POL-GA 1250:
Introduction to Quantitative Political Analysis I

Bernd Beber
New York University

Fall 2012

These are lecture notes as distributed to students. They do not include all in-class material and occasionally leave whitespace for classroom examples and exercises.

Quant I: Lecture 2

Bernd Beber
New York University

Fall 2012

Probability theory

- We know how to formulate expectations about samples on the basis of population properties (population → sample)

- But usually, we will work the other way around: infer population characteristics from sample features (sample → population)

- More precisely, we usually ask: how likely is it that we observe a particular sample if the population of interest is hypothesized to have a particular characteristic?

So let's figure out how to make statements about the probability of observing a particular sample.

First, terminology

- Observation: realized outcome of a data-generating process
- Events: set of potential outcomes; any subset of the sample space. E.g. "Romney wins" & "Obama wins" are events, "Obama wins" would be an observation
- Sample space: a set containing all possible events. Call this S. Note that S itself is an event which occurs whenever data is generated
- Sample points: elements of S. Also, sample events. Call these E_i for event i; mutually exclusive

- Compound event: a set containing more than one E_i. A subset of S. Call this A

- Discrete sample space: sample space containing a finite number of sample points

Remember, we want to estimate the likelihood of an event given a sample space of particular properties that we hypothesize/propose.

So, let's assign a number to each event that expresses this likelihood: $P(A)$ for A in S. This is the probability of A.

3 axioms:

1. $P(A) \geq 0$
2. $P(S) = 1$
3. $P(A_1 \cup A_2 \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$ for mutually exclusive A_i

How do we compute $P(A)$?

1. Let's count: Sample-point method

- $P(A)$: number of events in A, divided by the number of events in S, if events are equiprobable: $P(A) = \frac{|A|}{|S|}$
- In general: $P(A) = \frac{\sum_{i \in A} P(E_i)}{\sum_{i \in S} P(E_i)}$

Tools

- Ordered sample with repetition (mn-rule):
  - number of pairwise (ordered) combinations of two lists of size m and n is mn.
  - similarly, the number of 3-tuples for lists of size m, n and p is mnp.
  - for $m = n$, we have n^r for r draws

- Ordered sample without repetition (permutation theorem)

  E.g. suppose I draw up a random list of $r = 100$ people in a town of $n = 1000$ people to call for canvassing (and will start to call at the top). How many potential lists?

  By the mn-rule: $\overbrace{n}^{\text{1st pos. in 100-tuple}} \quad (n-1)(n-2)\cdots \quad \overbrace{(n-r+1)}^{\text{900th pos. in 100-tuple}}$

  Multiply by 1: $n(n-1)(n-2)\cdots(n-r+1)\frac{(n-r)!}{(n-r)!} =$

  $$= \frac{n!}{(n-r)!}$$

  Multiplying by $(n-r)!$ completes the factorial $n!$

- Unordered sample without repetition (combination theorem):

$$\frac{n!}{r!(n-r)!} = \binom{n}{r}$$

- Unordered sample with repetition (multi-set, or multi-choose):

$$\binom{n+r-1}{r}$$

2. Let's partition: Event composition method

- Conditional probability

- Independence

- Multiplicative law

- Associative property

- Additive law

- Complement

Law of total probability
S as a union of mutually exclusive subsets: $S = B_1 \cup B_2 \cup B_3 ... \cup B_k$,
such that $B_i \cap B_j = \emptyset, \forall i \neq j$.

Partition: $\{B_1, B_2, ..., B_k\}$ – a collection of sets

Decomposition of A: $A = (A \cap B_1) \cup (A \cap B_2) \cup ...(A \cap B_k)$

Let the collection $\{B_1, B_2, ..., B_k\}$ be a partition of S such that $P(B_i) > 0$ for $i = 1, 2, ..., k$. Then, $P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$.

Proof. Note that:

$$\begin{aligned}(A \cap B_i) \cap (A \cap B_j) \\ = A \cap (B_i \cap B_j) \\ = A \cap \emptyset = \emptyset\end{aligned}$$

So, no overlap. But that means:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + ...P(A \cap B_k) \quad \text{(Additive law)} \\ P(A) &= P(A|B_1)P(B_1) + ...P(A|B_k)P(B_k) \quad \text{(Multiplicative law)} \\ P(A) &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \end{aligned}$$

□

Bayes' Rule

$$\begin{aligned} P(B_j|A) &= \frac{P(A \cap B_j)}{P(A)} \quad \text{(def. of conditional probability)} \\ &= \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} \quad \begin{array}{l}\text{(def. of cond. prob.)} \\ \text{(law of total prob.)}\end{array} \end{aligned}$$

For two parts:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\neg B)P(\neg B)}$$

Quant I: Lecture 3

Bernd Beber
New York University

Fall 2012

Bayes' Theorem: Exercises

Exercise 1: *The political scientist and the politician each tell the truth one third of the time. The political scientist says of the politician "she just told the truth." What's the probability that the politician told the truth?*

We can use Bayes' rule to solve this problem. Let p represent that the political scientist says the politician told the truth, and t that the politician told the truth. We want to know $\Pr(t|p)$, which is:

$$\Pr(t|p) = \frac{\Pr(p|t)\Pr(t)}{\Pr(p|t)\Pr(t)+\Pr(p|\text{not } t)\Pr(\text{not } t)} = \frac{\frac{1}{3}\frac{1}{3}}{\frac{1}{3}\frac{1}{3}+\frac{2}{3}\frac{2}{3}} = \frac{1}{5}$$

The probability that the politician told the truth is lower than it was before the political scientist said she told the truth.

Exercise 2: *There are 500 parent-filed applications for 100 slots at a prestigious public school. The slots are filled by lottery, and the drawing has taken place, but the results have not yet been announced. Four of the applications come from different key supporters of the mayor, and the mayor wants to know if they are on the list of admitted candidates. His source at the school board has told him that at least one of the mayor's key supporters is on the list. The mayor wants to properly weigh the benefits and costs of (possibly illegally) meddling with the lottery results, and he asks you: What is the probability that at least two of his key supporters are on the list?*

Let A represent the event that at least one supporter has been admitted, and let T be the event that at least two supporters are on the list. Using Bayes' rule we then have:

$$\Pr(T|A) = \frac{\Pr(A|T)\Pr(T)}{\Pr(A)} = \frac{1\left(1-\frac{\binom{4}{0}\binom{496}{100}}{\binom{500}{100}}-\frac{\binom{4}{1}\binom{496}{99}}{\binom{500}{100}}\right)}{\left(1-\frac{\binom{4}{0}\binom{496}{100}}{\binom{500}{100}}\right)} \approx .305$$

Given that at least one supporter is on the list, the probability that at least two are on the list is about .305.

Now suppose the mayor's source at the school board adds that one particular key supporter, Mr. Smith, is on the list. Again, what is the probability that at least two of the mayor's key supporters are on the list?

Call S the event that Mr. Smith is on the list. We want to know

$$\Pr(T|S) = \frac{\Pr(T)\Pr(S|T)}{\Pr(S)} = \frac{\left(\frac{\binom{4}{2}\binom{496}{98}}{\binom{500}{100}} \quad \frac{\binom{4}{3}\binom{496}{97}}{\binom{500}{100}} \quad \frac{\binom{4}{4}\binom{496}{96}}{\binom{500}{100}}\right)\begin{pmatrix}.5\\.75\\1\end{pmatrix}}{\frac{\binom{1}{1}\binom{499}{99}}{\binom{500}{100}}}$$

$$= \frac{\frac{1}{2}\frac{\binom{4}{2}\binom{496}{98}}{\binom{500}{100}}+\frac{3}{4}\frac{\binom{4}{3}\binom{496}{97}}{\binom{500}{100}}+\frac{\binom{4}{4}\binom{496}{96}}{\binom{500}{100}}}{\frac{\binom{1}{1}\binom{499}{99}}{\binom{500}{100}}}$$

$$= \frac{3\binom{496}{98}+3\binom{496}{97}+\binom{496}{96}}{\binom{499}{99}} \approx .486$$

The posterior probability is .486. Although the difference in statements is subtle, the posterior probability increases substantially if the mayor's source says "Mr. Smith is on the list for sure, perhaps along with other key supporters" instead of "one key supporter is on the list for sure, perhaps along with others."

Addendum: A note on likelihoods and probabilities

Here's a bit of a clarification on notation and terminology. Generally, we distinguish the *probability* with which we observe certain data given some parameter from the *likelihood* that a parameter takes on a particular value given the data. Commonly people use the letter L to denote a likelihood (or likelihood function, which we will see soon), and the letter P for probabilities. For example, $L(\theta|y)$ denotes a likelihood of a variable θ, given some fixed data y.

Some of the confusion about probabilities and likelihoods stems from how we derive an expression for L from Bayes' rule:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

Since the denominator doesn't vary with θ, we typically write

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

The product on the right-hand side has two factors. We call $P(\theta)$ the prior distribution (or prior belief, or just prior). And $P(y|\theta)$ is often called the likelihood, or the likelihood component of the posterior. (See Wackerly et al., for example on page 798. They compound the confusion by using the letter L to denote this quantity.) Yes, this is confusing, because $P(y|\theta)$ is the probability of y given θ, i.e. the probability of observing varying data given a fixed parameter, not the other way around.

Let's try to clear up the confusion by taken one additional step: Let's say that the prior is uninformative, or flat, i.e. it doesn't vary in θ either. Then we're left with:

$$P(\theta|y) \propto P(y|\theta)$$

That's just where the likelihood function comes from:

$$L(\theta|y) = P(y|\theta)$$

So the reason why both the left-hand side and the right-hand side are sometimes referred to as likelihood is because we're talking about exactly the same expression, i.e. usually some probability density function. A density function will contain both y and θ; the key difference is what is allowed to vary. If θ is the variable (i.e. we're trying to find a parameter, given some fixed data we have), we talk about a likelihood. If y is the variable (i.e. we're trying to generate data, given some fixed parameter), we talk about a probability. So, the reason why you might see people refer to $P(y|\theta)$ as a likelihood is because doing so implies that we're analyzing $P(y|\theta)$ for the purpose of identifying $\hat{\theta}$, i.e. an estimate of θ.

Quant I: Lecture 4

Bernd Beber
New York University

Fall 2012

Plan for today:

- Random variables, distribution functions, and their properties
- Examples of discrete probability distributions

Next time:

- Examples of continuous distributions

Random variables and their properties

Random variable: A function that assigns one and only one value to each point in the sample space. For example, $Y : S \to \mathbb{N}$ or $Y : S \to \mathbb{R}$.

We denote random variables with capital letters, e.g. Y, as in $P(Y = 0)$.

Hypothetical or observed values are denoted by lower case letters, e.g. $P(Y = y)$.

We distinguish discrete random variables, which take on only a finite or countably infinite number of values, and continuous random variables, which can take on an uncountably infinite number of values.

