T_n| > z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$$

Obvious modifications for hypothesis tests with inequalities. Confidence intervals for β_j? Set $r = e_j = (0, \ldots, 1, \ldots, 0)'$ where there's a 1 in the j^{th} slot. Then just use duality between confidence intervals and hypothesis testing.

TESTING multiple linear restriction requires some more machinery. Suppose we want to test:

$$H_0 : R\beta = c \text{ vs. } H_A : R\beta \neq c$$

where R is a $p \times k+1$ matrix. Assume rows of R are linearly independent. Then:

$$R[\sqrt{n}(\hat{\beta}_n - \beta)] = [\sqrt{n}(R\hat{\beta}_n - R\beta)] \xrightarrow{d} R\mathcal{N}(0, \Omega) = \mathcal{N}(0, R\Omega R').$$

And $R\Omega R'$ is nonsingular. Why? For $\alpha \neq 0$, $\alpha' R\Omega R'\alpha > 0$ because $\alpha' R \neq 0$ by assumption. Then $R\hat{\Omega}_n R' \xrightarrow{p} R\Omega R'$, so, then by the CMT:

$$\mathcal{N}(R\hat{\beta}_n - R\beta)'(R\hat{\Omega}_n R')^{-1}(R\hat{\beta}_n - R\beta) \xrightarrow{d} \chi^2_p$$

[91] Azeem tosses up: $2(1 - \Phi(|T_n|))$. Why?

which suggests a test statistic T_n that's:

$$T_n = \mathcal{N}(R\hat{\beta}_n - c)'(R\hat{\Omega}_n R')^{-1}(R\hat{\beta}_n - c)$$

and critical value:

$$c_{p,1-\alpha} = F_p^{-1}(1-\alpha)$$

where $F_p(\cdot)$ is the c.d.f. of χ^2_p.

And if we really wanna go crazy we can test non-linear restrictions. So let's try to develop a test-statistic for:

$$H_0 : f(\beta) = c \text{ vs. } H_A : f(\beta) \neq c$$

for $f : \mathbb{R}^{k+1} \to \mathbb{R}^p$ where $f(\cdot)$ is continuously differentiable at β and assume that the rows are linearly independent. Then define $D_\beta f(\beta)$ to be the matrix of partials ($p \times k+1$). Then:

$$\sqrt{n}(f(\hat{\beta}_n) - f(\beta)) \xrightarrow{d} \mathcal{N}(0, D_\beta f(\beta)\Omega D_\beta f(\beta)')$$

and we can define the test-statistic we can use is:

$$T_n = n(f(\hat{\beta}_n - c)'((D_\beta f(\beta)\hat{\Omega}_n D_\beta f(\beta)'))^{-1} f(\hat{\beta}_n - c))$$

where the critical value we work with is:

$$c_{p,1-\alpha} = F_p^{-1}(1-\alpha).$$

Instrumental variables

We can think about IV as linear regression when $E[XU] \neq 0$. Let (Y, X, U) be s.t.:

$$Y = X'\beta + U$$

where:

$$X = (X_0, \dots, X_k)'$$

with $X_0 = 1$ and $E[XU] \neq 0$ and we'll assume that $E[U] = 0$[92] and any X_j with $E[X_jU] = 0$ is said to be *exogenous* and any X_j with $E[X_jU] \neq 0$ are said to be *endogenous*.

E.g. (Omitted variables): Let $k = 2$ with the following causal model for Y:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

with $E[XU] = 0$. The problem is that we don't observe X_2. Rewrite:

$$Y = \beta_0^* + \beta_1^* X_1 + U^*$$

with:

$$\begin{aligned}
&\beta_0^* = \beta_0 + \beta_2 E[X_2] \\
&\beta_1^* = \beta_1 \\
&U^* = \beta_2(X_2 - E[X_2]) + U
\end{aligned}$$

and $E[X_1U^*] = \beta_2 Cov(X_1, X_2)$. Obviously if we just run OLS we've got issues with our estimates.

E.g. (Measurement error): Partition X into X_0 and $X_1 \in \mathbb{R}^k$ with:

$$Y = \beta_0 + X_1'\beta_1 + U$$

with $E[XU] = 0$ but X_1 is unobserved. Instead observe:

$$\hat{X}_1 = X_1 + V$$

[92] Even if $E[U] \neq 0$ we can just have β_0 soak it up.

with $E[V] = 0$, $Cov(X_1, V) = 0$, & $Cov(V, U) = 0$. Then rewrite the model as:

$$\begin{aligned} Y &= \beta_0^* + \hat{X}_1' \beta_1^* + U^* \\ \text{with } \beta_0^* &= \beta_0 \\ \beta_1^* &= \beta_1 \\ U^* &= -V'\beta + 1 + U \end{aligned}$$

where $\hat{X}_1$ is endogenous because:

$$E[\hat{X}_1 U^*] = E[(X_1 + V)(-V\beta_1 + U)] = -E[VV']\beta_1$$

which is typically $\neq 0$.

E.g. (Simultaneity): The classic[93] supply and demand example goes as follows: Q^d is the quantity demanded and Q^s is the quantity supplied as a function of the (non-market clearing) price $\tilde{P}$. Assume:

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + U^s \end{aligned}$$

with: $E[U^d] = E[U^s] = E[U^d U^s] = 0$. All that we observe is the market clearing (Q, P) s.t. $Q^d = Q^s$. That is:

$$\beta_0^d + \beta_1^d P + U^d = \beta_0^s + \beta_1^s P + U^s$$

which implies that:

$$P = \frac{1}{\beta_1^d - \beta_1^s}(\beta_0^s - \beta_0^d + U^s - U^d)$$

and plugging in (Q, P) to the supply and demand equations we get that:

$$E[PU^d] = \frac{-Var(U^d)}{\beta_1^d - \beta_1^s} \neq 0.$$

Fuck. What should we do?

Solving for β

LET (Y, X, U) s.t. $Y = X'\beta + U$ and assume that there is a random vector $Z \in \mathbb{R}^{l+1}$ with:

$$l + 1 \geq k + 1 \ s.t. \ E[ZU] = 0$$

and assume that any exogenous X_j are included in Z. We'll also have to assume:

[93] CLASSIC!

1. No perfect colinearity in Z.
2. $E[ZX'] < \infty$.
3. $E[ZZ'] < \infty$.
4. $E[ZX']$ has rank $= k+1$.[94,95]

Now in order to solve for β, note that:

$$E[ZU] = 0 \implies E[Z(Y - X'\beta)] = 0 \implies E[ZX']\beta = E[ZY]$$

then if $l+1 = k+1$ then:

$$\beta = E[Z'X]^{-1}E[ZY]$$

but otherwise the system is over-determined, so to solve explicitly for β, the following lemma is useful:

Lem.: *Assume there is no perfect colinearity in Z. Then $E[ZX']$ has rank $k+1$ i.f.f. π has rank $k+1$ when:*

$$BLP(X|Z) = \pi' Z.$$

Moreover, $\pi' E[ZX'] = \pi' E[ZZ']\pi$ is invertible.

Proof. As is tradition, we'll start the proof with a note. Note: $X = \pi' Z + V$ when $E[ZV'] = 0$. So:

$$E[ZX'] = E[ZZ']\pi$$

Then using the useful Lemma to the right (→) we get that:

$$rank(E[ZX']) = rank(E[ZZ']\pi) \leq rank(\pi) = rank(E[ZZ']^{-1}E[ZZ']\pi) \leq rank(E[ZZ']\pi) = rank(E[ZX'])$$

which gives us one direction of the proof.

Next:

$$\pi' E[ZX'] = \pi' E[ZZ']\pi$$

and invertible using usual arguments.[96] □

Lem. (Rank inequality): *For any conformable matrices A, B, $rank(AB) \leq \min\{rank(A), rank(B)\}$.*

Proof. Huh? Azeem does something by the rows and then something by the columns. □

To solve for β explicitly, multiply both sides by π' to get:

$$\pi' E[ZX']\beta = \pi' E[ZY]$$

which gives us that:

$$\beta = (\pi' E[ZX'])^{-1}(\pi' E[ZY]) = (\pi' E[ZZ']\pi)^{-1}(\pi' E[ZY]).$$

If $l+1 = k+1$ then we say that β is exactly identified.[97] When things are exactly identified then π is invertible because it's square and has full rank, so:

$$\beta = E[ZX']^{-1}E[ZY].$$

[94] Often called instrument relevance.

[95] A sufficient condition for this to hold is $l+1 > k+1$.

[96] From some problem set?

[97] Because otherwise, it's overidentified.

Also, if $l+1 = k+1$ then we can derive an expression for the slope parameter alone:

$$Y = \beta_0 + X_1 \beta_1 + U$$

and:

$$E[U] = 0 \implies \beta_0 = E[Y] - E[X_1]'\beta_1$$

and write $Z = (1, Z_1')$ to get:

$$Y - E[Y] = (X_1 - E[X_1])'\beta_1 + U$$

which we can pre-multiply by Z_1 and take expectations of to get:

$$E[Z_1(Y - E[Y])] = E[Z_1(X_1 - E[X_1])']\beta_1$$

which gives us the familiar:

$$\beta_1 = Cov(Z_1, X_1)^{-1} Cov(Z_1, Y).$$

THIS gives us a way to think about reinterpreting the rank condition we imposed earlier. If $l+1 = k+1$ and all $X_0, \ldots, X_{k-1}$ are all exogenous and X_k is endogenous, so:

$$Z = (Z_0, \ldots, Z_k)'$$

with $Z_j = X_j$ for $0 \leq j \leq k-1$. Then:

$$\pi' = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \\ \pi_0 & \pi_1 & \pi_2 & \pi_3 & \ldots & \pi_{k-1} & \pi_k \end{bmatrix}$$

and in this case the rank condition means that $\pi_k \neq 0$. That is, the only way for π' to be invertible is for $\pi_k \neq 0$. In words this means that we require Z_k to be correlated with X_k after controlling for $X_0, \ldots, X_{k-1}$.

REVISITING our examples from above:

E.g. (Omitted variables: Part 2): Require Z_1 s.t. $Z = (1, Z_1)'$ satisfies an instrument exogeneity and relevance. Exogeneity requires Z_1 to be uncorrelated with X_2 and U. Relevance requires $\pi_1 \neq 0$ in the BLP regression $X_1 = \pi_0 + \pi_1 Z_1 + V$. In this particular case, relevance boils down to:[98]

$$Cov(X_1, Z_1) \neq 0.$$

[98] An example from the empirical literature would be $Y = log(wages)$ and $X_1 = yrs.school$ and $X_2 = unobservedability$ and Z_1 is something like mother's education or distance to nearest school or something else equally dumb.

E.g. (Measurement error: Part 2): Here we require Z_1 s.t. $(1, Z_1')'$ satisifies instrument exogeneity and relevance. Then suppose:

$$Z_1 = X_1 + W$$

with $E[W] = 0$ and $Cov(X_1, W) = 0 = Cov(U, W)$. We assume further that $Cov(V, W) = 0$. Instrument exogeneity here requires:

$$E[Z_1 U^*] = E[(X_1 + W)(-V'\beta_1 + U)] = 0$$

and instrument relevance requires:

$$E[Z(1, \hat{X}_1')] = \begin{bmatrix} 1 & E[\hat{X}_1'] \\ E[Z_1] & E[Z_1 \hat{X}_1'] \end{bmatrix} = E[XX']$$

where we just plug in the values for $Z_1 \hat{X}_1'$ in the matrix to get that result. This is cool because it gives us the sense in which repeated measurements are instruments.

E.g. (Simultaneity: Part 2): Augment models for Q^d, Q^s as follows:

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + \beta_2^d Z_1 + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + U^s \end{aligned}$$

with:

$$E[Z_1 U^d] = E[Z_1 U^s] = E[U^d] = E[U^s] = E[U^d U^s] = 0.$$

Then market clearing implies that the market clearing price, P:

$$P = \frac{1}{\beta_1^d - \beta_1^s}(\beta_0^s - \beta_0^d - \beta_2^d Z_1 + U^s - U^d)$$

and in the supply equation:

$$Q = \beta_0^s + \beta_1^s P + U^s$$

with $Z = (1, Z_1)'$ satisfying instrument exogeneity and relevance.[99]

Solving for subvectors of β

Just as we did with OLS there's some neat tools for solving subvectors of estimators in IV regressions. Recall that we're working with:

$$Y = X'\beta + U$$

with $E[ZU] = 0$, $E[ZZ']$ exists, there's no perfect colinearity in Z, $E[ZX']$ exists, and there's rank of $E[ZX'] = k + 1$. Then we'll split up our model into:

$$Y = X_1'\beta_1 + X_2'\beta_2 + U$$

[99] We get relevance if $\beta_2^d \neq 0$.

where X_2 is exogenous and Z_1 is our instrument for X_1 and $Z_2 = X_2$.
Then:

$$BLP(Y|Z_2) = BLP(X_1|Z_2)'\beta_1 + X_2'\beta_2 + \underbrace{BLP(U|Z_2)}_{=0}$$

and then:

$$Y^* = X_1^{*'}\beta_1 + U$$

with:

$$Y^* = Y - BLP(Y|Z_2)$$

and:

$$X_1^* = X_1 - BLP(X_1|Z_2)$$

and if exactly identified then:

$$\beta_1 = E[Z_1 X_1^*]^{-1} E[Z_1 Y]$$

and if overidentified, then let $\hat{X}_1^* = BLP(X_1^*|Z_1)$, and we get:

$$\beta_1 = E[\hat{X}_1^* X_1^*]^{-1} E[\hat{X}_1^* Y^*] = E[\hat{X}_1^* \hat{X}_1^{*'}]^{-1} E[\hat{X}_1^* Y]$$

because $X_1^* = \hat{X}_1^* + V$ with $E[\hat{X}_1^* V'] = 0$.