Discrete case: We can define a probability distribution for a discrete random variable if we know $P(Y = y)$ for all possible y. Formally,

$$P(Y = y) = \sum_{E_i : Y(E_i) = y} P(E_i).$$

Probability distribution could be a table or a formula mapping y to $P(Y = y) = p(y)$. We call $p(y)$ the probability mass function (pmf).

Properties of a pmf:

1. $p(y) \geq 0 \ \forall y$
2. $\sum_y p(y) = 1$

We call the set of y where the mass function is not zero the support of $p(y)$.

Continuous case: We cannot assign $P(Y = y) > 0$ for an uncountably infinite number of values while also satisfying $\sum_y P(Y = y) = 1$. Instead, we can think about ranges of values the RV can take on.

Define the cumulative distribution function (cdf) as:

$$F(y) \equiv P(Y \leq y)$$

for $-\infty < y < \infty$.

Properties of a cdf:

1. $F(-\infty) \equiv \lim_{y \to -\infty} F(y) = 0$
2. $F(\infty) \equiv \lim_{y \to \infty} F(y) = 1$
3. $F(y)$ is non-decreasing, i.e $y_1 < y_2 \to F(y_1) \leq F(y_2)$.

For discrete RVs, the cdf will be a step function. A random variable is continuous if its cdf is continuous.

For continuous RVs, we have

$$P(Y = y) = 0 \ \forall y \in \mathbb{R},$$

but the density is not zero. Define:

$$f(y) \equiv \frac{dF(y)}{dy} = F'(y)$$

as the probability density function (pdf). This is equivalent to the pmf in the discrete case.

The area under the curve defined by the pdf is the probability mass.

So,

$$F(y) = \int_{-\infty}^{y} f(t)dt$$

Properties of a pdf:

1. $f(y) \geq 0 \ \forall y$
2. $\int_{-\infty}^{\infty} f(y)dy = 1$

The set of y where the density function is not zero is the support of $f(y)$.

Remarks:

1. $P(a \leq Y \leq b) = P(Y \leq b) - P(Y \leq a) = F(b) - F(a) = \int_a^b f(y)dy$
2. $P(a < Y < b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b)$

Characterizing random variables

Expected value:

For a discrete random variable,

$$E(Y) \equiv \sum_y yp(y).$$

For a continuous random variable,

$$E(Y) \equiv \int_{-\infty}^{\infty} yf(y)dy.$$

We write μ to denote the true expectation of Y, a constant. For example, $E(\mu) = \mu$.

Properties of expected values

1. If b is a constant, then $E(b) = b$.
2. If a and b are constants, then $E(aY + b) = aE(Y) + b$.
3. $E(Y + X) = E(Y) + E(X)$ for any Y and X.
4. If Y and X are independent, then $E(YX) = E(Y)E(X)$.
5. If Y is a discrete random variable with pmf $p(y)$ and if $g(y)$ is any function of Y, then $E(g(Y)) = \sum_y g(y)p(y)$. Similarly, if Y is a continuous random variable with pdf $f(y)$, then $E(g(Y)) = \int_{-\infty}^{\infty} g(y)f(y)dy$.

We can use these properties to derive other helpful results. For example:

Claim. *We can distribute expectations, so that* $E[g_1(Y) + g_2(Y)] = E[g_1(Y)] + E[g_2(Y)]$.

Proof. Since $g_1(Y) + g_2(Y)$ is a function of Y, we can write

$$E[g_1(Y) + g_2(Y)] = \sum_y [g_1(y) + g_2(y)]p(y).$$

We can then distribute summations to get

$$\sum_y [g_1(Y)p(y)] + \sum_y [g_2(Y)p(y)],$$

and hence

$$E[g_1(Y)] + E[g_2(Y)].$$

□

Variance:

A measure of the distribution of values of Y around its expected value:

$$\text{Var}(Y) \equiv \sigma^2 \equiv E[(Y-\mu)^2]$$

The positive square root of σ^2 is the standard deviation, σ.

For a discrete random variable,

$$\text{Var}(Y) = \sum_y (Y-\mu)^2 f(y).$$

For a continuous random variable,

$$\text{Var}(Y) \equiv \int_{-\infty}^{\infty} (y-\mu)^2 f(y) dy$$

Sometimes it is more convenient to compute

$$\text{Var}(Y) = E[(Y-\mu)^2] = E[(Y)^2] - \mu^2 = E[(Y)^2] - [E(Y)]^2$$

Properties of variance:

1. If a and b are constants, then $\text{Var}(aY+b) = a^2\text{Var}(Y)$.
2. If Y and X are independent random variables, then $\text{Var}(Y+X) = \text{Var}(Y) + \text{Var}(X)$ and $\text{Var}(Y-X) = \text{Var}(Y) + \text{Var}(X)$.

Examples of discrete probability distributions

Keep in mind: it's easy to look up the exact formulas for the moments or pdfs of distributions, but understand which distributions are appropriate for which sample space!

Bernoulli

$y \in \{0, 1\}$

$Y \sim \text{Bernoulli}(\pi)$

Distribution is conditional on one parameter:

Probability of success $\pi \in [0, 1]$

$$p(y|\pi) = \begin{cases} \pi & \text{if } y = 1 \\ 1 - \pi & \text{if } y = 0 \end{cases}$$

$$= \pi^y (1 - \pi)^{(1-y)}$$

$E(Y) = \pi$

$\text{Var}(Y) = \pi(1 - \pi)$

Binomial

Suppose you run a series of Bernoulli trials and only observe the total number of successes (unordered draws, with independent replacement):

$y \in \{0, 1, \ldots, n\}$

$Y \sim \text{Binomial}(n, \pi)$

Now we have two parameters:

1. Number of trials $n \in \{1, 2, ...\}$
2. Probability of success $\pi \in [0, 1]$

$p(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$

$E(Y) = n\pi$

$\text{Var}(Y) = n\pi(1 - \pi)$

Note: $\text{Var}(Y) < E(Y)$ for binomial distributions.

If $\text{Var}(Y) > E(Y)$, then Negative binomial.
If $\text{Var}(Y) = E(Y)$, then Poisson.

Poisson

$y \in \{0, 1, \ldots\}$

$Y \sim \text{Poisson}(\lambda)$

Expected number of occurrences $\lambda > 0$

$p(y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$

$E(Y) = \lambda$

$\text{Var}(Y) = \lambda$

The probability mass function for a Poisson distribution can be derived as a limiting case of the binomial distribution:

$$\begin{aligned}
\frac{\lambda^y}{y!} e^{-\lambda} &= \lim_{n\to\infty} \binom{n}{y} \pi^y (1-\pi)^{(n-y)} \\
&= \lim_{n\to\infty} \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{(n-y)} \\
&= \lim_{n\to\infty} \frac{n!}{(n-y)!n^y} \left(\frac{\lambda^y}{y!}\right) \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-y}}_{1 \text{ for } n\to\infty} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{n}}_{e^{-\lambda} \text{ for } n\to\infty} \\
&= \frac{\lambda^y}{y!} e^{-\lambda} \lim_{n\to\infty} \frac{n!}{(n-y)!n^y} \\
&= \frac{\lambda^y}{y!} e^{-\lambda} \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-y+1)}{n^y} \\
&= \frac{\lambda^y}{y!} e^{-\lambda} \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-y+1)}{n \cdot n \cdot n \ldots n} \\
&= \frac{\lambda^y}{y!} e^{-\lambda} \lim_{n\to\infty} 1 \cdot (1-1/n)\ldots(1-(y+1)/n) \\
&= \frac{\lambda^y}{y!} e^{-\lambda}
\end{aligned}$$

Note: Bernoulli, binomial, Poisson, normal, and many other distributions are exponential families.

Quant I: Lecture 5

Bernd Beber
New York University

Fall 2012

Examples of continuous probability distributions

Normal (Univariate)

Important, because:

1. Many empirical distributions are approximately Normal
2. Sampling distribution of the sample mean is often Normal (Central Limit Theorem)!

$y \in \mathbb{R}$

$Y \sim \text{Normal}(\mu, \sigma^2)$

Two parameters:

1. Mean: $\mu \in \mathbb{R}$
2. Variance: $\sigma^2 > 0$

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$E(Y) = \mu$

$\text{Var}(Y) = \sigma^2$

Often use $\phi(y)$ for a normal density, especially the standard normal, with $\mu = 0$ and $\sigma^2 = 1$.

To standardize a normally distributed Y to obtain standard normally distributed Z:

$$Z \equiv \frac{Y - \mu_Y}{\sigma_Y}$$

No closed-form cdf exists for the normal distribution, but we refer to it frequently and sometimes denote it $\Phi(y) = \int_{-\infty}^{y} \phi(t)dt$.

Uniform

$y \in [\alpha, \beta]$

$Y \sim \text{Uniform}(\alpha, \beta)$

Interval $[\alpha, \beta]$, $\beta > \alpha$

$$f(y|\alpha, \beta) = \frac{1}{\beta - \alpha}$$

$E(Y) = \frac{\alpha+\beta}{2}$

$\text{Var}(Y) = \frac{(\beta-\alpha)^2}{12}$

Proof of variance.

$$\begin{aligned}
\text{Var}(Y) &= \int_\alpha^\beta \left(y - \frac{\alpha+\beta}{2}\right)^2 \frac{1}{\beta-\alpha} dy \\
&= \frac{\left(y - \frac{\alpha+\beta}{2}\right)^3}{3} \frac{1}{\beta-\alpha}\Bigg|_\alpha^\beta \\
&= \frac{\left(\frac{\beta-\alpha}{2}\right)^3}{3(\beta-\alpha)} - \frac{\left(\frac{\alpha-\beta}{2}\right)^3}{3(\beta-\alpha)} \\
&= \frac{(\beta-\alpha)^2}{24} + \frac{(\alpha-\beta)^2}{24} \\
&= \frac{(\beta-\alpha)^2}{12}
\end{aligned}$$

□

Gamma

$y \in [0, \infty)$, or alternatively we can write $y \in \mathbb{R}_0^+$

$Y \sim \text{Gamma}(\alpha, \beta)$

Shape $\alpha > 0$, inverse scale $\beta > 0$

$$f(y|\alpha, \beta) = \frac{\beta^\alpha y^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta y}$$

$E(Y) = \frac{\alpha}{\beta}$

$\text{Var}(Y) = \frac{\alpha}{\beta^2}$

Example: Let α be a positive integer, and $\beta = 1$. For integer α, we have $\Gamma(\alpha) = (\alpha - 1)!$ and hence:

$$f(y|\alpha, 1) = \frac{y^{\alpha-1}}{(\alpha-1)!} e^{-y}$$

Note that $E(Y) = \text{Var}(Y) = \alpha$.

Recall the pmf for the Poisson distribution, $p(y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$.

The Gamma distribution gives us a density of the parameter of the Poisson distribution ("conjugate prior").

Beta

$y \in [0, 1]$

$Y \sim \text{Beta}(\alpha, \beta)$

Shape parameters $\alpha, \beta > 0$

$$\begin{aligned} f(y|\alpha, \beta) &= \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt} \\ &= \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \end{aligned}$$

where $B(\alpha, \beta)$ is the Beta function. Conjugate prior for Binomial, Bernoulli.

$E(Y) = \frac{\alpha}{\alpha+\beta}$

$\text{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Uniform distribution on unit interval is the special case where $\alpha, \beta = 1$.

t-Distribution

Also known as Student's t-distribution

$y \in \mathbb{R}$

$Y \sim t(n)$

n degrees of freedom

$$f(y) = \frac{1}{\sqrt{n}\mathrm{B}\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

$E(Y) = 0$ for $n > 1$

$\mathrm{Var}(Y) = \frac{n}{n-2}$ for $n > 2$

As $n \to \infty$, t approaches the standard normal

Addendum: A note on the Gamma distribution

A student asked about the relative roles of shape α and inverse scale β in determining what the graph of the density function for the Gamma distribution looks like. I indicated that it's really α that determines the nature of the curve, and also pointed to the expected value and variance to give some intuition as to why β is a scale parameter. Take a look at the graphs on the next page for some more intuition.

The leftmost graphs in both rows are identical. They show the density for a random variable that is distributed Gamma with shape $\alpha = 10$ and inverse scale $\beta = 1$. (I used R for these graphs, which by default uses x to denote a variable in plotting a curve, and refers to the inverse scale parameter as the rate.) For the remaining two graphs in the top row, I then change the shape parameter, to 6 and 2. For the graphs in the bottom row, I change the inverse scale to 2 and 3 *and I correspondingly rescale the horizontal axis.* As you can see, the changes in α produce different curves, while the graphs in the bottom row all look exactly the same.

See below for the R code I used to create these graphs:

```
pdf("Gamma.pdf", width = 11, height = 8.5)
    par(mfrow=c(2, 3))
    curve(dgamma(x, shape = 10, rate = 1), xlim = c(0, 20))
    curve(dgamma(x, shape = 6, rate = 1), xlim = c(0, 20))
    curve(dgamma(x, shape = 2, rate = 1), xlim = c(0, 20))

    curve(dgamma(x, shape = 10, rate = 1), xlim = c(0, 20))
    curve(dgamma(x, shape = 10, rate = 2), xlim = c(0, 20/2))
    curve(dgamma(x, shape = 10, rate = 3), xlim = c(0, 20/3))
dev.off()
```

dgamma(x, shape = 10, rate = 1)
dgamma(x, shape = 10, rate = 2)
dgamma(x, shape = 6, rate = 1)
dgamma(x, shape = 10, rate = 3)
dgamma(x, shape = 2, rate = 1)
x

Quant I: Lecture 6

Bernd Beber
New York University

Fall 2012

Maximum likelihood

Identify a distribution that could have plausibly generated your data. Write down the likelihood function, and maximize with respect to any unknown parameters of the distribution. That is, select parameter values that maximize the probability of seeing the data that you've seen.

What is the likelihood function?

Recall Bayes' rule:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

$P(y)$ is used to scale the posterior, and we can write

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

Usually we leave aside $P(\theta)$ (in Bayesian terms, we assume an uninformative prior) and have

$$L(\theta|y) \propto P(y|\theta)$$

$P(y|\theta)$ is a probability mass or density function, but we now vary θ instead of y.

Note that $L(\theta|y)$ allows us to identify a best estimate θ by maximum likelihood estimation (MLE), but we cannot compare the magnitude of L across models on different data.

Example: Normal distribution with unknown mean and variance.

Recall:

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Then:

$$f(y_1, y_2, \ldots, y_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

Now, maximize with respect to μ, σ^2. This is easier when we take the log, which doesn't change the maximization problem because we are applying a monotonic function:

$$\ln L(\mu, \sigma^2|y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}$$

Taking partial derivatives yields:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0 \tag{1}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0 \tag{2}$$

Multiply equation 1 by σ^2:

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_{ML}) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \hat{\mu}_{ML} = \bar{y}$$

Rearrange equation 2, and substitute for μ:

$$\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = \frac{n}{2\sigma^2}$$

$$\begin{aligned}\hat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\end{aligned}$$

Another example of MLE: Bernoulli distribution with unknown parameter π

We have probability mass function

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

and hence

$$p(y_1, y_2, \ldots, y_n|\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{(1-y_i)}.$$

This gives us the log-likelihood

$$\begin{aligned} \ln L(\pi|y) &= \sum_{i=1}^{n}[y_i \ln \pi + (1-y_i)\ln(1-\pi)] \\ &= \left(\sum_{i=1}^{n} y_i\right)\ln \pi + \left(n - \sum_{i=1}^{n} y_i\right)\ln(1-\pi). \end{aligned}$$

Taking the derivative yields

$$\frac{\partial \ln L}{\partial \pi} = \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{n - \sum_{i=1}^{n} y_i}{1-\pi} = 0.$$

Let $\sum_{i=1}^{n} y_i = k$. Then

$$\begin{aligned} \frac{k}{\pi} - \frac{n-k}{1-\pi} &= 0 \\ \frac{k(1-\pi) - (n-k)\pi}{\pi(1-\pi)} &= 0 \\ k - \pi k - n\pi + k\pi &= 0 \\ \pi &= \frac{k}{n} \\ \hat{\pi}_{ML} &= \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}. \end{aligned}$$

We can also say something about the precision of this estimate using the result from problem set 2 that $\text{Var}(E(Y)) = \frac{Var(Y)}{n}$. Hence,

$$\text{Var}(\hat{\pi}_{ML}) = \frac{\hat{\pi}_{ML}(1-\hat{\pi}_{ML})}{n}.$$

As you may recall, the proof of this equality required that we consider the possibility that we are dealing with many different distributions, namely one for each y_i. In particular, we

considered $\text{Var}(Y_i)$ for each i. Note that Y_i is a random variable, i.e. the same observation i can take on different values in repeated samples.

This underpins, for the most part, how we'll make inferences about the precision of our estimates in this course: We take a primarily frequentist approach (parameters are fixed, data is variable), as opposed to a Bayesian approach (parameters are variable, data is fixed).

In order to get a better handle on this, we need a better understanding of multivariate probability distributions. This will also be crucial for when we condition on other data, say X. Often, we're interested not just in a characteristic of y, but how it relates to X, i.e. a multivariate statistic.

Multivariate probability distributions

Example:

Let's think about an example of jointly distributed variables in the context of this year's election.

$Y_1 \sim \text{Bernoulli}(\pi_1)$ "predict House results and record whether Republicans retain majority."
$Y_2 \sim \text{Bernoulli}(\pi_2)$ "predict Senate results and record whether Republicans win majority."

Suppose we estimate π_1 and π_2 from current polling data. Let's say $\hat{\pi}_1 = .8$ and $\hat{\pi}_2 = .5$. Do we know now the probability that Republicans will control both chambers? No! Unless we are willing to assume Y_1 and Y_2 are independent.

We have:
$P(Y_1 = 1 \cap Y_2 = 1) = P(Y_1 = 1|Y_2 = 1)P(Y_2 = 1)$ (by multiplicative law)

And we have:
$P(Y_1 = 1|Y_2 = 1) = P(Y_1 = 1)$ if Y_1, Y_2 are independent.

In this case, with independence [an example of dependence is in square brackets]:

		Y_1: GOP holds House		Total
		yes $P(Y_1 = 1)$	no $P(Y_1 = 0)$	
Y_2: GOP holds Senate	yes $P(Y_2 = 1)$	.4 [.5]	.1 [0]	.5
	no $P(Y_2 = 0)$	.4 [.3]	.1 [.2]	.5
	Total	.8	.2	

This describes a joint probability distribution, with $p(y_1, y_2)$ for each possible y_1, y_2. Generically, for discrete variables, we have the joint pmf

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2),$$

and joint cmf:

$$\begin{aligned} F(y_1, y_2) &= P(Y_1 \leq y_1, Y_2 \leq y_2) \\ &= \sum_{t_1 \leq y_1} \sum_{t_2 \leq y_2} p(t_1, t_2) \end{aligned}$$

RVs are jointly continuous if their joint pdf is continuous, so that:

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2) \, \mathrm{d}t_2 \, \mathrm{d}t_1,$$

for non-negative f and $-\infty < y_1 < \infty$, $-\infty < y_2 < \infty$.

Hence

$$P(a < Y_1 < b, c < Y_2 < d) = \int_a^b \int_c^d f(y_1, y_2) \, \mathrm{d}y_2 \, \mathrm{d}y_1.$$

Properties:

$$F(-\infty, -\infty) = F(-\infty, y_2) = F(y_1, -\infty) = 0$$
$$F(\infty, \infty) = 1$$

Marginal distribution

- Discrete:
$$p_i(y_i) = \sum_{\text{all } y_j} p(y_i, y_j)$$
This also works with additional RVs:
$$p_i(y_i) = \sum_{\text{all } y_k} \sum_{\text{all } y_j} p(y_i, y_j, y_k)$$

- Continuous:
$$f_i(y_i) = \int_{-\infty}^{\infty} f(y_i, y_j)\, \mathrm{d}y_j$$

Conditional distribution

- Discrete:
$$p(y_1|y_2) = \frac{p(y_1, y_2)}{p_2(y_2)}$$

- Continuous:
$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$
For the cdf, $P(Y_1 \leq y_1 | Y_2 = y_2) = F(y_1|y_2)$.

Independence

Now we can talk about independence a bit more formally. Recall that the events A and B are independent if $P(A \cap B) = P(A)P(B)$. Similarly, RVs are independent if

$$F(y_i, y_j) = F_i(y_i)F_j(y_j)$$

for every pair (y_i, y_j). Hence:

- Discrete: $p(y_1, y_2) = p_1(y_1)p_2(y_2)$
- Continuous: $f(y_1, y_2) = f_1(y_1)f_2(y_2)$

All of this can be extended to many RVs, e.g. $f(y_1, y_2, ..., y_k) = f_1(y_1)f_2(y_2)...f_k(y_k)$ for k independent continuous random variables.

Expected values

- Discrete:

$$E(g(Y_1, Y_2, \ldots, Y_k)) = \sum_{y_k} \sum_{y_{k-1}} \cdots \sum_{y_1} g(y_1, y_2, \ldots, y_k) p(y_1, y_2, \ldots, y_k)$$

- Continuous:

$$E(g(Y_1, Y_2, \ldots, Y_k)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_k) f(y_1, y_2, \ldots, y_k) \, \mathrm{d}y_1 \, \mathrm{d}y_2 \ldots \, \mathrm{d}y_k$$

Rules of expectations hold similarly to the univariate case.