Estimating β

With the same setup as above, with sample observations $(Y_1, X_1, Z_1), \ldots, (Y_1, X_n, Z_n)$ i.i.d. (Y, X, Z). If exactly identified then we can just write the sample analog for $\beta = E[ZX']^{-1}E[ZY]$:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right)$$

and equivalently, $\hat{\beta}_n$ satisfies:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n} Z_i \hat{U}_i = 0.$$

In matrix notation, define:

$$\mathbb{X} = (X_1, \ldots, X_n)'$$

$$\mathbb{Z} = (Z_1, \ldots, Z_n)'$$

$$\mathbb{Y} = (Y_1, \ldots, Y_n)'$$

then:

$$\hat{\beta}_n = (\mathbb{Z}'\mathbb{X})^{-1}\mathbb{Z}\mathbb{Y}.$$

ANOTHER approach to estimating β is the two-stage least squares[100] estimator. Recall:

$$\beta = E[\pi' Z X']^{-1} E[\pi' Z Y] = E[\pi' Z Z' \pi]^{-1} E[\pi' Z Y]$$

and so by analogy:

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_n' Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_n' Z_i Y_i \right)$$

and because $X_i = \hat{X}_i + \hat{V}_i = \hat{\pi}_n' Z_i + \hat{V}_i$ when $\frac{1}{n} Z_i \hat{V}_i' = 0$ we can write:[101]

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_n' Z_i Z_i' \hat{\pi}_n \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_n' Z_i Y_i \right)$$

and to get the analog for $\pi = E[ZZ']^{-1} E[ZX']$ we use:

$$\hat{\pi}_n = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i' \right)^{-1} \left(\frac{1}{n} Z_i X_i' \right).$$

We can think of $\hat{\beta}_n$ as the IV estimator using $\hat{X}_i$ as instruments. $\hat{U}_n$ is orthogonal to included exogenous regressions. Furthermore, if $\hat{\pi}_n$ is invertible, the formula reduces to the IV formula, which allows us to see why it's called two-stage least squares. In matrix notation:

$$\hat{\mathbb{X}} = (\hat{X}_1, \ldots, \hat{X}_n)' = \mathbb{P}_2 \mathbb{X}$$

where $\mathbb{P}_2 = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$, so:

$$\hat{\beta}_n = (\hat{\mathbb{X}}'\mathbb{X})^{-1}(\hat{\mathbb{X}}'\mathbb{Y}) = (\mathbb{X}'\mathbb{P}_2\mathbb{X})^{-1}(\mathbb{X}'\mathbb{P}_2\mathbb{Y}).$$

TO ESTIMATE subvectors of β in this framework, partition X into X_1, X_2 where X_2 is exogenous and Z into Z_1, Z_2 where $Z_2 = X_2$:

$$Y = X_1'\beta_1 + X_2'\beta_2 + U$$

and define:

$$\hat{\mathbb{U}} = (\hat{U}_1, \ldots, \hat{U}_n)'$$

$$\mathbb{X}_1 = (X_{1,1}, \ldots, X_{1,n})'$$

with $\mathbb{X}_2, \mathbb{Z}_1, \mathbb{Z}_2, \mathbb{P}_Z, \mathbb{P}_2, \mathbb{M}_Z$ defined equivalently and:

$$\mathbb{P}_1 = \mathbb{Z}_1(\mathbb{Z}_1'\mathbb{Z}_1)^{-1}\mathbb{Z}_1$$

$$\mathbb{M}_1 = \mathbb{I} - \mathbb{P}_1$$

and note that:

$$\mathbb{Y} = \mathbb{X}_1\hat{\beta}_{1,n} + \mathbb{X}_2\hat{\beta}_{2,n} + \hat{\mathbb{U}}$$

[100] TSLS.

[101] This definition of $\hat{\beta}_n$ satisifes:
$$\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_n' Z_i (Y_i - X_i'\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_n' Z_i \hat{U}_i = 0.$$

and:

$$\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \hat{\mathbb{U}}$$

and if exactly identified, multiply through by $\mathbb{Z}_1$ to get:

$$\hat{\beta}_{1,n} = (\mathbb{Z}_1'\mathbb{M}_2\mathbb{X}_1)^{-1}(\mathbb{Z}_1'\mathbb{M}_2\mathbb{Y})$$

and if overidentified, multiply through by $(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1)'$ to get:

$$\mathbb{X}_1\mathbb{M}_2\mathbb{P}_1\mathbb{M}_2\mathbb{Y} = \mathbb{X}_1'\mathbb{M}_2\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \mathbb{X}_1'\mathbb{M}_2\mathbb{P}_1\hat{\mathbb{U}}$$

and $\mathbb{X}_1'\mathbb{M}_2\mathbb{P}_1\hat{\mathbb{U}} = 0$ because:

$$\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1 = \mathbb{P}_Z\mathbb{M}_2\mathbb{X}_1$$

so enough to show that:

$$\mathbb{X}_1'\mathbb{M}_2\mathbb{P}_Z\hat{\mathbb{U}} = 0$$

and $(\mathbb{P}_Z\mathbb{X})'\hat{\mathbb{U}} = 0$ because we can add and subtract the same number:[102]

$$(\mathbb{P}_Z\mathbb{X})'\hat{\mathbb{U}} = (\mathbb{P}_Z\mathbb{M}_2\mathbb{X} + \mathbb{P}_Z\mathbb{P}_2\mathbb{X})'\hat{\mathbb{U}} = \mathbb{X}'\mathbb{M}_2\mathbb{P}_Z'\hat{\mathbb{U}} + \mathbb{X}'\mathbb{P}_2\mathbb{P}_Z\hat{\mathbb{U}} = 0$$

because:

$$\mathbb{X}'\mathbb{P}_2\mathbb{P}_Z\hat{\mathbb{U}} = \mathbb{X}'\mathbb{P}_2\hat{\mathbb{U}} = 0$$

because $\mathbb{P}_2\hat{\mathbb{U}} = 0$.

Properties of TSLS estimator

SAME setup as before.

Excercise: IV is not unbiased.

Lem. (Consistency): *Under the assumptions of IV $\hat{\beta}_n \xrightarrow{p} \beta$.*

Proof. For $\beta = E[\pi' Z Z' \pi]^{-1} E[\pi' Z Y]$ and:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_n' Z_i Z_i' \hat{\pi}_n\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_n' Z_i Y_i\right)$$

and from OLS we have that $\hat{\pi}_n \xrightarrow{p} \pi$ and that the same averages are going to converge to their analogs and by CMT we get that $\hat{\beta}_n \xrightarrow{p} \beta$. □

Lem. (Limiting distribution): *On top of the standard IV assumptions, if we assume further that:*

$$Var[ZU] = E[ZZ'U^2]$$

[102] That is,

$$\mathbb{P}_Z\mathbb{I}\mathbb{X} = \mathbb{P}_Z(\mathbb{M}_2 + \mathbb{P}_2)\mathbb{X}.$$

exists, then:

$$\sqrt{n}(\hat{\beta}_n - \beta) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_n' Z_i Z_i' \hat{\pi}_n\right)^{-1}}_{\xrightarrow{p} (\pi' E[ZZ']\pi)^{-1}} \quad \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\pi}_n' Z_i U_i\right)}_{= \hat{\pi}_n' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i U_i \xrightarrow{d} \mathcal{N}(0, \pi' Var(ZU)\pi).}$$

so:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

where:

$$\Omega = (\pi' E[ZZ']\pi)^{-1} \pi' Var[ZU] \pi (\pi' E[ZZ']\pi)^{-1}.$$

Excercise: Simplify if $E[U|Z] = 0$ and $Var(U|Z) = \sigma^2$. Homoskedasticity.

Excercise: Write out estimator under homoskedasticity.

Lem. (Consistent estimation of Ω): *More generally, without homoskedasticity, need to estimate $Var(ZU)$. Use:*

$$Var(ZU) = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \hat{U}_i^2$$

where:

$$\hat{U}_i = Y_i - X_i' \hat{\beta}_n.$$

Weak insturments

Asymptotic approximation above for $\sqrt{n}(\hat{\beta}_n - \beta)$ may be poor in finite-sample when $E[ZX']$ is "close" to having rank $< k + 1$.[103] Let's consider a simple version for the regression case:

$$Y_i = \beta X_i + U_i$$

and:

$$X_i = \pi Z_i + V_i$$

with:

$$(U_1, V_1), \ldots, (U_n, V_n) \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$$

where:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

and $Z_1, \ldots, Z_n$ are non-random[104] and $\pi \neq 0$. Then in this case:

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_i Z_i Y_i}{\frac{1}{n}\sum_i Z_i X_i} = \beta + \frac{\frac{1}{n}\sum_i Z_i U_i}{\left(\frac{1}{n}\sum_i Z_i^2\right) + \frac{1}{n}\sum_i Z_i V_i}$$

and define $\overline{Z_n^2} := \frac{1}{n}\sum_i Z_i^2$ so:

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_i Z_i U_i}{\overline{Z_n^2}\pi + \frac{1}{n}\sum_i Z_i V_i}$$

[103] Analogy: Confidence intervals for Bernoulli(p).

[104] Deterministic.

and the joint distribution of the numerator/denominator[105] is:

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_i Z_i U_i \\ \overline{Z^2_n}\pi + \frac{1}{n}\sum_i Z_i V_i \end{pmatrix} \sim \mathcal{N}\left(\begin{matrix} 0 \\ \overline{Z^2_n}\pi \end{matrix}, \begin{bmatrix} \overline{Z^2_n}\sigma_1^2 & \frac{1}{\sqrt{n}}\overline{Z^2_n}\sigma_{1,2} \\ \frac{1}{\sqrt{n}}\overline{Z^2_n}\sigma_{1,2} & \frac{1}{n}\overline{Z^2_n}\sigma_2^2 \end{bmatrix} \right)$$

and if $\overline{Z^2_n} \to \overline{Z^2}$ then:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2}{\pi^2 \overline{Z^2}}\right)$$

and we expect this to be a "good" approximation when:

$$\overline{Z^2_n}\pi >> \frac{1}{\sqrt{n}}\sigma_2 \sqrt{\overline{Z^2_n}}.$$

Something about how you can always pick a π s.t. this is bad? Probably related to the Bernoulli example.

METHODS that don't suffer from this problem have been developed, though.[106] To see this, we'll return to our general framework and suppose we wanted to test:

$$H_0 : \beta = c \text{ vs. } H_A : \beta \neq c$$

and define:

$$W_i(c) := Z_i(Y_i - X_i'c).$$

Then under H_0, $W_i(c) = Z_i U_i$ and we know that in this case $W_i(c)$ is an i.i.d. sequence of mean zero random vectors. Under H_A:

$$W_i(c) = Z_i(Y_i - X_i'\beta + X_i'(\beta - c)) = Z_i U_i + Z_i X_i'(\beta - c)$$

which may not have mean zero. This suggests a test we've already developed:

$$T_n = n\bar{W}_n(c)'\hat{\Sigma}_n^{-1}(c)\bar{W}_n(c)$$

where:

$$W_n(c) = \frac{1}{n}\sum_{i=1}^{n} W_i(c)$$
$$\hat{\Sigma}_n(c) = \frac{1}{n}\sum_{i=1}^{n}(W_i(c) - \bar{W}_n(c))(W_i(c) - \bar{W}_n(c))'$$

and critical value $c_{l+1,1-\alpha}$ which is the $(1-\alpha)^{th}$ quantile of the χ^2_{l+1}.[107] A closely related variant is the *Anderson-Rubin test*. The idea is to regress:

$$Y_i - X_i'c \text{ on } Z_i$$

and test whether the coefficients are all equal to zero. When the model is exactly identified[108] some recent research suggests that these tests have "good" power. But if the model is overidentified then

[105] Upstairs and downstairs.

[106] Robust methods for weak instruments.

[107] This is legit, but probably very underpowered.

[108] I.e., $l + 1 = k + 1$.

it's possible that you can do better. Also, if you're only interested in certain components of β, e.g., β_1, it's possible that you can do better.

Another approach is to think about a two-step procedure. For example, you might consider testing the null that:

$$H_0 : rankE[ZX'] < k+1 \text{ vs. } H_A : rankE[ZX'] = k+1$$

and in some cases this is easy to do, like when there's only one endogenous regressor. To do this, you'd just test whether the coefficients in the first-stage are all zero. However, this doesn't solve the problem. Will still behave poorly in finite samples.