Also, if Y_1 and Y_2 are independent:

$$E(g(Y_1)h(Y_2)) = E(g(Y_1))E(h(Y_2))$$

Proof.

$$\begin{aligned} E(g(Y_1)h(Y_2)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1)h(y_2)f(y_1, y_2) \, \mathrm{d}y_1 \, \mathrm{d}y_2 \\ \text{Because of independence:} & \\ &= \int_{-\infty}^{\infty} g(y_1)f_1(y_1) \Big[\int_{-\infty}^{\infty} h(y_2)f_2(y_2) \, \mathrm{d}y_2 \Big] \, \mathrm{d}y_1 \\ &= E(h(y_2)) \int_{-\infty}^{\infty} g(y_1)f_1(y_1) \, \mathrm{d}y_1 \\ &= E(g(Y_1))E(h(Y_2)) \end{aligned}$$

□

Covariance

If Y_1 and Y_2 are not independent, we can measure how related they are:

$$\mathrm{Cov}(Y_1, Y_2) \equiv E((Y_1 - \mu_1)(Y_2 - \mu_2)),$$

where μ_1 and μ_2 are the means of Y_1 and Y_2. Covariance is the product of deviations from the means.

What if we change the scale, e.g. multiply by 10? The covariance increases. To make the measure comparable across scales, standardize it:

Correlation coefficient ("rho"):

$$\rho \equiv \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1 \sigma_2},$$

where $-1 \leq \rho \leq 1$.

Having defined the covariance, we can identify two additional properties of expectations and variances (including for non-independent distributions):

1. $E(Y_1 Y_2) = E(Y_1)E(Y_2) + \text{Cov}(Y_1, Y_2)$
2. $\text{Var}(Y_1 + Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2) + 2\text{Cov}(Y_1, Y_2)$

Quant I: Lecture 7

Bernd Beber
New York University

Fall 2012

Now that we know a few things about multivariate distributions, let's get back to the following: how do we measure the precision of our estimates?

Chebychev's inequality

Recall that we can compute the variance $\text{Var}(E(Y)) = \frac{\text{Var}(Y)}{n}$, so we can make use of Chebychev's inequality:

$$P(|Y - \mu| \geq \kappa\sigma) \leq \frac{1}{\kappa^2},$$

where Y is an RV with mean μ and standard deviation σ, and κ is any real number.

Proof.

$$\begin{aligned} \text{Write } P(|Y - \mu| \geq \kappa\sigma) &= E(I(|Y - \mu| \geq \kappa\sigma)) \\ &= E\left(I\left(\frac{(Y-\mu)^2}{(\kappa\sigma)^2} \geq 1\right)\right) \end{aligned}$$

But $\left(\frac{Y-\mu}{\kappa\sigma}\right)^2$ is positive, and $I(\cdot)$ assigns 0 $\forall$ values between 0 and 1, and 1 for values at 1 or above. Hence:

$$\begin{aligned} P(|Y - \mu| \geq \kappa\sigma) &\leq E\left(\left(\frac{Y-\mu}{\kappa\sigma}\right)^2\right) \\ &= \frac{1}{\kappa^2}\frac{E((Y-\mu)^2)}{\sigma^2} \\ &= \frac{1}{\kappa^2} \end{aligned}$$

□

Note that we can also write:

$$P(|Y - \mu| \leq \kappa\sigma) \geq 1 - \frac{1}{\kappa^2},$$

and

$$P\left(\frac{|Y - \mu|}{\sigma} \geq \kappa\right) \leq \frac{1}{\kappa^2},$$

for a standardized variable. This puts remarkable but loose bounds on Y. E.g., let's say $\kappa = 2$. The probability that a variable is more than 2 standard deviations away from its mean is less than $\frac{1}{2^2} = \frac{1}{4}$.

But often this isn't enough. What if we want to make more precise statements about the precision of our estimates?

For this we need to think about the distribution of our statistics—the sampling distribution. As the name suggests, we are now thinking about inferences made on the basis of (random) samples from some population. If we are interested in the characteristics of the sample itself, it makes little sense to talk about variability or uncertainty of those characteristics!

If at all possible, we would like to avoid making strong distributional assumptions, and it turns out we can!

The Central Limit Theorem

$\overline{Y}$ has a sampling distribution that's approximately normal, for sufficiently large samples.

Formally:

Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. RVs with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$. Define

$$U_n \equiv \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

as a standardization of $\overline{Y}$ to mean 0 and standard deviation 1.

Then the cdf of U_n converges to the cdf of the standard normal:

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t, \forall u$$

Another way of saying this is that $\overline{Y}$ is asymptotically normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

Estimation

Now with a sampling distribution in hand, let's think again about estimation.

There are two kinds of estimates (which are generated through rules/formulas called estimators):

1. Point estimates: a single value
2. Interval estimates: two values that define an interval around the parameter of interest.

How do we know what's a good estimate? For example, why might $\frac{1}{n}\sum_{i=1}^{n} Y_i$ be a better estimate than $\frac{1}{n}\sum_{i=1}^{n}(Y_i + 1)$ for μ?

Two properties:

1. Unbiasedness / consistency
2. Small variance / efficiency

(Here, expectations and their rules will come in handy!)

Let's call an estimator $\hat{\theta}$ for a parameter θ.

First, let's define unbiasedness. An estimator is unbiased if it is equal to the parameter it claims to estimate, in expectation. That is:

$\hat{\theta}$ is an unbiased estimator for θ if $E(\hat{\theta}) = \theta$.

$\hat{\theta}$ is biased if $E(\hat{\theta}) \neq \theta$.

Bias is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Example:

Consider $\overline{Y_B} = \frac{1}{n}\sum_{i=1}^{n}(Y_i + 1)$ as an estimate for μ. One way to show that this isn't a good estimator is to show that $E(\overline{Y_B}) \neq \mu$ and hence $B(\overline{Y_B}) \neq 0$.

$$\begin{aligned} E(\overline{Y_B}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i + 1)\right] \\ &= \frac{1}{n}\left[\left(\sum_{i=1}^{n} E(Y_i)\right) + nE(1)\right] \\ &= \frac{1}{n}(n\mu + n) \\ &= \mu + 1 \neq \mu, \text{ and } B(\overline{Y_B}) = 1. \end{aligned}$$

Second, efficiency. We would like our estimator to be as close as possible to θ in repeated samples. That is, we want the variance of the sampling distribution of our estimator, which we can write $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$, to be as small as possible.

Sometimes we face a trade-off between reducing an estimator's bias and reducing its variance. We can see this when we try to minimize an estimator's mean square error (MSE), which is defined as the expected value of the squared distance between the estimator and the parameter:

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

In particular, we can show that

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (B(\hat{\theta}))^2.$$

Proof. Note that

$$\begin{aligned} (\hat{\theta} - \theta) &= (\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta) \\ &= (\hat{\theta} - E(\hat{\theta})) + B(\hat{\theta}). \end{aligned}$$

Hence,

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[(\hat{\theta} - E(\hat{\theta}))^2 + B(\hat{\theta})^2 + 2B(\hat{\theta})(\hat{\theta} - E(\hat{\theta}))] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + E[B(\hat{\theta})^2] + 2E[B(\hat{\theta})(\hat{\theta} - E(\hat{\theta}))] \\ &= \text{Var}(\hat{\theta}) + B(\hat{\theta})^2. \end{aligned}$$

□

Note that the intuitive estimator is not always the best estimator! In particular, we'll show that a particular intuitive estimator is in fact biased.

Suppose you want to estimate population variance σ^2. It might seem natural to use the sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

as an estimator.

Let's check if s^2 would be an unbiased estimator of the population variance σ^2:

$$
\begin{aligned}
E(s^2) &= E\left(\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n}\right) \\
&= E\left(\frac{\sum_{i=1}^{n}(Y_i^2+\overline{Y}^2-2Y_i\overline{Y})}{n}\right) \\
&= E\left(\frac{\sum_{i=1}^{n}Y_i^2}{n}+\frac{n\overline{Y}^2}{n}-2\overline{Y}^2\right) \\
&= E\left(\frac{\sum_{i=1}^{n}Y_i^2}{n}-\overline{Y}^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2\right) \\
&= \frac{1}{n}\left[\left(\sum_{i=1}^{n}E(Y_i^2)-nE(\overline{Y}^2)\right)\right]
\end{aligned}
$$

Since $\mathrm{Var}(Y)=E(Y^2)-E(Y)^2$, we have:

$$
\begin{aligned}
E(s^2) &= \frac{1}{n}\left[\left(\sum_{i=1}^{n}\mathrm{Var}(Y)+E(Y)^2\right)-n\left(\mathrm{Var}(\overline{Y})+E(\overline{Y})^2\right)\right] \\
&= \frac{1}{n}\left[\left(\sum_{i=1}^{n}\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \frac{1}{n}\left[n(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \frac{1}{n}(n\sigma^2-\sigma^2) \\
&= \frac{n-1}{n}\sigma^2 \\
&\neq \sigma^2
\end{aligned}
$$

We can correct this bias by multiplying:

$$
\frac{n}{n-1}E\left(\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n}\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)
$$

We'll refer to this corrected quantity as s_u.

Sampling distribution for small samples

Suppose we compute the sample average $\overline{Y}$ as an estimator for the true population mean μ, and we compute $\frac{s_u^2}{n}$ as an unbiased estimator for $Var(\overline{Y}) = \frac{Var(Y)}{n}$. However, our sample size (that is, n) is small. Can we still say something about the sampling distribution of $\overline{Y}$?

In particular, we may be interested in the sampling distribution of the standardized sample average, $\frac{\overline{Y} - \mu}{s_u/\sqrt{n}}$ which is itself a random variable. We standardize by subtracting the (often hypothesized) true mean from our observed realizations of $\overline{Y}$ and then divide by the standard deviation of the sample average. This gives us a random variable with mean 0 and standard deviation 1.

It is often attractive to standardize random variables in this fashion. For example, if you want to plot different empirical distributions on the same scale (for example, SAT and ACT scores), simply subtract from each set of data its mean and divide by its standard deviation to put both sets of scores on the same scale. We want to standardize the distribution of $\overline{Y}$ so we can make scale-independent statements about it.

So, how might $\frac{\overline{Y} - \mu}{s_u/\sqrt{n}}$ be distributed? Let's prove the following claim:

Claim. *If $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed Normal, then $\frac{\overline{Y} - \mu}{s_u/\sqrt{n}}$ is distributed t with ν degrees of freedom.*

Proof. First, let's rewrite

$$\begin{aligned}
\frac{\overline{Y} - \mu}{s_u/\sqrt{n}} &= \frac{\sqrt{n}(\overline{Y} - \mu)}{s_u} \\
&= \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{s_u/\sigma} && \left(\text{Multiply by } \frac{1/\sigma}{1/\sigma}\right) \\
&= \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s_u^2/\sigma^2}} && \text{(Square and square root)} \\
&= \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s_u^2}{\sigma^2}/(n-1)}} && \left(\text{Multiply part by } \frac{n-1}{n-1}\right)
\end{aligned}$$

The numerator is the standardized sampling distribution for $\overline{Y}$ with known population variance (i.e. with σ, not s_u). We know that $\overline{Y}$ is a linear combination of i.i.d. and normally distributed variables, which implies that $\overline{Y}$ is also normally distributed. Subtracting and

dividing by constants does not change the distribution, so the numerator $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ is distributed Normal. Let's call this expression Z.

What about the denominator? The only random variable here is s_u^2, and we know how certain functions of s_u^2 are distributed. In particular, we know that

$$\frac{(n-1)s_u^2}{\sigma^2}$$

has a χ^2 (chi-squared) distribution with $\nu = n-1$ degrees of freedom (see Wackerly et al., Theorem 7.3 on p. 357, as well as section 8.9 on pp. 434-435). Call this expression W.

This leaves us with

$$\frac{\overline{Y}-\mu}{s_u/\sqrt{n}} = \frac{Z}{\sqrt{W/\nu}}$$

which is known to have a t distribution with ν degrees of freedom (see Wackerly et al., p. 360). □

This means that the standardized sample average $\overline{Y}$ is distributed t in small samples, provided we are working with data that is distributed Normal. We will use this fact when we construct confidence intervals for small samples.

Quant I: Lecture 8

Bernd Beber
New York University

Fall 2012

Confidence intervals

Last time we talked (mostly) about point estimators, and two desirable properties of estimators: Unbiasedness and efficiency (small variance). Today we'll focus on an interval estimator: A rule that specifies how to use data to calculate two numbers that form the endpoints of an interval containing a parameter of interest.

Two desirable properties:

1. Contain the parameter of interest θ
2. Be relatively narrow

As with point estimators, the length and location of the interval are random quantities, so our goal is to find an interval estimator that generates narrow intervals with a high probability of trapping θ.

In particular, we'll focus on confidence intervals (CIs), the most commonly used interval estimator:

- CIs are constructed of two quantities called the upper and lower confidence limits, or upper and lower bounds.
- The probability that a random CI will include θ is called the confidence coefficient, or confidence level. It is the fraction of the time, in repeated sampling, that the CI will contain θ. We thus like the confidence coefficient associated with our CI to be high.
- The confidence coefficient is written $1 - \alpha$. If $\hat{\theta}_L$ and $\hat{\theta}_U$ are the random lower and upper confidence limits, then

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha.$$

We typically call a CI with confidence coefficient $1 - \alpha$ a "$100 \cdot (1 - \alpha)$-percent confidence interval." Sometimes we also call such CI a "confidence interval with significance level α."

Our task then is to find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = (1 - \alpha).$$

First, let's construct a confidence interval for a sample statistic $\hat{\theta}$ that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. For example, $\hat{\theta} = \overline{Y}$ in large samples.

From the Central Limit Theorem, we recall that

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is distributed approximately standard Normal in this case.

To construct a confidence interval for $\hat{\theta}$, we pick two values $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha,$$

or

$$\int_{-z_{\alpha/2}}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, \mathrm{d}t.$$

We substitute for Z, and get

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2}\sigma_{\hat{\theta}} \leq \hat{\theta} - \theta \leq z_{\alpha/2}\sigma_{\hat{\theta}}) \\
&= P(-z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta} \leq -\theta \leq z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta}) \\
&= P(z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta} \geq \theta \geq -z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta}) \\
&= P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}})
\end{aligned}$$

Hence, $\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}$ and $\hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$.

But what is $z_{\alpha/2}$?

It's the value satisfying $P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$, that is

$$z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right),$$

where Φ is the inverse standard normal cdf which returns a value of z given some probability.

For example:

$$-\Phi^{-1}(.025) \approx 1.96$$
$$-\Phi^{-1}(.05) \approx 1.64$$

In Stata, type `display invnormal(.05)`.

So, a 95% confidence interval is approximately

$$[\hat{\theta} - 1.96\sigma_{\hat{\theta}}, \hat{\theta} + 1.96\sigma_{\hat{\theta}}].$$

What is $\sigma_{\hat{\theta}}$?

$$\sigma_{\hat{\theta}} = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

That is, $\sigma_{\hat{\theta}}$ has two components:

1. variability in the population
2. variability due to the sample

Remember: the variance of our estimator is the population variance divided by the number of observations.

So what is σ^2? We usually don't know σ^2, but recall that we derived an unbiased estimator for it when we were thinking about $\overline{Y}$:

$$s_u^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Unfortunately, unbiasedness is not quite what we need, because we want to make a statement not about $E(s_u^2)$ but the distribution of an expression containing s_u^2. In particular, we require:

$$F\left(\frac{\overline{Y} - \mu}{s_u/\sqrt{n}}\right) \xrightarrow{p} \Phi$$

For this, we need s_u^2 to be <u>consistent</u>.

<u>Definition</u>: $\hat{\theta}_n$ is a consistent estimator for θ if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

Equivalently, "$\hat{\theta}_n$ converges in probability to θ":

$$\hat{\theta}_n \xrightarrow{p} \theta$$

Note: If $\hat{\theta}$ is unbiased for θ and $\lim_{n\to\infty} \text{Var}(\hat{\theta}_n) = 0$, then $\hat{\theta}$ is consistent for θ.

Example: $\overline{Y}$ is unbiased for μ.

Since $\text{Var}(\overline{Y}) = \frac{\sigma^2}{n}$, $\lim_{n\to\infty} \text{Var}(\overline{Y}) = 0$, and $\overline{Y}$ is consistent for μ.

$\Rightarrow$ This is known as the law of large numbers: $\overline{Y} \xrightarrow{p} \mu$

Note:

- Consistency does not imply unbiasedness
- Unbiasedness does not imply consistency

Example: s^2 is a biased but consistent estimator for σ^2, where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

If some unbiased estimator's variance does not decline in n, it won't be consistent.

Properties:

1. If $\hat{\theta} \xrightarrow{p} \theta$, and $\hat{\theta}' \xrightarrow{p} \theta'$, then:
   - $\hat{\theta} + \hat{\theta}' \xrightarrow{p} \theta + \theta'$
   - $\hat{\theta}\hat{\theta}' \xrightarrow{p} \theta\theta'$
   - $\frac{\hat{\theta}}{\hat{\theta}'} \xrightarrow{p} \frac{\theta}{\theta'}$, for $\theta' \neq 0$
2. If $g(\cdot)$ is continuous at θ, $g(\hat{\theta}) \xrightarrow{p} g(\theta)$

Claim. *s_u^2 is a consistent estimator for σ^2.*

Proof.

$$\begin{aligned} s_u^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \quad \text{(as shown earlier)} \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2\right) \end{aligned}$$

Now let's consider:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2 &= \overbrace{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Y_i^2}^{\text{Law of large numbers}}-\overbrace{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\overline{Y}^2}^{\overline{Y}\text{ constant over }i}\\ &= \mu_{Y^2}-(\mu_Y)^2\\ &= E(Y^2)-\mu^2=\sigma^2\end{aligned}$$

Now, the multiplicand:

$$\lim_{n\to\infty}\frac{n}{n-1}=1$$

Thus:

$$\begin{aligned}\lim_{n\to\infty}\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2\right) &= 1\sigma^2\\ &= \sigma^2.\end{aligned}$$

□

Slutsky's Theorem:

If $F(U_n)\xrightarrow{p}\Phi$, and $F(W_n)\xrightarrow{p}1$, then $F\left(\frac{U_n}{W_n}\right)\xrightarrow{p}\Phi$.

Back to what we required...

Claim. $F\left(\frac{\overline{Y}-\mu}{s_u/\sqrt{n}}\right)\xrightarrow{p}\Phi$

Proof.

$$F\left(\frac{\overline{Y}-\mu}{s_u/\sqrt{n}}\right)=F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{s_u/\sigma}\right)\quad\Big|\cdot\frac{1/\sigma}{1/\sigma}$$

From CLT, $F\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)\xrightarrow{p}\Phi$

From above, $s_u^2\xrightarrow{p}\sigma^2$, which implies

$$\sqrt{\frac{s_u^2}{\sigma^2}}\xrightarrow{p}\sqrt{\frac{\sigma^2}{\sigma^2}}=1$$

because $g(x) = +\sqrt{\frac{x}{c}}$ is continuous in positive x and c, and hence we invoke the rule that $g(\hat{\theta}) \xrightarrow{p} g(\theta)$.

But $\sqrt{\frac{s_u^2}{\sigma^2}} = \frac{s_u}{\sigma}$ and so $\frac{s_u}{\sigma} \xrightarrow{p} 1$.

Therefore, by Slutsky's Theorem, $F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{s_u/\sigma}\right) \xrightarrow{p} \Phi$, as required. □

Implication:

For large samples, confidence intervals can be specified as:

$$\left[\hat{\theta} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{s_u}{\sqrt{n}}, \hat{\theta} + \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{s_u}{\sqrt{n}}\right]$$

For example, for a 95% confidence interval:

$$\left[\hat{\theta} - 1.96\frac{s_u}{\sqrt{n}}, \hat{\theta} + 1.96\frac{s_u}{\sqrt{n}}\right]$$

Example: Suppose mean household income in a particular zip code is 150K, with unbiased standard deviation of 100K and sample size of 10,000. What is the 95% confidence interval? 90% confidence interval?

95%: $\left[150K - 1.96\frac{100K}{100}, 150K + 1.96\frac{100K}{100}\right]$

90%: $\left[150K - \Phi^{-1}\left(\frac{.1}{2}\right)\frac{100K}{100}, 150K + \Phi^{-1}\left(\frac{.1}{2}\right)\frac{100K}{100}\right]$,

where $\Phi^{-1}\left(\frac{.1}{2}\right) \approx -1.64$.

Quant I: Lecture 9

Bernd Beber
New York University

Fall 2012

Some comments on the midterm exam

- Rescheduled for next Monday, November 12. The exam will be in class, 10am–11:50am.
- Open book, open notes, open problem sets, but not open Internet. You can bring a calculator, although it is not required and you should not need one. No smartphones or other electronic devices.
- Topics covered include discrete and continuous probability distributions, Bayes' rule, maximum likelihood, expected values and variances, etc.
- No confidence intervals. No Stata component.
- The next homework will be distributed the same day as the midterm and will be due on November 21. Answers have to be typed starting with this problem set!

An example question from last year's midterm

- Suppose you are sampling NYU undergraduates for a lab experiment, and you are able to randomly and independently select students with equal probability. Assume that each student can participate in the experiment as many times as he or she is randomly chosen. (It is sufficient to provide numeric expressions for probabilities in this problem; it is not necessary to compute the exact values.)
    1. Let's say that 10% of the undergraduate population at NYU identify as Republicans. What is the probability that a sample of 10 students will include at least one Republican?

2. Suppose you have drawn a sample of 10 students and found that it includes no Republicans. You now draw an additional sample of 990 students to increase your total sample size to 1000. What is the probability that the full sample will exactly reflect population proportions and include 10% Republicans? What is the probability that your full sample will include no more than 8% Republicans?