Efficiency

We could have solved for β using any $l+1 \times k+1$ matrix Γ s.t. $E[\Gamma' ZX']$ has rank $k+1$. Then:

$$\beta = E[\Gamma' ZX']^{-1} E[\Gamma' ZY]$$

which would have given us the following estimator:

$$\tilde{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \Gamma' Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \Gamma' Z_i Y_i \right)$$

and you can work out the limiting distribution of this estimator to be:

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \tilde{\Omega})$$

where:

$$\tilde{\Omega} = E[\Gamma' ZX']^{-1} \Gamma' Var(ZU) \Gamma E[\Gamma' ZX']^{-1'}.$$

Recall that for our original estimator the limiting distribution was:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

with:

$$\Omega = E[\pi' ZZ' \pi]^{-1} \pi' Var(ZU) \pi E[\pi' ZZ' \pi]^{-1}$$

and under certain assumptions $\Gamma = \pi$ is the "best." I.e.:

$$\Omega \leq \tilde{\Omega}.$$

The assumptions we need are:

$$E[U|Z] = 0 \; \& \; Var(U|Z) = \sigma^2$$

which gives us that $Var(ZU) = \sigma^2 E[ZZ']$. Then define $W := \Gamma' Z$ and $W^* := \pi' Z$. Then our two variance-covariance terms reduce to:[109]

[109] The following trick helps us get the expression for $\tilde{\Omega}$, $E[\Gamma' ZX'] = E[\Gamma' ZZ' \pi]$ because $X = \pi' Z + V$.

$$\Omega = \sigma^2 E[W^* W^{*'}]^{-1}$$
$$\tilde{\Omega} = \sigma^2 E[WW^{*'}]^{-1} E[WW'] E[WW^{*'}]^{-1'}$$

Then we can use the following trick:

$$\Omega \leq \tilde{\Omega} \iff \Omega^{-1} \geq \tilde{\Omega}^{-1}.$$

Then:

$$\Omega^{-1} \geq \tilde{\Omega}^{-1} \iff E[W^* W^{*'}] - E[WW^{*'}]' E[WW']^{-1} E[WW^{*'}] \geq 0 \iff E[W'] \geq 0$$

where $V = W^* - BLP(W^*|W)$.[110]

Heterogeneity

RECALL our model is just: $Y = X'\beta + U$. This implies that a change in X from $X = x$ to $X = \tilde{x}$ holding everything else constant is the *same* for everybody. How do we relax this? One approach is to treat β as random.[111] Then we can absorb the error term into β and our model becomes:

$$Y = X'\beta.$$

This is a hot area of research right now. We'll just consider a simple case, though. Take $k = 1$ and $X_1 = D$ where $D \in \{0, 1\}$. So:

$$Y = \beta_0 + \beta_1 D$$

where we can think about D as being some treatment.[112] Neyman introduced the potential outcomes framework to think about this problem.[113] The potential outcomes are (Y_0, Y_1) where Y_0 is the outcome if not treated and Y_1 is outcome is treated. In this notation:

$$Y = DY_1 + (1 - D)Y_0.$$

And the following words tend to be used to describe objects that emerge from this setup:

1. $Y_1 - Y_0 =$ treatment effect.

2. $E[Y_1 - Y_0] =$ average treatment effect.

3. $E[Y_1 - Y_0 | D = 1] =$ average treatment effect on the treated.

[114] Suppose that (Y_1, Y_0) is independent of D and $P\{D = 1\} \in (0, 1)$.[115] In this case, OLS of Y on $(1, D)$ will consistently estimate the average treatment effect. Why?[116]

[110] Azeems says: Check!

[111] Random coefficients model.

[112] Job training program? Red-pill vs. blue-pill.

[113] Suck it, Rubin!

[114] In this equation:

$$Y = \beta_0 + \beta_1 D$$

take $\beta_0 = Y_0$ and $\beta_1 = Y_1 - Y_0$.

[115] E.g., a randomized trial.

[116] Supposedly we showed this first-step on a problem set.

$$\begin{aligned}\frac{Cov(Y,D)}{Var(D)} &= E[Y|D=1] - E[Y|D=0] \\ &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= E[Y_1 - Y_0].\end{aligned}$$

What if (Y_1, Y_0) is not independent of D? Then suppose there is a binary instrument Z on D. Then the TSLS regression of Y on D with Z as an instrument. Then the slope estimand in this regression is:

$$\begin{aligned}\frac{Cov(Y,Z)}{Cov(D,Z)} &= \frac{\frac{Cov(Y,Z)}{Var(Z)}}{\frac{Cov(D,Z)}{Var(Z)}} \\ &= \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}\end{aligned}$$

and it's useful to introduce potential treatments D_1, D_0 where:

$$D = D_1 Z + D_0(1-Z)$$

and instrument relevance requires:

$$P\{D_1 \neq D_0\} > 0$$

while instrument exogeneity requires:

$$(Y_1, Y_0, D_1, D_0) \text{ independent } Z.$$

We also assume monotonicity,[117] which requires:

$$P\{D_1 \geq D_0\} = 1.$$

OK. Now that we've got all this notation, let's work to simplify our slope estimand. First, note:

$$\begin{aligned}E[Y|Z=1] - E[Y|Z=0] &= E[Y_1 D + Y_0(1-D)|Z=1] - E[Y|Z=0] \\ &= E[Y_1 D_1 + Y_0(1-D+1)|Z=1] - E[Y|Z=0] \\ &= E[Y_1 D_1 + Y_0(1-D+1)|Z=1] - E[Y_1 D_0 + Y_0(1-D_0)] \\ &= E[(Y_1 - Y_0)(D_1 - D_0)] \\ &= E[Y_1 - Y_0 | D_1 - D_0 \neq 0] P\{D_1 - D_0 \neq 0\} \\ &= E[Y_1 - Y_0 | D_1 > D_0] P\{D_1 > D_0\} \text{ by uniformity.}\end{aligned}$$

Then by the same type of steps:

$$E[D|Z=1] - E[D|Z=0] = E[D_1 - D_0] = P\{D_1 > D_0\}$$

and we can express our estimand as:

$$E[Y_1 - Y_0 | D_1 > D_0] =: LATE.$$

[117] Sometimes called uniformity, because it applies to every person.

That is, we only observe the population where $D_1 > D_0$ which is unobserved.[118] Is this interesting? Another important take-away is that different instruments estimate different treatment effects. This observation is unique to a world with random effects to treatment. If treatment effects are constant then this isn't an issue.[119] Also, Heckman and Vytlacil have unified this stuff.[120]

[118] The compliers.

[119] Can you include covariates? People do, but there's not a great justification.

[120] World congress paper in ECMA is the best.

Maximum Likelihood Estimation

Paradigm shift time: Let's venture into the world of parameterized models.

Unconditional Maximum Likelihood Estimators

Suppose $X_1, \dots, X_n$ is an i.i.d. sequence of random vectors with distribution P and we'll assume $P = P_{\theta_0}$ and $\theta_0 \in \Theta \subseteq \mathbb{R}^d$. That is,

$$\{P_\theta : \ \theta \in \Theta\}$$

is a parametric model for P. Assume P_θ has a density q_θ with respect to a common measure μ. The likelihood of $X_1, \dots, X_n$ is simply the joint density of $X_1, \dots, X_n$ evaluated at $X_1, \dots, X_n$. Mathematically we can define the likelihood function:

$$l_n(\theta) := \prod_{1 \le i \le n} q_\theta(X_i)$$

and the ML estimator of θ, $\hat{\theta}_n$ is simply any:[121]

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} l_n(\theta).$$

Typically, though, we work with the follow transformation:

$$L_n(\theta) = \frac{1}{n} \log(l_n(\theta)) = \frac{1}{n} \sum_{i=1}^{n} \log q_\theta(X_i)$$

which is convenient because it's now just a sample average and we've got a ton of tools at our disposal to work with that sort of object.[122]

E.g. (ML Estimation with Bernoulli RVs): $X_1, \dots, X_n$ i.i.d. $P = P_{\theta_0} = \mathcal{B}(\theta_0)$, $\theta_0 \in (0, 1)$. Then we know our density is:

$$q_\theta(x) = \begin{cases} \theta, & x = 1 \\ 1 - \theta, & x = 0 \\ 0, & o/w \end{cases}$$

[121] ML estimators need not exist. For that case we just take the "near" maximizer. Also, ML estimators need not be unique. Just take any one.

[122] See: Everything above.

which we can write as:

$$q_\theta(x) = \theta^x(1-\theta)^{1-x}$$

so:

$$l_n(\theta) = \prod_{i=1}^n \left(\theta^{X_i}(1-\theta)^{1-X_i}\right)$$

and then:

$$\begin{aligned} L_n(\theta) =& \frac{1}{n}\sum_{i=1}^n X_i \log\theta + (1-X_i)\log(1-\theta) \\ =& \bar{X}_n \log\theta + (1-\bar{X}_n)\log(1-\theta) \end{aligned}$$

and the first-order conditions imply that $\hat{\theta}_n = \bar{X}_n$.

E.g. (Uniform): $X_1, \ldots, X_n$ i.i.d. $P = P_{\theta_0} = \mathcal{U}(0, \theta_0)$ where $0 \leq \theta_0 < \infty$. Then:[123]

$$q_\theta(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta \\ 0, & o/w \end{cases}$$

giving us that:

$$l_n(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 \leq X_i \leq \theta \ \forall i \\ 0, & o/w \end{cases}$$

which implies that:

$$\hat{\theta}_n = \max_{1\leq i\leq n} X_i.$$

Why? Well because $l_n(\theta)$ is decreasing in θ for θ s.t. $0 \leq X_i \leq \theta$, $\forall i$ and $X_j \leq \max_i\{X_i\} \leq \theta$.[124]

E.g. (Mixed distributions): Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P = P_{\theta_0}$ where $\theta_0 \in (-\infty, \infty)$ and P_θ is distribution of $X = \max\{Z - \theta, 0\}$ where $Z \sim \mathcal{N}(0,1)$. Then:

$$P\{X = 0\} = \Phi(\theta)$$

so the density of $Z - \theta$ is $\phi(x+\theta)$, so:

$$q_\theta = \begin{cases} \Phi(\theta) & , \text{ if } \ x = 0 \\ \phi(x+\theta) & , \text{ if } \ x > 0 \end{cases}$$

then:

$$l_n(\theta) = \prod_{1\leq i\leq n: X_i = 0} \Phi(\theta) \prod_{1\leq i\leq n: X_i > 0} \phi(X_i + \theta)$$

so in this case there's no easy closed form expression for $\hat{\theta}_n$. But we can still study its properties using its implied characterizations as a (near) maximizer.

[123] Working with this MLE was on the 2013 Core.

[124] Where X_j is all X_i that aren't the maximum of all the X_is.

Conditional ML Estimators

SUPPOSE now that we have $(Y_1, X_1), \ldots, (Y_n, X_n)$ is an i.i.d. sequence of random vectors with distribution P. Write $P \to (P_{Y|X}, P_X)$. Assume that:

$$P_{Y|X} = P_{\theta_0} \text{ where } \theta_0 \in \Theta$$

and assume P_θ has a density (w/r/t some μ) q_θ. The conditional likelihood of $(Y_1, X_1), \ldots, (Y_n, X_n)$ is just the joint conditional density evaluated at $(Y_1, X_1), \ldots, (Y_n, X_n)$. So,

$$l_n(\theta) = \prod_{1 \le i \le n} q_\theta(Y_i|X_i)$$

and again, it's easier to work with:

$$L_n(\theta) = \frac{1}{n} \log l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log q_\theta(Y_i|X_i)$$

and a conditional ML estimator $\hat{\theta}_n$ is any (near) maximum of $l_n(\theta)$ by taking X_i =constant, we get unconditional ML estimators.

E.g. (Probit model): $(Y_1, X_1), \ldots, (Y_n, X_n)$ i.i.d. P and assume the marginal distribution $P_{Y|X} = P_{\theta_0}$ with $\theta_0 \in \mathbb{R}^{k+1}$, $Y_i \in \{0, 1\}$, $X_i \in \mathbb{R}^{k+1}$. Then we'll further assume that:[125]

$$q_\theta(y|x) = \begin{cases} \Phi(x'\theta) & , \text{ if } y = 1 \\ 1 - \Phi(x'\theta) & , \text{ if } y = 0 \end{cases}$$

which is a lot easier to work with if we just write:

$$q_\theta(y|x) = \Phi(x'\theta)^y (1 - \Phi(x'\theta))^{1-y}$$

which gives us the following log-likelihood function:

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i \log \Phi(X_i'\theta) + (1 - Y_i) \log(1 - \Phi(X_i'\theta)).$$

E.g. (Mul):

Properties of ML Estimators

HOPEFULLY this new approach to estimation has the sorts of features of an estimator that we like.

[125] Can think of $Y_i = I\{X_i'\theta \ge \epsilon_i\}$ where $\epsilon_i \sim \mathcal{N}(0, 1)$.

Consistency

WITH $L_n(\theta) = \frac{1}{n}\sum_i q_\theta(Y_i|X_i)$ with $\hat{\theta}_n$ a "near" maximizer of this function $L_n(\theta)$ and there's no general closed form solution to $\hat{\theta}_n$ so it'll take a little bit of work to see whether $\hat{\theta}_n \xrightarrow{p} \theta_0$. First, define:

$$L(\theta) := E[\log q_\theta(Y_i|X_i)]$$

where we're taking the expectation over P_{θ_0}, P_X, not P_θ. Second, here's a Lem.:

Lem.: *If $\forall\ \theta \neq \theta_0$: $P\{q_\theta(Y_i|X_i) \neq q_{\theta_0}(Y_i|X_i)\} > 0$ then $L(\theta)$ is uniquely maximized at $\theta = \theta_0$.*

Proof. The proof is essentially an application of Jensen's inequality.[126] Define:

$$\begin{aligned} M(\theta) :=& L(\theta) - L(\theta_0) \\ =& E\left[\log \frac{q_\theta(Y_i|X_i)}{q_{\theta_0}(Y_i|X_i)}\right] \end{aligned}$$

and then by Jensen's inequality we have that:

$$M(\theta) \leq \log E\left[\frac{q_\theta(Y_i|X_i)}{q_{\theta_0}(Y_i|X_i)}\right]$$

and we'll express the expectation as an integral to simply:

$$E\left[\frac{q_\theta(Y_i|X_i)}{q_{\theta_0}(Y_i|X_i)}\right] = \int\int \frac{q_\theta(y|x)}{q_{\theta_0}(y|x)} q_{\theta_0}(y|x) d\mu(y) dP_X(x)$$

which we can write as:

$$\int\int q_\theta(y|x) d\mu(y) dP_X(x) = 1$$

which is equal to 1 because we're just integrating over a density.[127]

Because $M(\theta_0) = 0$, we need to show that for $\theta \neq \theta_0$, $M(\theta) < 0$. And, since log is strictly concave $M(\theta) < 0$ unless for some c:

$$P\left\{\frac{q_\theta(Y_i|X_i)}{q_{\theta_0}(Y_i|X_i)} = c\right\} = 1$$

and there are three possible cases:

1. If $c > 1$ then we've violated the result from J's Inequality above.
2. If $c < 1$ then $M(\theta) < 0$ and we're good.
3. If $c = 1$ then we're good because the original assumption of the Lemma ruled this out.

[126] Full circle, bitches.

[127] That shit better equal 1.

□

Lem.: *Let $\hat{\theta}_n$ for $n \geq 1$ be such that:*[128]

$$L_n(\hat{\theta}_n) \geq L_n(\theta_0) - o_P(1)$$

and if:[129]

$$\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{p} 0$$

and if:[130]

$$\forall\, \delta > 0, \quad \sup_{\theta \in \Theta \setminus B_\delta(\theta_0)} L(\theta) < L(\theta_0)$$

then:

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

Proof. As is tradition, here we want to show that:

$$\forall \epsilon > 0, \; P\{|\hat{\theta}_n - \theta_0| > \epsilon\} \to 0.$$

When $|\hat{\theta}_n - \theta_0| > \epsilon$ we can use the unique maximizer result from above to write,

$$\begin{aligned} L(\theta_0) - L(\hat{\theta}_n) \geq L(\theta_0) - & \sup_{\theta \in \Theta \setminus B_\delta(\theta_0)} L(\theta) \\ &=: \eta \\ &> 0 \;\text{ by well-seperated condition.} \end{aligned}$$

Thus far we've shown that:

$$P\{|\hat{\theta}_n - \theta_0| > \epsilon\} \leq P\{L(\theta_0) - L(\hat{\theta}_n) \geq \eta\}$$

and we want to show that that RHS is gonna go to 0. Well, one useful fact is that:

$$L_n(\theta_0) \xrightarrow{p} L(\theta_0)$$

so:

$$L_n(\hat{\theta}_n) \geq L(\theta_0) - o_P(1)$$

so:

$$\begin{aligned} L(\theta_0) - L(\hat{\theta}_n) \leq & L_n(\hat{\theta}_n) - L(\hat{\theta}_n) + o_P(1) \;\text{ by def. of } \hat{\theta}_n \\ \leq & \sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| + o_P(1) \\ \xrightarrow{p} & 0 \;\text{ by uniform conv.} \end{aligned}$$

□

Now we've got useful conditions for consistency,[131] but it'd be nice to have some sufficient conditions that are easier to work with to get the consistency result.

[128] Which is clearly satisifed if:

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} L_n(\theta) - o_p(1).$$

[129] We call this condition the uniform convergence condition.

[130] We call this condition the well seperating condition.

[131] Well-seperating maximum and uniform convergence.

Lem.: *Let Θ be compact, $L(\theta)$ is continuous, and $L(\theta)$ is uniquely maximized at $\theta = \theta_0$. Then the well-separating condition holds.*

Proof. Let $\delta > 0$ be given. Then $\Theta \setminus B_\delta(\theta_0)$ is compact, so there exists $\theta^* \in \Theta \setminus B_\delta(\theta_0)$ s.t.:

$$\sup_{\theta \in \Theta \setminus B_\delta(\theta_0)} L(\theta) = L(\theta^*) < L(\theta_0)$$

and that $\theta^* \neq \theta_0$ where the last inequality holds by the unique maximum of continuous functions on compact sets. So θ_0 is the unique maximizer. □

Lem.: *$X_1, \ldots, X_n$ i.i.d. with distribution P, $\Theta \subseteq \mathbb{R}^k$ is compact. If $f(x, \theta)$ is continuous in θ for each value of x and there exists a function $F(x)$ s.t.:*[132]

$$|f(x, \theta)| \leq F(x) \ \forall x, \ \theta$$

and $E[F(X_i)] < \infty$ then:

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) - E[f(X_i, \theta)] \right| \xrightarrow{p} 0.$$

By applying this result with $f(x, \theta)$ given by $\log q_\theta(y|x)$ then we can get conditions for:

$$\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{p} 0.$$

An example of applying these might be useful:

E.g. (Consistency of Probit): Want to show that $\hat{\theta}_n \xrightarrow{p} \theta_0$. Recall that

$$\log q_\theta(y|x) = y \log \Phi(x'\theta) + (1 - y) \log(1 - \Phi(x'\theta)).$$

Assume that Θ is compact, there's no perfect collinearity in X_i and X_i is bounded.

1. *Well-separating*: Then to verify that it's well-separating we need:

$$L(\theta) = E[\log q_\theta(Y_i|X_i)]$$

is continuous in θ, Θ is compact (which we've just assumed), and to check unique maximizer. That is, we need to show $\theta \neq \theta_0$:

$$P\{q_\theta(Y_i|X_i) \neq q_{\theta_0}(Y_i|X_i)\} > 0$$

and:

$$P\{q_\theta(Y_i|X_i) \neq q_{\theta_0}(Y_i|X_i)\} = P\{\Phi(X_i'\theta) \neq \Phi(X_i'\theta_0)\} = P\{X_i'\theta \neq X_i'\theta_0\} = P\{X_i'(\theta - \theta_0) \neq 0\} > 0$$

which holds by the no perfect colinearity assumption. Then we have well-separating.

[132] Dominated.

2. *Uniform convergence*: We already have that Θ is compact, $\log q_\theta(y|x)$ is continuous in θ for all (y, x). Then all we have to check is the supremum condition:

$$\sup_{y,x,\theta} |\log q_\theta(y|x)| = \max \left\{ \sup_{x,\theta} |\log \Phi(x'\theta)|, \sup_{x,\theta} |\log(1 - \Phi(x'\theta))| \right\} \le M < \infty$$

where we can bound the second term because of the compactness of Θ and because X_i is bounded. Take $F(x) = M$ and apply the uniform law of large numbers.[133]

Some examples are hard to show the sufficient conditions for consistency but you can still show it more directly.

E.g. (Uniform): $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{U}(0, \theta_0),\ \theta_0 \ge 0$ where:

$$\hat{\theta}_n = \max_{1 \le i \le n} X_i$$

and we can show directly that $\hat{\theta}_n \xrightarrow{p} \theta_0$.

$$\begin{aligned} P\{|\hat{\theta}_n - \theta_0| > \epsilon\} &= P\{\theta_0 - \theta_n > \epsilon\} \text{ because } \theta_0 \ge \theta_n \\ &= P\{\max_i X_i < \theta_0 - \epsilon\} \\ &= P\{X_i < \theta_0 - \epsilon\}^n \\ &= \left(\frac{\theta_0 - \epsilon}{\theta_0}\right)^n \\ &\to 0 \end{aligned}$$

Mis-specification

What if the distribution of $Y|X \ne P_\theta$ for any θ?[134] Still reasonable to expect $\hat{\theta}_n \xrightarrow{p} \theta^*$ which maximizes $L(\theta)$. Why? Well to see, let $f(y|x)$ be the unknown density of $Y|X$ w/r/t μ. Then the θ^* that maximizes:

$$L(\theta) = E[\log q_\theta(Y_i|X_i)]$$

is the same as the θ^* that minimizes:

$$E[\log f(Y_i|X_i)] - E[\log q_\theta(Y_i|X_i)] = E\left[\log \frac{f(Y_i|X_i)}{q_\theta(Y_i|X_i)}\right] = d(f, q_\theta)$$

where we're taking the expectation over the true distribution, $f(y|x)$. The last term is always ≥ 0.[135] Also, it's $= 0$ for $f = q_\theta$, but it's not symmetric. Anyways, this conveys the sense in which you can estimate the parameters with the wrong model. It's similar to Interp. #2 of OLS.

[133] U.L.L.N.

[134] There's a common idea that MLE is robust to misspecifications.

[135] This comes off of Jensen's inequality. Take the negative of the $\log(\cdot)$.

Limiting distribution

UNDER fairly general conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega).$$

Lem.: *Suppose:*

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} L_n(\theta)$$

that:

$$\theta_0 \in \arg\max_{\theta \in \Theta} L(\theta)$$

that $\hat{\theta}_n \xrightarrow{p} \theta_0$, and that $\theta_0 \in Int(\Theta)$. Also assume that $\log q_\theta(y|x)$ is twice continuously differentiable in θ, $\forall (y, x)$. Also, for some $\delta > 0$:

$$|D_{\theta_j} \log q_\theta(y|x)| \leq M_1(y, x) \; \forall \theta \in B_\delta(\theta_0), \; j, \; (y, x)$$

and $E[M_1(Y_i, X_i)] < \infty$, where D_{θ_j} is the derivative w/r/t θ_j, where θ_j is the j^{th} component of θ. Also need a second-derivative equivalent. That is, suppose:

$$|D^2_{\theta_j, \theta_l} \log q_\theta(y|x)| \leq M_2(y, x) \; \forall \theta \in B_\delta(\theta_0), \; (j, l), \; (y, x)$$

and $E[M_2(Y_i, X_i)] < \infty$. Define:

$$A := E[D_\theta \log q_{\theta_0}(Y_i|X_i) D_{\theta'} \log q_{\theta_0}(Y_i|X_i)]$$

and:

$$B := E[D^2_{\theta, \theta'} \log q_{\theta_0}(Y_i|X_i)]$$

and assume that A, B both exist and B is non-singular. Then:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, B^{-1} A B^{-1}).$$

Proof. We'll do this proof in steps that we'll use numbers to demarcate.

1. W.T.S.

   $$E[D_\theta \log q_{\theta_0}(Y_i|X_i)] = 0$$

   which holds because $\theta_0 \in Int(\Theta)$ and is in the $\arg\max L(\theta)$ so $D_\theta L(\theta_0) = 0$ and $L(\theta_0) = \log q_{\theta_0}(y|x)$, which you can take expectations of and pass the D_θ through to get the desired expression.

2. W.T.S.

   $$D_\theta L_n(\hat{\theta}_n) = 0.$$

   We have that L_n is twice differentiable because $\log q_\theta$ is differentiable. Also, $\hat{\theta}_n \xrightarrow{p} \theta_0 \in Int(\Theta)$ so with probability approaching 1

$\hat{\theta}_n \in Int(\Theta)$ so because $\hat{\theta}_n \in \arg\max L_n(\theta)$ we have $D_\theta L_n(\hat{\theta}_n) = 0$. Now we can apply the Mean Value Theorem.[136] Then:

$$0 = D_\theta L_n(\hat{\theta}_n) = D_\theta L_n(\theta_0) + H_n(\hat{\theta}_n - \theta_0)$$

where H_n is the matrix whose j^{th} component equals the row (??) of:

$$D^2_{\theta,\theta'} L_n(\tilde{\theta}_{n,j})$$

where $\tilde{\theta}_{n,j}$ lies between θ_0 and $\hat{\theta}_n$. Then:

$$-H_n\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n} D_\theta L_n(\theta_0)$$

and then:

$$\sqrt{n} D_\theta L_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_\theta \log q_{\theta_0}(Y_i|X_i) \xrightarrow{d} \mathcal{N}(0, A).$$

Now to complete the proof, show that $H_n \xrightarrow{p} B$ and it's enough to show that $D^2_{\theta,\theta'} L_n(\tilde{\theta}_{n,j}) \xrightarrow{p} B$ since $\hat{\theta}_n \xrightarrow{p} \theta_0 \implies \tilde{\theta}_{n,j} \xrightarrow{p} \theta_0$. Then observe that:

$$\begin{aligned} D^2_{\theta,\theta'} L_n(\tilde{\theta}_{n,j}) =& \frac{1}{n} D^2_{\theta,\theta'} \log q_{\tilde{\theta}_{n,j}}(Y_i|X_i) \\ =& E[D^2_{\theta,\theta'} \log q_{\tilde{\theta}_{n,j}}(Y_i|X_i)] + o_p(1) \text{ by U.L.L.N.} \\ \to& E[D^2_{\theta,\theta'} \log q_{\theta_0}(Y_i|X_i)] \text{ because } \tilde{\theta}_{n,j} \xrightarrow{p} \theta_0 \end{aligned}$$