Now, back to where we were ...

Confidence intervals in small samples

We've already shown that the standardized sample average $\overline{Y}$ is distributed t with degrees of freedom $\nu = n - 1$, provided $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed Normal.

So, in small samples, the sampling distribution for $\overline{Y}$ is t for (approximately) normally distributed Y, and we can then show that the endpoints of a $100(1-\alpha)\%$ confidence interval are:

$$\overline{Y} \pm t_{\alpha/2,\nu} \frac{s_u}{\sqrt{n}}.$$

For comparison, in large samples we have:

$$\overline{Y} \pm z_{\alpha/2} \frac{s_u}{\sqrt{n}}.$$

Examples

1. The latest Gallup poll on presidential approval (October 26–28, 2012) shows that 51% of $n = 1,500$ respondents approve of "the job Barack Obama is doing as President." Let's identify the 95% confidence interval around this estimate.

   - Assume that the number of people Y who approve is binomially distributed with parameter p. Let's show that $\hat{p} = \dfrac{Y}{n}$ is an unbiased estimator for p.

   - Now, what is the variance of $\hat{p}$? (Note that $\hat{p} = \dfrac{Y}{n}$, not $\dfrac{\sum_i Y_i}{n}$.)

- Next, assume that the sampling distribution of $\hat{p}$ is approximately Normal. Why is this reasonable?

- Finally, what is the 95% confidence interval around the estimated approval rate of 51%?

- We refer to one half of the confidence interval as the study's margin of error. At which level of approval is the margin of error largest? At which is it smallest?

2. Another recent Gallup poll on the presidential candidates (October 27–28, 2012) showed that Barack Obama received a 60% favorable rating among registered voters, compared to 56% for Mitt Romney. The sample size was $n_1 = n_2 = 1,063$ in both cases. Let's answer the question whether this is a meaningful difference in ratings.

   - Let the estimated share of people who view each candidate favorably be given by $\hat{p}_1 = \dfrac{Y_1}{n_1}$ for Obama and $\hat{p}_2 = \dfrac{Y_2}{n_2}$ for Romney. Assume that Y_1 and Y_2 are identically and independently distributed Binomial, with parameters p_1 and p_2 respectively. Let's show that $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator for $p_1 - p_2$.

- What is the variance of $\hat{p}_1 - \hat{p}_2$?

- What is the 95% confidence interval around the estimated difference in favorability ratings?

Bootstrapping

What if we want to derive confidence intervals for other estimators? If we can't find an algebraic solution, computational approaches exist (bootstrap, jackknife).

Key idea for these approaches is to generate additional datasets just like the one we started with, compute the statistic of interest for each set of data, and use the resulting set of values as a proxy for the sampling distribution of the statistic.

But where are those other datasets going to come from? Let's pull ourselves up by our bootstraps! Treat the data we have as a proxy for all of the data we could have, and simulate additional samples. This is known as resampling.

Basic steps:

1. Draw a sample (usually of the same size as your dataset) from your data, with replacement. Since you are sampling with replacement, some observations may be included several times and some not at all.
2. Calculate and save the statistic of interest $\hat{\theta}$ for the sample you have drawn.
3. Repeat steps (1) and (2) a set number of times.
4. You now have many instances of $\hat{\theta}$, which you can use to see how this statistic is distributed.

Stata provides command `bootstrap`, which automates much of this process. See also the addendum to this lecture, which includes detailed example code. There will be a problem that asks you to perform a bootstrap on the next homework.

Hypothesis tests

Let's ask a slightly different question: Given some estimate $\hat{\theta}$, how sure am I that θ is not equal to θ_0? Often $\theta_0 = 0$, but can be anything.

Reject the null hypothesis $\theta = \theta_0$ when the confidence interval does not include θ_0. Reject in favor of alternative hypothesis such as $\theta \neq \theta_0$ (or perhaps $\theta > \theta_0$)

A hypothesis test consists of four elements:

1. A null hypothesis about a parameter,

$$H_0 : \theta = \theta_0.$$

   This is typically zero, or some other value that captures the "conventional wisdom."

2. An alternative hypothesis about a parameter,

$$H_a : \begin{cases} \theta > \theta_0 & \text{(upper tail)} \\ \theta < \theta_0 & \text{(lower tail)} \\ \theta \neq \theta_0 & \text{(two-tailed).} \end{cases}$$

3. A test statistic derived from an estimator $\hat{\theta}$ of the parameter.

4. A rejection region: The range of values of the test statistic for which the null is to be rejected in favor of the alternative.

Given our previously specified conditions for large- and small-sample estimators, we then have:

Large-sample α-level hypothesis test:

$$\text{Test statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

$$\text{Rejection region} : \begin{cases} \{z > z_\alpha\} & \text{(upper tail RR)} \\ \{z < -z_\alpha\} & \text{(lower tail RR)} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$$

Small-sample α-level hypothesis test:

$$\text{Test statistic: } T = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

$$\text{Rejection region}: \begin{cases} \{t > t_\alpha\} & \text{(upper tail RR)} \\ \{t < -t_\alpha\} & \text{(lower tail RR)} \\ \{|t| > t_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$$

For example, if $Y_1, Y_2, ..., Y_n$ is a random sample from the normal distribution with $E(Y_i) = \mu$:

$$H_0 : \mu = \mu_0$$

$$H_a : \begin{cases} \mu > \mu_0 & \text{(upper tail)} \\ \mu < \mu_0 & \text{(lower tail)} \\ \mu \neq \mu_0 & \text{(two-tailed)} \end{cases}$$

$$\text{Test statistic: } T = \frac{\overline{Y} - \mu_0}{s_u/\sqrt{n}}$$

$$\text{Rejection region}: \begin{cases} \{t > t_\alpha\} & \text{(upper tail RR)} \\ \{t < -t_\alpha\} & \text{(lower tail RR)} \\ \{|t| > t_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$$

Choosing the rejection region:

- RRs are associated with two kinds of error:
  - Type I error ("false positive") is made if H_0 is rejected when H_0 is actually true.

    $$\Pr(\text{Type I error}) = \alpha$$

  - Type II error ("false negative") is made if H_0 is accepted when H_a is actually true.

    $$\Pr(\text{Type II error}) = \beta$$

- α and β are two practical ways to measure the goodness of a statistical test. We call α the test's level of significance. We call the quantity $1 - \beta$ the test's statistical power. In the best of all worlds, we want a test's level of significance to be low and its power to be high. In reality, we always face a tradeoff between these two things.

Addendum: Bootstrapping

So far, we have calculated our uncertainty around a parameter estimate by figuring out the (approximate) distribution of the estimator: (1) We said that the sampling distribution of the sample average $\overline{Y}$ approximates a normal distribution as our sample becomes large, and (2) the sampling distribution of $\overline{Y}$ is given by a t distribution for normally distributed data, even in small samples.

But what if we are estimating parameters other than $\overline{Y}$? Sometimes we can derive properties of their sampling distributions, but what if we can't?

Another option is to take a computational rather than an algebraic approach. What this means is that we will simulate a whole bunch of datasets just like the one we started with, compute the statistic of interest for each of those datasets, and voilà, we have a sampling distribution for our statistic of interest. Draw a histogram of all of the values we computed for our statistic to see the distribution.

But wait, we only have one dataset. Where are all of those other datasets going to come from? Well, we're going to pull ourselves up by our bootstraps. The key idea is that we'll treat the data we have as a proxy for all of the data we could have. But if the data we have is like the data we could have, then we can generate additional samples from the data we got. This is known as resampling.

Here are the basic steps:

1. Draw a sample (usually of the same size as your dataset) from your data, with replacement. Since you are sampling with replacement, some observations may be included several times and some not at all.
2. Calculate and save the statistic of interest $\hat{\theta}$ for the sample you have drawn.
3. Repeat steps (1) and (2) a set number of times.
4. You now have many instances of $\hat{\theta}$, which you can use to see how this statistic is distributed.

Stata provides command `bootstrap`, which automates much of this process. (Many estimation commands also have an option `bootstrap`.) Let's do an example.

Let's look at life expectancies in a sample of countries in 1998, using one of Stata's exam-

ple datasets. Let's say we want to know the standard error of the estimated average life expectancy across countries:

```
sysuse lifeexp
bootstrap meanlexp = r(mean): sum lexp
```

After the colon, we tell `bootstrap` how we want the resampled data to be analyzed. Here, we want the variable for life expectancy to be summarized. Before the colon, we note which statistic we are interested in, i.e. which output we want to save from the program specified after the colon. Here, that is the average life expectancy. We give this statistic a name and call it `meanlexp`, but we could call it anything. That's it! Stata tells us that our estimate of the mean life expectancy has a standard error of about .52.

Let's do a slightly more complicated example. Let's say we are interested in the gap in life expectancies between the 25th and the 75th percentile (also known as the interquartile range). Obviously we can calculate what this gap is in the sample we have. But we may want to know how much this gap would vary if we drew a different sample of countries. How different an answer should we expect to get with a different sample? From what we've learned, we can't answer this question algebraically. But we can simulate additional datasets and answer this question computationally. Let's also take this opportunity to introduce additional features of Stata's programming syntax.

```
sysuse lifeexp
program drop gap
program define gap, rclass
syntax varname [if] [in], [Lower(integer 25) Upper(integer 75)]
    quietly centile `varlist' `if' `in', centile(`lower' `upper')
    return scalar diff = r(c_2) - r(c_1)
end

set seed 984785
bootstrap avggap = r(diff), reps(500): gap lexp, l(25) u(75)
```

The command `syntax` lets Stata know what input to expect when the program is called. Bracketed input is optional, for example `if` and `in` statements specifying which observations to use. The program `gap` has to be called with exactly one variable (`varname`), which is saved in macro `varlist`. Options `lower` and `upper` (with capital letters showing minimal

abbreviations) correspond to the lower and upper ends of the range we are interested in and are optional integer inputs with default values 25 and 75, respectively. The program returns a scalar, which is the difference in the value of `varname` between the upper and the lower end of the range.

Our `bootstrap` then calls this program for 500 resampled datasets and saves the interquartile range for each of them. We learn that our estimate of 7 years that divide the upper end from the lower end of the middle fifty comes with a bootstrapped standard error of about .87.

Note the similarities and differences between `simulate` and `bootstrap`: The former repeatedly executes instructions, while the latter repeatedly executes instructions on resampled datasets. The command `jackknife` is also related. Just like `bootstrap`, it generates additional samples from the existing data, but it does so by "shaving off" one observation per sample.

The kind of bootstrap shown here is not the only way to bootstrap. Another popular bootstrapping method is parametric, in which case we make an assumption about the distribution of a variable, draw repeatedly from this distribution, and use these draws to compute whatever statistic we are interested in.

Addendum: Sampling without replacement

The problem at the beginning of this lecture involves a sample drawn with replacement. What if we were sampling without replacement?