□

This is all kind of a mess, but often Ω simplifies. Suppose $D_\theta L(\theta_0) = 0$. Then if we pass the derivative through the expectation we get:

$$\begin{aligned} D_\theta L(\theta_0) =& E[D_\theta \log q_{\theta_0}(Y_i|X_i)] \\ =& \int\int D_\theta \log q_{\theta_0}(y|x) q_{\theta_0}(y|x) d\mu(y) dP_X(x) \\ =& \int\int D_\theta \log q_\theta(y|x) q_\theta(y|x) d\mu(y) dP_X(x) \text{ because functional relationship (huh?)} \\ =& 0 \text{ because of the original identity.} \end{aligned}$$

Then we can differentiate w.r.t. θ to get:

$$D_{\theta'} \int\int D_\theta \log q_\theta(y|x) q_\theta d\mu(y) dP_X(x) = 0$$

which turns into:

$$\int\int D^2_{\theta,\theta'} \log q_\theta(y|x) q_\theta(y|x) d\mu(y) dP_X(x) + \int\int D_\theta \log q_\theta(y|x) D_{\theta'} q_\theta(y|x) d\mu(y) dP_X(x) = 0$$

and note that:

$$D_{\theta'} q_\theta(y|x) = \frac{D_{\theta'} q_\theta(y|x)}{q_\theta(y|x)} q_\theta(y|x) = D_{\theta'} \log q_\theta(y|x) q_\theta(y|x)$$

[136] Are you fucking shitting me? Wow. Anyways, MVT says:

$$f(x+h) = f(x) + f'(\tilde{x})h$$

with $\tilde{x}$ on line segment between $x, x+h$.

so evaluating at $\theta = \theta_0$ we get:

$$-E[D^2_{\theta,\theta'} \log q_{\theta_0}(Y_i|X_i)] = E[D_\theta \log q_{\theta_0}(Y_i|X_i) D_{\theta'} q_{\theta_0}(Y_i|X_i)]$$

which is just:

$$-B = A$$

using our terminology from above. So plugging in:

$$\Omega = -B^{-1} = A^{-1}.$$

Then quantity $-B$ is called the *Fisher information matrix* and:

$$D_\theta \log q_\theta(Y_i|X_i) = i^{th} \text{ score}.$$

THE EFFICIENCY of MLE is a much cited result. The Cramer-Rao Lower Bound argument is that:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, -B^{-1})$$

when this is true, the ML is often "efficient." Azeem isn't a of the Cramer-Rao argument. He's a fan of the convolution argument, but this is hard to show.

E.g.: Previous limiting distribution may not always hold. Suppose $X_1, \ldots X_n \overset{i.i.d.}{\sim} \mathcal{U}(0, \theta_0)$. Then we have that:

$$\hat{\theta}_n = \max_{1 \leq i \leq n} X_i$$

and here:[137] [138]

$$-n(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{E}(\theta_0).$$

Then we need to show that for all t:

$$P\{-n(\hat{\theta}_n - \theta_0) \leq t\} \to F(t).$$

If $t < 0$ then we're done because $0 \to 0$. For $t \geq 0$ we have that:

$$\begin{aligned} P\{-n(\hat{\theta}_n - \theta_0) \leq t\} &= P\left\{\hat{\theta}_n \geq \theta_0 - \frac{t}{n}\right\} \\ &= 1 - P\left\{\hat{\theta}_n < \theta_0 - \frac{t}{n}\right\} \\ &= 1 - P\left\{X_i < \theta_0 - \frac{t}{n}\right\}^n \\ &= 1 - \left(\frac{\theta_0 - \frac{t}{n}}{\theta_0}\right)^n \\ &= 1 - \left(1 - \frac{\frac{t}{\theta_0}}{n}\right)^n \\ &\to 1 - \exp\left(-\frac{t}{\theta_0}\right) \\ &= F(t). \end{aligned}$$

[137] The C.D.F. of an exponential is:

$$F(t) = \begin{cases} 0, & \text{if } t < 0 \\ 1 - \exp(-t/\theta_0), & \text{if } t \geq 0. \end{cases}$$

[138] *Useful fact:*

$$\lim_{n \to \infty} \left(1 - \frac{c}{n}\right)^n \to \exp(-c).$$

Inference

THERE are three approaches to inference with MLE.[139] Here we'll consider tests of the form:

$$H_0 : f(\theta_0) = 0 \text{ vs. } H_A : f(\theta_0) \neq 0$$

for $f : \mathbb{R}^k \to \mathbb{R}^p$ where $D_\theta f(\theta_0)$, $(p \times k)$, has rank p.

Wald tests

BY simply applying the Delta Method we get that:

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta_0)) \xrightarrow{d} \mathcal{N}(0, D_\theta f(\theta_0)\Omega D_\theta f(\theta_0)')$$

and we can construct a χ^2_p test but need consistent estimate of variance. To do this we'll just estimate $D_\theta f(\theta_0)$ with $D_\theta f(\hat{\theta}_n)$ and to estimate $\Omega = -B^{-1} = A^{-1}$ we could use:

$$\hat{\Omega}_n = -\left(\frac{1}{n}\sum_{i=1}^n D^2_{\theta,\theta} \log q_{\hat{\theta}_n}(Y_i|X_i)\right)^{-1}$$

or:

$$\hat{\Omega}_n = \left(\frac{1}{n}\sum_{i=1}^n D_\theta \log q_{\hat{\theta}_n}(Y_i|X_i) D_{\theta'} \log q_{\hat{\theta}_n}(Y_i|X_i)\right)^{-1}.$$

Score tests

GIVEN the same testing problem define:[140] [141]

$$\tilde{\theta}_n \in \arg \max_{\theta \in \Theta,\ f(\theta)=0} L_n(\theta)$$

and the hope is that if the null is true then $\tilde{\theta}_n \xrightarrow{p} \theta_0$. The idea is that under the null $D_\theta L_n(\tilde{\theta}_n) \approx 0$. A sketch of formalization would be that by the MVT:[142]

$$D_\theta L_n(\tilde{\theta}_n) = D_\theta L_n(\theta_0) + H_n(\tilde{\theta}_n - \theta_0)$$

where H_n has j^{th} row given by the j^{th} row of $D^2_{\theta,\theta'} L_n(\theta)$ evaluated at $\theta = \tilde{\theta}_{n,j}$ between $\tilde{\theta}_n$ and θ_0. If $\tilde{\theta}_n \xrightarrow{p} \theta_0$ then $\tilde{\theta}_{n,j} \xrightarrow{p} \theta_0$, so by arguing as before that:

$$H_n \xrightarrow{p} E[D^2_{\theta,\theta} \log q_{\theta_0}(Y_i|X_i)] = -B$$

we can use the MVT again to get:

$$f(\tilde{\theta}_n) = f(\theta_0) + F_n(\tilde{\theta}_n - \theta_0)$$

[139] The trinity!

[140] Score!

[141] This is also called a Lagrange multiplier test, a characterization that Azeem eschews for some pedantic reason.

[142] Mean value theorem!

where F_n has j^{th} row $= j^{th}$ row of $D_\theta f(\theta)$ evaluated at $\theta = \theta^*_{n,j}$ which is between $\tilde{\theta}_n$ and θ_0. Then the CMT[143] gives us that:

$$F_n \xrightarrow{p} D_\theta f(\theta_0).$$

Then under the null:

$$F_n(\tilde{\theta}_n - \theta_0) = 0$$

which gives us that:

$$\underbrace{F_n H_n^{-1}}_{\text{use } D_\theta f(\tilde{\theta}_n)\hat{B}^{-1}} \sqrt{n} D_\theta L_n(\tilde{\theta}_n) = \underbrace{F_n H_n^{-1}}_{\xrightarrow{p} D_\theta f(\theta_0) B^{-1}} \underbrace{\sqrt{n} D_\theta L_n(\theta_0)}_{\xrightarrow{d} \mathcal{N}(0,A)} + o_P(1)$$

where $\hat{B} = \frac{1}{n}\sum_{i=1}^n D'_{\theta,\theta} \log q_{\tilde{\theta}_n}(Y_i|X_i)$. Then combining the limiting properties of the RHS we get that:

$$RHS \xrightarrow{d} \mathcal{N}(0, D_\theta f(\theta_0) B^{-1} A B^{-1} D_\theta f(\theta_0)')$$

where $B^{-1}AB^{-1} = A^{-1}$ or $-B^{-1}$.

Likelihood ratio test

Given the same setup as above with $\tilde{\theta}_n$ is as before. Then under the null:

$$\frac{l_n(\hat{\theta}_n)}{l_n(\tilde{\theta}_n)} = \frac{\prod_{i=1}^n q_{\hat{\theta}_n}(Y_i|X_i)}{\prod_{i=1}^n q_{\tilde{\theta}_n}(Y_i|X_i)}$$

shouldn't be too big. Equivalently:

$$L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n) = 0.$$

Then using arguments similar to those shown above:

$$2(L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n)) \xrightarrow{d} \chi^2_p.$$

Big picture what to use when?

1. Wald test is not invariant to parameterization.

2. Score test's advantage is that it's

3. LRT suffers less from the invariance issue and in simple problems[144] it's the most powerful.[145]

[143] Country Music Television? jk. nr.

[144] E.g., testing $\theta_0 = \theta^*$ vs. $\theta_0 = \theta^{**}$.

[145] Neyman-Pearson Lemma.

Part II

Empirical Analysis 31100—Harald Uhlig

Measure Theory

Measure Spaces

> Topology is about *open* sets. The characterizing property of a *continuous* function f is that the inverse image $f^{-1}(G)$ of an open set G is open. Measure theory is about *measurable* sets. The characterizing property of a *measurable* function f is that the inverse image $f^{-1}(A)$ of any measurable set is measurable.
>
> In topology, one axiomatizes the notion of 'open set', insisting in particular that the union of *any* collection of open sets is open, and that the intersection of a *finite* collection of open sets is open.
>
> In measure theory, on axiomatizes the notion of 'measurable set', insisting that the union of a *countable* collection of measurable sets is measurable, and that the intersection of a *countable* collection of measurable sets is also measurable. Also, the complement of a measurable set must be measurable, and the whole space must be measurable. Thus the measurable sets form a σ-algebra, a structure stable (or 'closed') under countably many set operations. Without the insistence that 'only countable many operations are allowed', measure theory would be self-contradictory—a point lost on certain philosophers of probability.
>
> —Williams (1991)

A measure space is a composed of a triple, $(\Omega, \mathcal{F}, \mu)$, which consists of:

1. Ω: A set of points which are different states of nature, ω.
2. $\mathcal{F}$: A set of subsets called that you can think of as events of Ω,[146] that form a $\sigma-$algebra:
   (a) $\Omega \in \mathcal{F}$
   (b) If $A \in \mathcal{F}$, then so is its complement, $A^c = \Omega \setminus A \in \mathcal{F}$.
   (c) $A_j \in \mathcal{F},\ j = 1, 2, \ldots$ implies $\cup_{j=1}^{\infty} A_j \in \mathcal{F}$
3. A measure μ that is a mapping, $\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$ with:[147]
   (a) Positivity: $\mu(A) \geq 0$.

[146] The idea is that Ω has all the possible combination of states of nature and different combinations of those can occur. Those are called events and are just subsets of Ω.

[147] Is every set measurable?

E.g. (Non-measurable set): Pick $\Omega = [0,\ 1]$. Also, define equivalence, $x \sim y$ for $x, y \in [0,\ 1]$ if $x - y \in \mathbb{Q}$. Then for any $x \in [0,\ 1]$ there is $A_x = \{y :\ x \sim y\}$, e.g., $A_0 = \mathbb{Q} \cap [0,\ 1]$. Then each A_x has countably many numbers. Furthermore, pick the set B s.t. $[0,\ 1] = \cup_{x \in B} A_x$. Then B has uncountably many elements in it because we've used it to construct the reals from the rationals. Then B is not measurable.

(b) $\sigma-$additivity: If $A_j \in \mathcal{F},\ j = 1, 2, \ldots$ are disjoint, then:

$$\mu\left(\cup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

(c) $\mu(\emptyset) = 0$

4. Probability space or probability measure: $\mu(\Omega) = 1$.[148]

Geometrically, you can think of this triple as some amorphous space, Ω. Within Ω there's points ω that can be collected into subsets, $\mathcal{F}$. The measure, μ, maps the events, $\mathcal{F}$, into the real number line. Discrete examples also help to convey the setup.

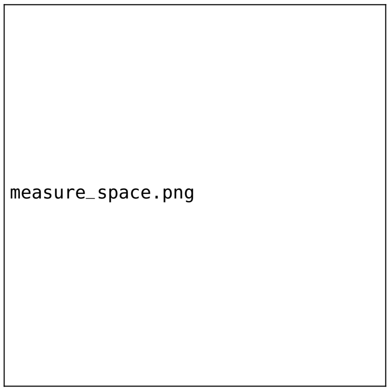

Figure 20: An example of a measure space, Ω, with events, A_1 and A_2, and a measure, μ that maps events to the real number line.

E.g. (Flipping two coins): Suppose you're flipping two coins at random. Then the possible states of nature are:

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Some examples of sets of the subsets of Ω, $\mathcal{F}$, that meet the criteria of a $\sigma-$algebra are:

$$\mathcal{F}_0 = \{\{\emptyset\}, \{\Omega\}\} = \{\{\emptyset\}, \{(H, H), (H, T), (T, H), (T, T)\}\}$$

and:

$$\mathcal{F}_1 = \{\{\emptyset\}, \{(H, H), (H, T)\}, \{(T, H), (T, T)\}, \{\Omega\}\}$$

and:

$$\begin{aligned} \mathcal{F}_2 = 2^{\Omega} &= \{\{\emptyset\}, \{(H, H), (H, T)\}, \{(H, H), (T, H)\}, \{(H, H), (T, T)\}, \{(H, H), (H, T), (T, H)\}, \ldots, \{\Omega\}\} \\ &= \{A_x \times \{H, T\} :\ A_x \subseteq \{H, T\}\} \end{aligned}$$

Are we sure these are $\sigma-$algebras? Well Ω is in each $\mathcal{F}$, if there's something in $\mathcal{F}$ then so is its complement, and the subsets implies that the union of subsets is in $\mathcal{F}$. One measure we could use would be:

$$\mu(A) = \sum_{\omega \in A} \frac{1}{2}$$

which is a probability measure.

E.g. (Rolling two dice): Suppose you're rolling two dice. Then:

$$\Omega = \{\omega = (x, y) :\ x, y\ \in \{1, \ldots, 6\}\}$$

and then three $\sigma-$algebras would be:

1. $\mathcal{F}_0 = \{\emptyset, \Omega\}$
2. $\mathcal{F}_1 = \{A_x \times \{1, \ldots, 6\} :\ A_x \subseteq \{1, \ldots, 6\}\}$
3. $\mathcal{F}_2 = \{A \subseteq \Omega\}$

[148] Not all measures have to be a probability measure, but a probability measure ought to satisfy this criteria.

where $\mathcal{F}_1$ is just the set where the first element is first 1 and then that's crossed with every possible outcome for the second die, second the first element is 2 and then that's crossed with every possible outcome for the second die, etc., etc. The probability measure would be:

$$\mu(A) = \sum_{\omega \in A} \frac{1}{36}.$$

Notice that here each subsequent $\sigma-$algebra is contained in one another. That is:

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$$

which we call a *filtration*. You can think of each $\sigma-$algebra being indexed by time, t, $\mathcal{F}_t$, where $\mathcal{F}_t$ is the set of events "known" or the information at t. The idea here is that $\mathcal{F}_0$ is the information you have regarding states of the world when you've rolled neither dice, etc., etc.[149]

E.g. (Infinite sequence): A more mathematical example would be an the natural numbers. Here we'd define the triple according to:

1. $\Omega = \mathbb{N}$

2. $\mathcal{F} = \{A \subseteq \Omega\}$

3. Let $\alpha_j = \mu(\{j\})$. Then $\mu(A) = \sum_{j \in A} \alpha_j$.

E.g. (Lebesgue measure): The Lebesgue measure, which is just a generalization of the Uniform distribution ($\mathcal{U}$) is defined by the following triple:

1. $\Omega = \mathbb{R}^m$

2. $\mathcal{F} = \mathcal{B}(\Omega)$: The Borel-$\sigma$ algebra, i.e., the smallest $\sigma-$algebra, which contains all open subsets of Ω.

3. Let $I_j = [a_j, b_j]$, $a_j \leq b_j \in \mathbb{R}$ be intervals. Define the box:

   $$B = I_1 \times \cdots \times I_n.$$

   and define:

   $$\mu(B) = (b_1 - a_1)(b_2 - a_2) \ldots (b_m - a_m).$$

4. Can define Lebesgue-measurable sets. See lecture notes if interested.

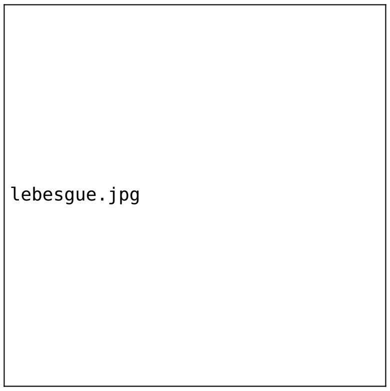

Figure 21: Henri Lebesgue wearing an early prototype of Morpheus's sunglasses in the Matrix.

[149] Recall Lars's notation for information available at time t in a conditional expectation statement. #martingale

Integration

ONCE you have measures, you can define the sense in which you integrate over a measure. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then:

1. A function $f : \Omega \to \mathbb{R}^k$ is called $\mathcal{F}-$measurable if $f^{-1}(B) \in \mathcal{F}$ for ever Borel set $B \in \mathcal{B}(\mathbb{R}^k)$.

2. Suppose $f = I\{\omega \in A\}$. Then define the integral:

$$\int f d\mu = \int f(\omega)\mu(d\omega) = \mu(A).$$

3. Suppose f is a linear combination of indicator functions:

$$f(\omega) = \sum_{j=1}^{n} \psi_j I\{\omega \in A_j\},\ A_j \in \mathcal{F}.$$

Then define the integral per line extension:[150]

$$\int f d\mu = \sum_{j=1}^{n} \psi_j \int I\{\omega \in A_j\} d\mu = \sum_{j=1}^{n} \psi_j \mu(A_j).$$

Additionally we can extend this to all positive measurable functions and can then extend to all measurable functions, f, with:

$$\int f d\mu = \int \max\{f, 0\} d\mu - \int \max\{-f, 0\} d\mu$$

provided at least one of the integrals is finite. For $A \in \mathcal{F}$ define:

$$\int_A f d\mu = \int I\{\omega \in A\} f(\omega)\mu(d\omega).$$

If μ is a probability measure then we can define the expectation of f as:

$$E[f] = \int f d\mu.$$

Thm. (Radon-Nikodym Theorem): *Let $\mathcal{F}$ be a $\sigma-$algebra on Ω. Let μ and ν be two measures on $\mathcal{F}$. Suppose that $\mu(\Omega) < \infty$ and $\nu(\Omega) < \infty$. Suppose that ν is absolutely continuous with respect to μ.*[151,152] *I.e., $\nu << \mu$. Then there exists a positive measurable function g, called the Radon-Nikodym derivative:*[153]

$$g : \Omega \to \mathbb{R}_+ \ \text{or}\ g = \frac{d\nu}{d\mu} \ \text{with}\ \nu(A) = \int_A g d\mu = \int_A \frac{d\nu}{d\mu} d\mu$$

for all $A \in \mathcal{F}$.[154]

To see the value of the Radon-Nikodym Theorem we'll work through three examples:

[150] An example could be: $\psi_1 = 1,\ \psi_2 = 3.5$. Then:

$$f = \sum_{j=1}^{2} \psi_j I\{\omega \in A_1\}$$

and:

$$\int f d\mu = \mu(A_1) + 3.5\mu(A_1) = 4.5\mu(A_1)$$

which is a result we get for both approaches to integrating.

[151] I.e., $\mu(A) = 0$ implies $\nu(A) = 0$ for $A \in \mathcal{F}$, which we write as $\nu << \mu$.

[152] Here's a theorem on this topic:

Thm.: *Suppose there is a function g so that:*

$$\nu(A) = \int_A g d\mu$$

for all $A \in \mathcal{F}$. Then ν is absolutely continuous w/r/t μ, i.e., $\nu << \mu$.

[153] If you have an expression like:

$$\int_A d\nu = \int_A \frac{d\nu}{d\mu} d\mu$$

you should note that you can't cancel out the $d\mu$s because the $\frac{d\nu}{d\mu}$ is a function, not a ratio and the $d\mu$ is the thing you're measuring over, not some small change. The point is just that you can't use typical operations on this stuff and that seems to be because it's pretty bad notation.

E.g. (Conditional expectations with rolling two dice): Pick $\Omega = \{\omega = (x,y) : \; x,y \in \{1,\ldots,6\}\}$, $\mathcal{F} = \mathcal{F}_j$ for $j = 0,1,2$, $\mu(A) = \sum_{\omega \in A} \frac{1}{36}$. Also, let $f : \; \Omega \to \mathbb{R}$ be measurable:

$$\int f d\mu = \sum_{\omega \in \Omega} \frac{f(\omega)}{36}.$$

Furthermore, let f be $\mathcal{F}_2$-measurable.[155] To get a sense of conditional expectations in this framework we want to find a $\mathcal{F}_1$-measurable function $g : \Omega \to \mathbb{R}$, $gE[g|\mathcal{F}_1] = E[f|\mathcal{F}_1] = E_1[f]$ because we construct g s.t.:

$$\int_A g d\mu = \int_A f d\mu, \quad \text{for all} \quad A \in \mathcal{F}_1$$

and we're certain that g exists due to the Rad.-Nik. Thm. for signed measures:[156]

$$g(x,y) = E_1[f(x,y)] = E[f(x,y)|\mathcal{F}_1] = E[f(x,y)|x] = \sum_{j=1}^{6} \frac{1}{6} f(x,j).$$

E.g. (Integration as summation): If we revisit the infinite sequence example above, which had:

$$(\Omega, \mathcal{F}, \mu) = \left(\mathbb{N}, \{A \subseteq \Omega\}, \mu(A) = \sum_{j \in A} \alpha_j \right)$$

for $\alpha_j \geq 0$. Then for $f : \Omega \to \mathbb{R}$:

$$\int f d\mu = \sum_{j=1}^{\infty} \alpha_j f(j)$$

where $\alpha_j = \mu(\{j\})$.

E.g. (Lebesgue measure revisited): For $(\Omega, \mathcal{F}, \mu) = (\text{An open subset of some } \mathbb{R}^n, \mathcal{B}(\Omega), \text{ Lebesgue measure })$ then for a measurable function, $f : \Omega \to \mathbb{R}$, the integral over f with respect to μ is what we expect. E.g., for:

$$f(\omega) = \kappa I\{\omega \in B\}$$

then:

$$\int f d\mu = \kappa(b_1 - a_1)(b_2 - a_2) \ldots (b_n - a_n).$$

[155] $\mathcal{F}_2$ is the power set, so it includes every open set in Ω and the inverse of any open set is an open set, so all functions, f, are $\mathcal{F}_2$-measurable.

[156] Why are we certain? Define $\nu(A)$ for $A \in \mathcal{F}_1$ per:

$$\nu(A) = \int_A f(x,y) d\mu$$

and define $\tilde{\mu}$ on $\mathcal{F}_1$ per: $\tilde{\mu}(A) = \mu(A), \; \forall A \in \mathcal{F}_1$. Then $\nu << \tilde{\mu}$ and then we can apply Rad.-Nik. which implies that there is a $\mathcal{F}_1$-measurable function g so that:

$$\nu(A) = \int g d\tilde{\mu} = \int g d\mu.$$

Maximum Likelihood Estimation

Framework

Idea is that there's some unknown parameter, $\theta \in \Theta$, with measure $\mu(d\theta)$.[157] We observe $y \in Y$ with measure $\nu(dy)$ and probability density $f(y|\theta)$ with respect to the measure ν. That is:

$$\int f(y|\theta)\nu(dy) = 1 \ \forall \ \theta \in \Theta.$$

Conceptually we think about conducting an experiment on θ which leads to an observation $y \sim f(y|\theta)$ for an assumed density, $f(\cdot)$.[158]

Def. (Likelihood function): The likelihood function, $L(\cdot)$ is $L(\theta|y) := f(y|\theta)$.

Def. (Log-likelihood function): The log-likelihood function, $\ell(\cdot)$ is $\ell(\theta|y) := \log L(\theta|y)$.

E.g. (Linear Regression): Suppose $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times k}$, $\beta \in \mathbb{R}^k$, $\Sigma \in \mathbb{R}^{n\times n}$ where Σ is positive definite with:

$$y = X\beta + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, \Sigma)$. Solving for ϵ yields:

$$\epsilon = y - X\beta \sim \mathcal{N}(0, \Sigma).$$

The vector of parameters to identify (θ) could be just β or might be both β, Σ, etc., etc. The key assumption is that X does not depend on Θ, which gives us the following conditional likelihood:

$$L(\theta|y, X) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta)\right\}$$

and we get a nice result if you assume $\Sigma = \sigma^2 I_n$:[159]

$$L(\theta|y, X) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2\right\}.$$

[157] Often $\Theta \subseteq \mathbb{R}^m$.

[158] Sometimes the distinction between conditional and unconditional likelihoods is drawn. We typically think about conditional likelihoods when there is some set of covariates x that do not depend on θ, giving us:

$$f(x, y|\theta) = f(y|\theta, x)f(x)$$

which is propoertional to $f(y|theta, x)$.

[159] Which you get if the errors are i.i.d. and homoskedastic.

E.g. (Logit and Probit): Suppose $y \in \{0, 1\}$ according to:

$$y = \begin{cases} 1, & \text{if } \epsilon \leq X\beta \\ 0, & \text{if } \epsilon > X\beta \end{cases}$$

then the strategy that both Logit and Probits will take to writing down a likelihood is to partition the two states of the world (when the outcome equals 1 or 0). Pick $\theta = \beta$. Then if $y = 1$ we have the following likelihood:

$$L(\theta | y = 1, X) = G(X\beta)$$

where G is the CDF of the error term, ϵ. For $y = 1$ the likelihood is just $1 - G$. Then the Probit picks $\epsilon \sim \mathcal{N}(0, 1)$ and the Logit picks:

$$G(v) = \frac{e^v}{1 + e^v}.$$

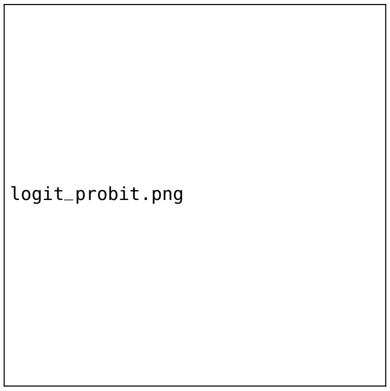

Figure 22: The CDF of a logit and probit CDF are plotted. Additionally a Normal CDF with a slightly higher variance is plotted to show how to transform a Probit CDF to a Logit CDF.