That is, let's answer the following question: Suppose you're drawing a sample of 1000 NYU students, without replacement. You've already drawn 10 students, none of which were Republicans. The probability with which a randomly drawn student is a Republican is .1. What is the probability that the sample will include exactly 80 Republicans?

There are two ways to answer this question:

1. The fact that we're sampling without replacement makes no difference. We use the same formula we used in the example with replacement, and the probability is

$$\binom{990}{80} \cdot .1^{80} \cdot .9^{910} \approx .005393.$$

   There are two ways to motivate this response:

   (a) Party identification is not a fixed but a probabilistic attribute of each potential subject. That is, the probability with which each particular student will identify as Republican when he or she is sampled is .1. If this is the case, the probability of drawing a Republican doesn't change as we take students out of the sampling frame (i.e. the population of students from which we are sampling). And if the probability of drawing a Republican doesn't change with the composition of the sampling frame, it doesn't matter if we sample with or without replacement. This could be a defensible assumption in some settings, but probably not here.

   (b) Sampling without replacement approximates sampling with replacement as the sampling frame becomes large relative to the sample. Let's say there are a million students to sample from. Then taking less than a thousand students from this population won't have a substantial impact on the probability with which we'll sample a Republican. But if this probability doesn't change (much) as we're sampling, we can proceed as if we were sampling with replacement.

   This raises the question: How large can the sample be relative to the sampling frame before we have to worry? As a rule of thumb, about 5%. If you're drawing

more than 5% of your sampling frame without replacement, don't treat the sampling design as if it was with replacement (and you can make use of something called a finite population correction to adjust estimates).

In our example, NYU has an undergrad enrollment of more than 22,000. So if all undergrads were included in our sampling frame, we'd sample about 4.5% of them, so we could plausibly claim that sampling without replacement approximates a design with replacement.

This approximation has real bite when it comes to multi-stage designs. For example, we could conduct a survey of students, where we sample school districts first, then schools within the selected districts, then classrooms within the selected schools, and then students within the selected classrooms. Clearly observations will be clustered in the sense that the responses of students from the same classroom or the same school will be correlated, but adjusting for such clustering usually diminishes the efficiency of our estimates. But it turns out that if the number of sampling units at the first stage (here, school districts) is large relative to the number of units sampled, we can estimate variances without adjusting for clustering at any of the subsequent stages. In other words, it can be very convenient to proceed as if sampling occurred with replacement even if it didn't. (I'm happy to send some intuition for this result to anyone who's interested. Note also that we'll want to sample a minimum number of clusters, usually at least thirty or so.)

2. The second way to answer the question is to say that we do not have enough information to solve this problem. If we do want to adjust for the fact that we are sampling without replacement, we need to know how the probability of selecting a Republican changes as we draw our sample. And that means we need to know how large the sampling frame is, which is not a piece of information that was provided in the prompt.

   Let's suppose we do know that the sampling frame of NYU undergraduates consists of 22,000 individuals, with 2200 Republicans among them. Then the probability of selecting exactly 80 Republicans in a sample of 990 students that are randomly drawn without replacement would be:

$$\frac{\binom{22000-2200}{910}\binom{2200}{80}}{\binom{22000}{990}} \approx .004998.$$

Now, there's one final wrinkle. The prompt said that we'd already drawn ten students, none of whom were Republicans. If we sampled these students also without replacement, we'll need to change the binomial coefficients slightly to

$$\frac{\binom{22000-10-2200}{910}\binom{2200}{80}}{\binom{22000-10}{990}} \approx .004948.$$

Quant I: Lecture 10

Bernd Beber
New York University

Fall 2012

Statistical power

Recall that β is the Type II error rate, and $1-\beta$ is a test's statistical power. So how do we compute β, given some significance level α?

We have

$$\beta = \Pr(\text{Accept } H_0 | H_a \text{ is true})$$

Suppose H_a is that $\theta = \theta_a$. Then

$$\beta = \Pr(\hat{\theta} < \theta_0 + z_\alpha \sigma_{\hat{\theta}} | \theta = \theta_a)$$

Let's see what we get if we assume that $\hat{\theta}$ is normally distributed with mean θ_a and standard deviation $\sigma_{\hat{\theta}}$. (When would this be plausible?)

We first standardize and then make use of our distributional assumption, so that

$$\begin{aligned}
\beta &= \Pr\left(\frac{\hat{\theta}-\theta_a}{\sigma_{\hat{\theta}}} < \frac{\theta_0 + z_\alpha \sigma_{\hat{\theta}} - \theta_a}{\sigma_{\hat{\theta}}} | \theta = \theta_a\right) \\
&= \Phi\left(\frac{\theta_0 + z_\alpha \sigma_{\hat{\theta}} - \theta_a}{\sigma_{\hat{\theta}}}\right) \\
&= \Phi\left(\frac{\theta_0 - \theta_a}{\sigma_{\hat{\theta}}} + z_\alpha\right).
\end{aligned}$$

(We don't need to assume $\theta_a > \theta_0$. Why?)

The test's power then is

$$\begin{aligned}
1-\beta &= 1 - \Phi\left(\frac{\theta_0 - \theta_a}{\sigma_{\hat{\theta}}} + z_\alpha\right) \\
&= 1 - \Phi\left(\frac{\theta_0 - \theta_a}{s_u/\sqrt{n}} - \Phi^{-1}(\alpha)\right).
\end{aligned}$$

We have specified θ_0, θ_a, and α, and we can compute s_u from our data. Usually we will want power to be at about 80% in a survey or experimental design.

P-values

In hypothesis testing, we

1. pick a significance level,
2. determine the critical value(s) of the test statistic associated with this significance level,
3. and compare the test statistic to the critical value to accept or reject the null hypothesis.

This is useful because it prevents us from shifting the goal posts of what constitutes a significant finding, but it does not provide much information beyond acceptance or rejection of the null hypothesis.

An alternative is to report the p-value, or attained significance level, which is the smallest level of significance α at which the observed data indicate that the null hypothesis should be rejected.

The smaller the p-value, the more compelling the evidence that the null hypothesis should be rejected.

Ordinary least squares

Suppose we have data (x_i, y_i) for $i = 1, ..., N$, and we want to express the bivariate relationship between X and Y. We'll refer to X as the independent variable and Y as the dependent or outcome variable. The labels suggest that we may be interested in a causal relationship between the two, but nothing that follows in this lecture either assumes or justifies this interpretation. For now, we'll just offer a model of the relationship between the two variables.

There are many ways to illustrate the relationship between X and Y; we could even do so in a scatter plot or table. What we'll do here is plot x_i and y_i and draw a line:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Deviations from the line are

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

These are positive and negative, so to have an expression that is monotonic in the magnitude of deviations:

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

We want to find β_0, β_1 such that our errors are minimal:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

First-order conditions:

$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \Rightarrow \sum_{i=1}^{N} e_i = 0 \tag{3}$$

$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \Rightarrow \sum_{i=1}^{N} x_i e_i = 0 \tag{4}$$

We switched to $\hat{\beta}_0, \hat{\beta}_1$ and e_i to denote that these are estimates.

From equation 3:

$$\begin{aligned} \sum_{i=1}^{N} y_i &= \sum_{i=1}^{N} \hat{\beta}_0 + \sum_{i=1}^{N} \hat{\beta}_1 x_i \\ &= n\hat{\beta}_0 + \left(\sum_{i=1}^{N} x_i \right) \hat{\beta}_1 \\ \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \text{(divide by } n\text{)} \end{aligned}$$

OLS regression line passes through the mean of the data. Not necessarily true without a constant $\hat{\beta}_0$!

We can solve for the constant term and get $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

From equation 4:

$$\begin{aligned}
\sum_{i=1}^{N} x_i y_i &= \sum_{i=1}^{N} x_i \hat{\beta}_0 + \sum_{i=1}^{N} x_i^2 \hat{\beta}_1 \\
&= \left(\sum_{i=1}^{N} x_i \right) \hat{\beta}_0 + \left(\sum_{i=1}^{N} x_i^2 \right) \hat{\beta}_1 \\
\text{Substituting for } \hat{\beta}_0 & \\
&= \left(\sum_{i=1}^{N} x_i \right) (\bar{y} - \hat{\beta}_1 \bar{x}) + \left(\sum_{i=1}^{N} x_i^2 \right) \hat{\beta}_1 \\
\text{Note that } \sum_{i=1}^{N} x_i &= n\bar{x}, \text{ and collect terms:} \\
\sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y} &= \hat{\beta}_1 \left(\sum_{i=1}^{N} x_i^2 - n\bar{x}^2 \right) \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - n\bar{x}^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}
\end{aligned}$$

Check:

$$\begin{aligned}
\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \bar{y} - \sum_{i=1}^{N} y_i \bar{x} + \sum_{i=1}^{N} \bar{x}\bar{y} \\
&= \sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

Are we at a minimum? We'll skip the proof, but this follows from the convexity of the function we are minimizing (another reason to look at squared deviations!)

So,

$$\hat{\beta}_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Some terminology:

- We "regress y on x."
- x is a "regressor."
- x is on the right-hand side (RHS) of the regression.

Some characteristics of OLS (by construction)

1. $\hat{\beta}_1$ is not defined if $\text{Var}(x) = 0$; $\hat{\beta}_1 = 0$ if $\text{Var}(y) = 0$
2. $\hat{\beta}_1$ is the change in $\hat{y}$ associated with a one-unit change in x (not necessarily causal!):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$$\frac{\partial \hat{y}}{\partial x} = \hat{\beta}_1$$

3. The sum of residuals is zero, $\sum_{i=1}^{N} e_i = 0$
   - From equation 3: note that this reflects the mechanics of fitting a line. It does not reflect the property of our data; in fact, we'll assume that our data is appropriate!
4. The previous also implies $\bar{e} = 0$
5. Sample covariance between the regressor and residuals is zero, $\text{Cov}(x, e) = 0$

$$\begin{aligned}
\text{Cov}(x, e) &= \frac{\sum_i (e_i - \bar{e})(x_i - \bar{x})}{n} \\
&= \frac{\sum_i e_i (x_i - \bar{x})}{n} \\
&= \frac{\sum_i e_i x_i - \sum_i e_i \bar{x}}{n} \\
&= \frac{0 - \bar{x} \sum_i e_i}{n} \\
&= 0
\end{aligned}$$

   - Residuals are not correlated with the regressor! Not a freebie, but imposed by fitting a line.

6. Sample covariance between fitted values and residuals is zero, $\text{Cov}(\hat{y}, e) = 0$

$$\begin{aligned}
\text{Cov}(\hat{y}, e) &= \frac{\sum_i (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}})}{n} \\
&= \frac{\sum_i e_i \hat{y}_i - \sum_i e_i \bar{\hat{y}}}{n} \quad \text{(because } \bar{e} = 0\text{)} \\
&= \frac{\sum_i e_i \hat{y}_i}{n} \quad \text{(because } \sum_i e_i = 0 \text{ and } \bar{\hat{y}} \text{ constant)} \\
&= \frac{\sum_i e_i (\beta_0 + \beta_1 x_i)}{n} = 0 \\
&\quad \text{(because } \sum_i e_i = 0 \text{ and } \beta_0 \text{ const.,} \\
&\quad \text{and } \sum_i x_i e_i = 0 \text{ and } \beta_1 \text{ const.)}
\end{aligned}$$

7. Point $(\bar{x}, \bar{y})$ is always on the regression line. Follows from equation 3 and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$.