MLE, Score, and Information Matrix

The maximum likelihood estimator (MLE) is:

$$\hat{\theta} := \arg\max_{\theta} L(\theta | y)$$

and to analytically solve for this estimator we probably want to think about first-order conditions. This sort of approach turns out to be so important that the vector it yields has a name: The Score.

Def. (Score): For $\theta \in \Theta \subseteq \mathbb{R}^m$, an open set, we define the score as the first derivative of the log-likelihood function:

$$s(\theta) = s(\theta | y) = \frac{\partial \ell(\theta | y)}{\partial \theta}$$

which is a column vector that's as long as the number of parameters in θ.

Thm.: $E_{\theta_0}[s(\theta_0 | y)] = 0$ *where θ_0 is the true value of the parameter.*[160]

Proof. Densities must integrate to 1, so we can write:

$$\forall \theta, \int f(y | \theta) \nu(dy) = 1$$

differentiating both sides w/r/t θ yields:

$$\int \frac{\partial f(y | \theta)}{\partial \theta} \nu(dy) = 0$$

which holds for all θ. Then we can multiply and divide within the integral by $f(y | \theta_0)$ to get:[161]

[160] The sub-θ_0 on the expectation is to indicate that you're integrating with respect to the density at θ_0.

[161] Because:

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial L / \partial \theta}{L} = \frac{\partial f / \partial \theta}{f}.$$

$$0 = \int \frac{\partial \ell(\theta|y)}{\partial \theta} f(y|\theta_0)\nu(dy) = \int s(\theta_0|y) f(y|\theta_0)\nu(dy) = E_{\theta_0}[s(\theta_0|y)]$$

which gives us the desired identity. □

THE score function helps us identify $\hat{\theta}$ but to go further we'll need a sense of the variance of our parameter. Taking the second derivative proves useful here.

Def. (Information Matrix): The Information matrix is defined by :

$$\mathcal{I}(\theta) := E[s(\theta|y)s(\theta|y)'].$$

Thm.: *Assuming certain regularity conditions:*

$$\mathcal{I}(\theta_0) = E_{\theta_0}[s(\theta_0|y)s(\theta_0|y)'] = -E_{\theta_0}\left[\frac{\partial^2 \ell(\theta_0|y)}{\partial\theta\partial\theta'}\right].$$

Proof. Starting from:

$$E_{\theta_0}[s(\theta|y)] = 0$$

write out the integral form of the expectation and then differentiate both sides w/r/t θ'. Then we get:

$$0 = \int \frac{\partial^2 \ell(\theta|y)}{\partial\theta\partial\theta'} f(y|\theta)\nu(dy) + \int \frac{\partial \ell(\theta|y)}{\partial\theta} \frac{\partial f(y|\theta)/\partial\theta'}{f(y|\theta)} f(y|\theta)\nu(dy)$$

which we can re-write as:

$$0 = E\left[\frac{\partial^2 \ell(\theta|y)}{\partial\theta\partial\theta'}\right] + \mathcal{I}(\theta).$$

□

ONE OTHER useful tool for working with likelihoods is taking first and second order expansions around some value of θ. Taylor's Theorem gives us:

$$\ell(\tilde{\theta}) \approx \underbrace{\ell(\theta) + s(\theta)'(\tilde{\theta} - \theta)}_{\text{Some constant, } A} - \frac{1}{2}(\tilde{\theta} - \theta)' \underbrace{\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}}_{\approx -\ell(\theta)}(\tilde{\theta} - \theta)$$

so then the likelihood is:

$$L(\theta) \approx A \exp\left\{-\frac{1}{2}(\tilde{\theta} - \theta)'\ell(\theta)(\tilde{\theta} - \theta)\right\}$$

which is approximately proportional to the normal density. Pretty awesome. Similarly, we can take a first-order expansion around $s(\tilde{\theta})$:

$$s(\tilde{\theta}) \approx s(\theta) + \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}(\tilde{\theta} - \theta).$$

Pick $\theta = \theta_0$ for the value to perform the expansion around. Then:

$$E_{\theta_0}[s(\tilde{\theta})] \approx E_{\theta_0}[s(\theta_0)] - \mathcal{I}(\theta_0)(\tilde{\theta} - \theta_0) = -\mathcal{I}(\theta_0)$$

and the score function similarly drops out when taking expectations of the likelihood.

Asymptotics

To DEVELOP the asymptotic properties of likelihood estimation we'll first have to adjust our notation a bit. Define the following:

$$\ell_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell(\theta|y_i)$$
$$s_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell(\theta|y_i)}{\partial\theta}$$
$$H_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\theta|y_i)}{\partial\theta\partial\theta'}.$$

Then noting that the CLT gives us:

$$\sqrt{n}s_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, E[s(\theta_0)s(\theta_0)']) = \mathcal{N}(0, \mathcal{I}(\theta_0))$$

and that the MLE, $\hat{\theta}_n$ solves $s_n(\hat{\theta}_n) = 0$ we can take a first-order expansion around θ_0:

$$s_n(\hat{\theta}_n) = 0 \approx s_n(\hat{\theta}_n) \approx s_n(\theta_0) + H_n(\theta_0)(\hat{\theta}_n - \theta_0)$$

which we can rearrange and pre-multiply by $\sqrt{n}$ and $-\mathcal{I}(\theta_0)^{-1}$ to get:

$$\sqrt{n}\mathcal{I}(\theta_0)^{-1}H_n(\theta_0)(\hat{\theta}_n - \theta_0) \to \sqrt{n}(\hat{\theta}_n - \theta_0) \approx \sqrt{n}\mathcal{I}(\theta_0)^{-1}s_n(\theta_0)$$

so using the CLT we invoked above we get:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}\mathcal{I}(\theta_0)\mathcal{I}(\theta_0)^{-1}) = \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}.$$

Thm.: *If $\mathcal{I}(\theta)$ is invertible at the true θ then:*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1}).$$

Proof. See immediately above. □

A CLASSIC result for MLEs is the following theorem.

Thm. (Cramer-Rao Lower Bound): *For a statistic, T, that's a function, $T : y \in Y \to \mathbb{R}^k$, if $Var_\theta(T(y)) < \infty$ and $\mathcal{I}(\theta)$ is invertible then:*

$$Var_\theta(T(y)) \geq \left(\frac{\partial\psi(\theta)}{\partial\theta}\right)\mathcal{I}(\theta)^{-1}\left(\frac{\partial\psi(\theta)}{\partial\theta}\right)'$$

for $\psi(\theta) = E_\theta[T(y)]$.[162]

[162] In words, this is just saying that for an unbiased estimator of θ, $T(y)$, then the variance of that estimator is bounded below by the asymptotic variance of the MLE.

As we've done many times, now that we have a limiting distribution we have to figure out the sample analog of the variance-covariance matrix. In this context, first recall the following two equations:

$$\mathcal{I}(\theta) = E[s(\theta|y)s(\theta|y)'] = -E\left[\frac{\partial^2 \ell(\theta|y)}{\partial\theta\partial\theta'}\right]$$

which gives us two ways to construct a sample analog:

$$\hat{\mathcal{I}}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} s(\hat{\theta}_n|y_i)s(\hat{\theta}_n|y_i)'$$

and:

$$\hat{\mathcal{I}}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell(\hat{\theta}_n|y_i)}{\partial\hat{\theta}_n\partial\hat{\theta}_n'}$$

then if $\hat{\theta}_n \xrightarrow{p} \theta$ we have two consistent estimators for the information matrix.

Three hypothesis tests

. . .

Extremum Estimators

One way to generalize the types of estimators we've looked at so far is to classify them as extremum estimators. We call an estimator, $\hat{\theta}_n$, an extremum estimator if it solves:

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta \subseteq \mathbb{R}^m} Q_n(\theta)$$

for some objective function $Q_n(\theta)$. Examples of this type of estimator include:

1. MLE: $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta|Y_i)$.
2. NLLS:[163] $Q_n = -\frac{1}{n}\sum_{i=1}^{n}(Y_i - h(X_i;\theta))^2$.
3. M:[164] $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(Y_i;\theta)$.
4. GMM: $Q_n(\theta) = -\frac{1}{2}g_n(\theta)'\hat{W}_n g_n(\theta)$, for $g_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Y_i;\theta)$.

Thm. (Consistency of Extremum Estimators): *Suppose Q_n, $n = 1, 2, \ldots$ are continuous functions of $\theta \in \Theta$ where Θ is a compact set and also suppose that we have: (identification) $\theta_0 = \arg\max_{\theta \in \Theta} Q_0(\theta)$ is unique, (uniform convergence) $Q_n(\cdot)$ converges uniformly in probability to $Q_0(\cdot)$, i.e.:*

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0.$$

Then $\hat{\theta}_n \xrightarrow{p} \theta$.

M-Estimators

Working with M-Estimators is nearly identical to working with NLLS and we know how to do anything we want with MLEs, so in this section we just work through the conditions of M-Estimators.[165] For the M-Estimators we assume twice-differentiability, where if the truth is $\theta = \theta_0$, then we define the score:

$$s(Y_i;\theta) = \frac{\partial m(Y_i;\theta)}{\partial \theta}$$

[163] Non-linear least squares.

[164] Moment.

[165] We'll dive into GMM with different tools in a bit.

and the Hessian as:

$$H(Y_i;\theta) = \frac{\partial^2 m(Y_i;\theta)}{\partial\theta\partial\theta'}$$

then similar to MLE we get the following nice results:

$$Q_n(\theta) = \frac{1}{n}\sum_i m(Y_i;\theta) \xrightarrow{p} E_{\theta_0}[m(Y_i;\theta)] =: Q_0(\theta)$$

$$s_n(\theta) = \frac{\partial Q_n(\theta)}{\partial\theta} = \frac{1}{n}\sum_i s(Y_i;\theta) \xrightarrow{p} E_{\theta_0}[s(Y_i;\theta)] = 0$$

$$H_n(\theta) = \frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_i H(Y_i;\theta) \xrightarrow{p} E_{\theta_0}[H(Y_i;\theta)] =: \Psi \text{ at } \theta = \theta_0$$

And we can also assume that Y_i are correlated across "nearby" i, which is related to ergodicity.[166] Then:

$$\sqrt{n}s_n(\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

where the long-run variance of Σ of $s(Y_i;\theta_0)$ is:[167]

$$\Sigma = \sum_{k=-\infty}^{\infty}\Gamma_k$$

where $\Gamma_k = E[s(Y_i;\theta_0)s(Y_{j+k};\theta_0)']$.[168]

E.g.: Suppose $E[s_i s_{i+k}] = 0,\ \forall\, k \geq 3$. Then:

$$\Sigma = \frac{1}{n}\sum_{k=-2}^{2}\left(1 - \frac{|k|}{n}\right)\Gamma_k \approx \frac{1}{n}\sum_{k=-2}^{2}\Gamma_k.$$

E.g.: Suppose $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and:

$$s_i = (\epsilon_i,\ \epsilon_{i-1},\ \epsilon_{i-2})'$$

then if we take:

$$E[s_i s_i'] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \Gamma_0$$

because the expectation of a squared $\mathcal{N}(0,1)$ is 1. Next we take one more iteration forward:

$$E[s_i s_{i+1}'] = E\left[\begin{bmatrix} \epsilon_i \\ \epsilon_{i-1} \\ \epsilon_{i-2} \end{bmatrix}\begin{bmatrix} \epsilon_i & \epsilon_{i-1} & \epsilon_{i-2} \end{bmatrix}\right] = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \Gamma_1$$

and by the same logic:

$$\Gamma_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

[166] The concept of ergodicity that Uhlig is using here is that the past correlations eventually are overwhelmed by the asymptotics.

[167] So if we have a scalar score function where each observation is $s_i(Y_i;\theta)$:

$$E\left[\left(\frac{1}{n}\sum_i s_i\right)^2\right] = \frac{1}{n^2}\sum_i E[s_i(Y_i;\theta)] = \frac{1}{n}E[s_i(Y_i;\theta)^2]$$

but if not i.i.d. then it's not as easy:

$$E\left[\left(\frac{1}{n}\sum_i s_i\right)^2\right] = \frac{1}{n^2}\sum_i\sum_j E[s_i(Y_i;\theta)s_j(Y_j;\theta)'].$$

[168] Just as we did in Lars's class last year.

and:

$$\Gamma_3 = 0.$$

If we then did Γ_{-1}, Γ_{-2} we'd get that:

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

BUT WHAT ABOUT the estimator part of M-Estimators? Well, the estimator, $\hat{\theta}_n$ solves:

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = s_n(\hat{\theta}_n) = 0$$

and if we take a first-order expansion around θ_0, we get:

$$0 = s_n(\hat{\theta}_n) \approx s_n(\theta_0) + H_n(\theta_0)(\hat{\theta}_n - \theta_0)$$

and if we assume that ψ is invertible[169] then:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx -\sqrt{n}\Psi^{-1}H_n(\theta_0)(\hat{\theta}_n - \theta_0) \approx \sqrt{n}\Psi^{-1}s_n(\theta_0)$$

and if we take the limit we get the familiar looking result:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Psi^{-1}\Sigma\Psi^{-1}).$$

If $\Gamma = \Psi$ then this simplifies to:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Psi^{-1})$$

where $\Gamma = \Gamma(\theta_0)$ and $\Psi = \Psi(\theta_0)$.

GMM

FURTHER generalizing brings us to GMM. First we assume twice differentiability of $g(\cdot)$ and ergodicity as well. Furthermore, we assume that:

$$\hat{W}_n \xrightarrow{p} \mathcal{W}$$

and define:

$$\mathcal{S} := \sum_{k=-\infty}^{\infty} \Gamma_k$$

which is the long-run variance of $g(Y_i; \theta_0)$, where Γ_k is as we defined before:

$$\Gamma_k = E[g(Y_i; \theta_0)g(Y_{i+k}; \theta_0)']$$

and we lastly define:

$$G := E\left[\frac{\partial g(Y_i; \theta_0)}{\partial \theta'}\right].$$

[169] Which implies positive definiteness.

Then for our sample average of $g(\cdot)$:

$$g_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} g(Y_i;\theta)$$

gives us:

$$Q_n(\theta) := -\frac{1}{2} g_n(\theta)'\hat{W}_n g_n(\theta)$$

and:

$$G_n(\theta) := \frac{\partial g_n(\theta)}{\partial \theta'} = \frac{1}{n}\sum_i \frac{\partial g(Y_i;\theta)}{\partial \theta'} \xrightarrow{p} G \ \text{ at } \theta = \theta_0$$

and:

$$s_n(\theta) := \frac{\partial Q_n(\theta)}{\partial \theta} = -G_n(\theta)'\hat{W}_n g_n(\theta)$$

which gives us the asymptotics:

$$\sqrt{n} s_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\Sigma := G'\mathcal{W}\mathcal{S}\mathcal{W}G$.[170],[171] Then we can do the same expansion procedure we've done in the past. First note that $\hat{\theta}_n$ solves:

$$s_n(\hat{\theta}_n) = 0$$

so we can do a first-order expansion of $g_n(\theta)$ around θ_0:

$$0 = s_n(\hat{\theta}_n) = -G_n(\hat{\theta}_n)'\hat{W}_n g_n(\hat{\theta}_n) \approx -G_n(\hat{\theta}_n)'\hat{W}_n G_n(\hat{\theta}_n - \theta_0)$$

Note that:

$$G_n(\hat{\theta}_n)'\hat{W}_n G_n(\hat{\theta}_n) \xrightarrow{p} G'\mathcal{W}G =: \Psi$$

and if we assume that Ψ is invertible then:[172]

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \sqrt{n}\Psi^{-1} s_n(\theta_0)$$

and taking the limit gives us the asymptotics of the GMM estimator:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,\ \Psi^{-1}\Sigma\Psi^{-1})$$

and a good choice of $\mathcal{W}$? How about $\mathcal{W} = \mathcal{S}^{-1}$ because then the asymptotics reduces to:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,\ \Psi^{-1}).$$

Hypothesis testing with GMM isn't particularly unique. To see this, suppose we want to do constrained estimation as we did with MLEs. For:

$$\hat{\theta}_{c,n} = \arg\max_{\theta \in \Theta} Q_n \ \ s.t.\ a(\theta) = 0$$

where $\partial a(\theta_0)/\partial\theta$ has rank k. Then if $\Psi = \Sigma$ and $\hat{\Psi}_n \xrightarrow{p} \Psi$ we get the following tests:

[170] For $E[g(Y_i;\theta_0)] = 0$ we can also write:

$$\sqrt{n} g_n(Y_i;\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{S}).$$

[171] The sample analog of Σ is:

$$G_n'\hat{W}_n g_n g_n'\hat{W}_n G_n.$$

[172] Uhlig mentions that something's missing from this expansion that we dealt with with M-Estimators. Perhaps worth dwelling on.

1. "Likelihood" Ratio Test: $LR = 2n(Q_n(\hat{\theta}_n) - Q_n(\hat{\theta}_{c,n})) \xrightarrow{d} \chi^2_k$

2. Lagrange Multiplier Test: $LM = ns_n(\hat{\theta}_{c,n})'\hat{\Upsilon}_n^{-1}s_n(\hat{\theta}_{c,n})) \xrightarrow{d} \chi^2_k$

3. Wald: Define the object $A_n := \partial a(\hat{\theta}_n)/\partial\theta \xrightarrow{p} A = \partial a(\theta_0)/\partial$ then:

$$W = na(\hat{\theta}_n)'(A_n\hat{\Upsilon}_n^{-1}A_n')^{-1}a(\hat{\theta}_n) \xrightarrow{d} \chi^2_k$$

E.g. (IV as GMM): Suppose $Z_t,\ t = 1, \ldots, T$ are uncorrelated with ϵ_t. Then:

$$E[Z_t\epsilon_t] = E[Z_t(Y_t - X_t\beta)] = 0$$

is just our moment condition. We can generalize this using:

$$g([X_t, Z_t]; \theta) = E[Z_t' f(X_t; \theta)] = 0$$

and find $\hat{\theta}$ per GMM for some suitable weighting matrix, $\hat{W}_t$.

Bayesian Inference

Framework

The setup for Bayesian Inference goes as follows: There is an unknown parameter $\theta \in \Theta$ with measure $\mu(d\theta)$. We observe X with measure $\nu(dX)$ and X has the density $f(X|\theta)$ w/r/t ν which allows us to write down the likelihood function $L(\theta|X) = f(X|\theta)$. Conceptually we think about draws (experiment) on θ which leads to an observation $X \sim f(X|\theta)$ for some known $f(\cdot)$ if it is carried out.

Def. (Sufficient Statistic): A function or statistic T of X is sufficient if the distribution of X condition on $T(X)$ does not depend on θ.

E.g.: Suppose $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then a sufficient statistic for X is $T(X) = (\bar{X}_n, \hat{\sigma}_n^2)$.

Def. (Sufficiency Principle): Two observations X, Y which lead to the same value of a sufficient statistic, T, $T(X) = T(Y)$, lead to the same inference regarding θ.

Def. (Conditionality Principle): If two experiments can be carried out on θ and if exactly one of these experiments is actually carried out with some probability p, then the resulting inference on θ should only depend on the selected experiment and the resulting observation.[173]

Def. (Likelihood Principle): The information brought about by an observation X_i about θ is entirely contained in the likelihood function, $L(\theta|X_i)$ and if two observations X_1, X_2 lead to proportional likelihood functions:

$$L(\theta|X_1) = \alpha L(\theta|X_2),\ \alpha > 0$$

then they shall lead to the same inference regarding θ.

Inference under these principles can be done by maximizing the likelihood by choice of θ and then using an estimate of the information matrix to test hypotheses. A Bayesian would go a slightly

[173] Huh? The idea seems to just be that data you don't observe shouldn't impact your inference on θ which seems like a terrible assumption.

different route. They'd first define a prior, $\pi(\theta)$, which is a density w/r/t μ. Then the posterior would be:

$$\pi(\theta|X) = \frac{L(\theta|X)\pi(\theta)}{\int_{\Theta} L(\tilde{\theta}|X)\pi(\tilde{\theta})\mu(d\tilde{\theta})}.$$

How does this allow for us to do inference on θ? Well the idea is that $X \sim f(X|\theta_0)$ is given when the experiment is conducted and the true parameter, θ_0, is distributed according to $\pi(\theta_0|X)$, so the true parameter is a random variable. This is in stark contrast to the frequentist approach where θ_0 is unknown and the observation $X \sim f(X|\theta_0)$ is random.

Def. (Stopping Rule Principle): If a sequence of experiments is directed by a stopping rule, τ, that indicates when the experiments stop, then inference about θ shall depend on τ only through the resulting sample.

E.g.: Experimenter has 100 observations, $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$ with sample mean $\bar{X}_n = 0.2$. The frequentist wants to conduct the following test: $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$. Then there are the following stopping rules:

1. Stop always: If $\sqrt{100}\ \bar{X}_{100} > 1.96$ then reject.

2. If $\sqrt{100}\bar{X}_{100} \geq c$ then stop and reject. If not take another 100 draws and reject if: $\sqrt{200}\bar{X}_{200} \geq c$.

3. Now suppose that an RA is generating the data and shows up with probability p. As a classical econometrician you work out all the possibilities and make inference accordingly. Yikes.

Stopping rule principle avoids these issues.

E.g.: Suppose we have data, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}(\theta)$. To fix ideas suppose each observation is a realization of some casino game where θ is the probability that the gambler wins. Then define:

$$X^{(n)} := \sum_{i=1}^{n} X_i$$

which would give us the following likelihood:

$$L(\theta|X^{(n)}) = f(x^{(n)}|\theta, n)$$

where $f(\cdot)$ is the Binomial distribution for n draws. Now consider the following stopping rules:

1. Take 100 draws. None more.

2. Take draws until $X^{(n)} = \frac{n}{2}$ or $n = 1,000,000$ where the idea is that $1,000,000$ is the maximum number of draws you could observe.

3. Suppose $n = 100$ and $X^{(100)} = 50$. Then the stopping rule principle says that inference about θ does not depend on the stopping rule.

Clearly the second stopping rule is going to bias your estimate of θ and overweight the possibility that the game is fair.[174]

E.g.: Suppose $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$. Then a stopping rule like:

$$|\bar{X}_n| = \left| \frac{1}{n} \sum_i X_i \right| > \frac{1.96}{\sqrt{n}}$$

then classical inference that's done carelessly would always reject $H_0 : \theta = 0$ at the 5% level. The Bayesian approach doesn't suffer from these issues, though. Hmmmmmmmm.

To complete the Bayesian framework, we need two more ingredients: A decision $\delta(X) \in \mathcal{D}$ and a loss function, $\mathcal{L}(\theta, \delta(X))$.[175] Then risk from the frequentist's perspective is had with:

$$\mathcal{R}(\theta, \delta) := E_\theta[\mathcal{L}(\theta, \delta(X))] = \int_X \mathcal{L}(\theta, \delta(X)) f(X|\theta) dX$$

but from the Bayesian perspective, the posterior expected loss is:

$$\rho(\pi, \delta(X)) := E_\pi[\mathcal{L}(\theta, \delta(X))|X] = \int_\Theta \mathcal{L}(\theta, \delta(X)) \pi(\theta|X) d\theta$$

so the integrated risk is:

$$r(\theta, \delta) := E_\pi[\mathcal{R}(\theta, \delta)] = \int_\Theta \int_X \mathcal{L}(\theta, \delta(X)) f(X|\theta) \pi(\theta) dX d\theta = \int_X \rho(\theta, \delta(X)) m(X) dX$$

where $m(X) = \int_\Theta f(X|\theta)\pi(\theta) d\theta$.

Def. (Admissibility): An estimator δ_0 is *admissibile* if there is no estimator δ_1, which dominates δ_0. That is, there's no δ_1 that satisfies:

$$\mathcal{R}(\theta, \delta_0) \geq \mathcal{R}(\theta, \delta_1)$$

which holds with strictly for at least one value of θ_0.

Def. (Bayes Estimator): A Bayes estimator associated with a prior distribution π and a loss function $\mathcal{L}$ is any estimator δ^π which minimizes $r(\theta, \delta)$:

$$\delta^\pi \in \arg\min_{d \in \mathcal{D}} \rho(\pi, d|X).$$

Def. (Bayes Risk): The value $r(\pi) := r(\pi, \delta^\pi)$ is called the Bayes risk.

[174] That is, that $\theta_0 = 0.5$.

[175] E.g., Quadratic Loss: $\mathcal{L}(\theta, \delta(X)) = ||\theta - \delta(X)||^2$.

Thm.: *Bayes estimator δ^{π} is admissible given certain condition.*

Thm.: *All admissible estimators are limits of sequences of Bayes estimators.*

Thm.: *The MLE estimator is inadmissible and is dominated by the James-Stein estimator:*

$$\delta_{JS}(\hat{\theta}) = \left(1 - \frac{k-2}{||\hat{\theta}||^2}\right)\hat{\theta}$$

Conjugacy and Priors

A few words on this

Def.: If the prior π is a member of a parametric family of distributions, so that the posterior, $\pi(\theta|X)$, also belongs to that family, then the family is called conjugate to $\{f(\cdot|\theta) : \theta \in \Theta\}$.

E.g. (Flat Prior): Not invariant to reparamaterization.

E.g. (Jeffrey's Prior): Suppose $\theta \in \mathbb{R}$ and $\tilde{\theta} \in \mathbb{R}$ and the two parameters are related according to $h(\cdot)$: $\tilde{\theta} = h(\theta)$ and the function $h(\cdot)$ is one-to-one. Then we get the following two priors: $\tilde{\pi}(\tilde{\theta})$ and $\pi(\theta)$ and a desirable property for these priors is that they're invariant to reparameterization:

$$\int_{\tilde{A}} \tilde{\pi}(\tilde{\theta})d\tilde{\theta} = \int_{A} \pi(\theta)d\theta$$

where $A, h(A) = \tilde{A} \subseteq \mathbb{R}$. To get this invariance we need the following property to hold:

$$\tilde{\pi}(\tilde{\theta})d\tilde{\theta} = \pi(\theta)d\theta.$$

Suppose you chose $h(\cdot)$ so that:

$$X|\tilde{\theta} \sim \mathcal{N}(\tilde{\theta}, \sigma^2)$$

...

Numerical Methods for Bayesian Inference