8. Sample average of fitted values equals sample average of observed y_i, $\bar{y} = \bar{\hat{y}}$

$$\begin{aligned}
y_i &= \hat{y}_i + e_i \\
\sum_i y_i &= \sum_i \hat{y}_i + \sum_i e_i \\
\frac{\sum_i y_i}{n} &= \frac{\sum_i \hat{y}_i}{n} \quad \text{(because } \bar{e} = 0\text{)} \\
\Rightarrow \bar{y} &= \bar{\hat{y}}
\end{aligned}$$

Model fit

How well does the line fit? Fitting a line decomposes y_i into two parts: $\hat{y}_i$ and e_i. By construction, these are uncorrelated (see 4. above). Define the following:

- Total sum of squares:
$$SST = \sum_i (y_i - \bar{y})^2$$
How much does y vary about its mean?

- Explained sum of squares:
$$SSE = \sum_i (\hat{y}_i - \bar{y})^2$$

- We already know the error sum of squares:

$$SSR = \sum_i e_i^2$$

We can show:

$$SST = SSE + SSR$$

Proof.

$$\begin{aligned} SST &= \sum_i (y_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i (e_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i e_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i e_i(\hat{y}_i - \bar{y}) \\ &= SSR + SSE \end{aligned}$$

□

Note that we have (from 4.) $\text{Cov}(\hat{y}, e) = \frac{\sum_i (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}})}{n} = 0$, hence $\sum_i e_i(\hat{y}_i - \bar{y}) = 0$, and the last line follows.

We can use this to construct a measure of goodness of fit:

$$R^2 = \frac{SSE}{SST}$$

- Ranges from 0 to 1
- Gives the proportion of the variation in y explained by our model
- Note that $R^2 = 1 - \frac{SSR}{SST}$
- Name derives from the fact that in the bivariate context $R^2 = r^2$, where r is the sample correlation coefficient (i.e. ρ with substituted estimates)

Quant I: Lecture 11

Bernd Beber
New York University

Fall 2012

Multiple regression

What if we have more than one right-hand side variable x?

Suppose we have k regressors and N observations. We add normal equations

$$\frac{\partial \sum_i e_i^2}{\partial \hat{\beta}_j} = 0 \text{ for } j \in \{0, \ldots, k\}.$$

This is easier in matrix notation!

The sum of squared residuals is

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}})^2,$$

where $\mathbf{x}_i$ and $\hat{\boldsymbol{\beta}}$ (or alternatively $\vec{x}_i$ and $\vec{\beta}_i$) denote the column vectors

$$\mathbf{x}_i = \begin{bmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \text{ and } \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}.$$

Hence, $x_0 = 1$, usually. (Why?)

Minimization problem for OLS:

$$\min_{\hat{\boldsymbol{\beta}}} \underbrace{\mathbf{e}'\mathbf{e}}_{(1\times N)(N\times 1)} = \underbrace{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'}_{(1\times N)} \underbrace{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}_{(N\times 1)-(N\times k)(k\times 1)}$$

where

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \text{ and } \mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N0} & x_{N1} & \cdots & x_{Nk} \end{bmatrix}.$$

Recall that $(\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}'$ and $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, and so expanding gives

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

Note that $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ is a scalar, and hence $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = (\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}})' = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$. Then

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

The first order conditions now become

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

The solution then satisfies the least squares normal equations

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists (i.e. assuming full rank), the solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Example:

Consider

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \text{ and } \mathbf{x} = \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix}.$$

What is $\hat{\beta}_1$?

First, solve without matrix algebra. We have $\bar{y} = 0$ and $\bar{x} = 0$.

From last lecture we have $\hat{\beta}_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$, that is

$$\hat{\beta}_1 = \frac{2+4}{4+4} = \frac{6}{8} = \frac{3}{4}.$$

In matrix notation, we solve $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Add a column of ones for $\hat{\beta}_0$ and perform matrix matrix multiplication to get

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 1 & -2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 0 \\ 0 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 6 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} \begin{bmatrix} 0 \\ 6 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ \frac{3}{4} \end{bmatrix}.
\end{aligned}$$

Gauss-Markov assumptions

We can always fit a line through a cloud of data. But if we are willing to make the following five assumptions, then OLS is the best (i.e. most efficient) linear unbiased estimator. That is, OLS is BLUE. This result is known as the Gauss-Markov theorem.

Assumptions:

1. In the population, dependent variable $\mathbf{y}$ is a linear function of unknown parameters $\boldsymbol{\beta}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

2. Our data is a random sample of size N from this population.

3. $\mathbf{X}$ is $N \times k$ with rank k (i.e. has full rank). The absence of linear dependence ensures invertibility.

4. Zero conditional mean:

$$\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$$

   This is also known as the assumption of strict exogeneity. By the law of iterated expectations, we can further claim that the unconditional mean is zero:

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathrm{E}_{\mathbf{X}}(\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X})) = \mathbf{0}$$

5. Spherical errors:

$$\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_N$$

   We assume homoskedasticity and uncorrelated errors across observations. Note that

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) &= \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) - \mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X})\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X})' \\ &= \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})\end{aligned}$$

   We can again make an unconditional claim by way of the law of iterated expectations:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\varepsilon}) &= \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \mathrm{E}_{\mathbf{X}}(\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})) \\ &= \sigma^2\mathbf{I}_N\end{aligned}$$

What does $\sigma^2\mathbf{I}_N$ look like?

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} \mathrm{E}(\varepsilon_1^2) & \mathrm{E}(\varepsilon_1\varepsilon_2) & \cdots & \mathrm{E}(\varepsilon_1\varepsilon_N) \\ \mathrm{E}(\varepsilon_2\varepsilon_1) & \mathrm{E}(\varepsilon_2^2) & \cdots & \mathrm{E}(\varepsilon_2\varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\varepsilon_N\varepsilon_1) & \mathrm{E}(\varepsilon_N\varepsilon_2) & \cdots & \mathrm{E}(\varepsilon_N^2) \end{bmatrix},$$

where

$$\mathrm{Var}(\varepsilon_i^2) = \mathrm{E}[(\varepsilon_i - \mathrm{E}(\varepsilon_i))^2] = \mathrm{E}(\varepsilon_i^2),$$

and

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}[(\varepsilon_i - \mathrm{E}(\varepsilon_i))(\varepsilon_j - \mathrm{E}(\varepsilon_j))] = \mathrm{E}(\varepsilon_i\varepsilon_j).$$

That is, we can interpret

$$\sigma^2\mathbf{I}_N = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

as the variance-covariance matrix of the error terms.

Unbiasedness of the OLS estimator

Claim. *$\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased given the Gauss-Markov assumptions (1.-4.).*

Proof. We need to show $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. We have

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.\end{aligned}$$

Taking expectations yields

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) \\ &= \boldsymbol{\beta}.\end{aligned}$$

This means our coefficients are unbiased if our covariates $\mathbf{X}$ are fixed in repeated samples. For the unconditional expectation of $\hat{\boldsymbol{\beta}}$, we again rely on the law of iterated expectations and "average" over all of the samples of $\mathbf{X}$ that we could observe, so that

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}_{\mathbf{X}}[\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X})] \\ &= \boldsymbol{\beta} + \mathrm{E}_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X})] \\ &= \boldsymbol{\beta}.\end{aligned}$$

□

Quant I: Lecture 12

Bernd Beber
New York University

Fall 2012

Where we are:

- We derived $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$
- We showed unbiasedness $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$
- Next: Derive variance of $\hat{\boldsymbol{\beta}}$

What is the variance of $\hat{\boldsymbol{\beta}}$?

First, what is the variance-covariance matrix for a vector of random variables in general?

$$\mathrm{Var}(\mathbf{z}) = \mathrm{E}((\mathbf{z} - \mathrm{E}(\mathbf{z}))(\mathbf{z} - \mathrm{E}(\mathbf{z}))')$$

for $\mathbf{z}$ of size $k \times 1$.

$$\mathrm{E}\left(\begin{bmatrix} z_1 - \bar{z}_1 \\ z_2 - \bar{z}_2 \\ \vdots \\ z_k - \bar{z}_k \end{bmatrix} \begin{bmatrix} z_1 - \bar{z}_1 & \cdots & z_k - \bar{z}_k \end{bmatrix}\right)$$

$$= \mathrm{E}\left(\begin{bmatrix} (z_1 - \bar{z}_1)^2 & (z_1 - \bar{z}_1)(z_2 - \bar{z}_2) & \cdots & (z_1 - \bar{z}_1)(z_k - \bar{z}_k) \\ \vdots & \vdots & \ddots & \vdots \\ (z_k - \bar{z}_k)(z_1 - \bar{z}_1) & \cdots & \cdots & (z_k - \bar{z}_k)^2 \end{bmatrix}\right)$$

Similarly, we have $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})')$, where

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \end{aligned}$$

So, $\text{Var}(\hat{\boldsymbol{\beta}}) = \text{E}[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})']$.

By the transpose rule $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, applied iteratively:

$$((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})' = (\boldsymbol{\varepsilon}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]')$$

Note that the transpose of the inverse is equal to the inverse of the transpose:

$$[(\mathbf{X}'\mathbf{X})^{-1}]' = [(\mathbf{X}'\mathbf{X})']^{-1}$$

But the transpose of a symmetric matrix returns the matrix itself:

$$[(\mathbf{X}'\mathbf{X})']^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$$

Hence:

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{E}[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})] \\ &= \text{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \end{aligned}$$

Passing through the (conditional) expectations operator:

$$\begin{aligned} \text{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_N\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by Assumption 5: } \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_N\text{)} \\ &= \sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \quad \text{(since } \sigma^2 \text{ is a constant)} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

What about $\text{Var}(\hat{\boldsymbol{\beta}})$ instead of $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$?

From the law of total variance, we have $\text{Var}(Y) = \text{E}(\text{Var}(Y|X)) + \text{Var}(\text{E}(Y|X))$, and hence

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{E}_{\mathbf{X}}(\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})) + \text{Var}_{\mathbf{X}}[\text{E}(\hat{\boldsymbol{\beta}}|\mathbf{X})].$$

But $\text{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ for all $\mathbf{X}$, so $\text{Var}_{\mathbf{X}}[\text{E}(\hat{\boldsymbol{\beta}}|\mathbf{X})] = 0$ and

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

But we don't know σ^2, the true variance of the unit error!

Finding an estimate for σ^2

Again, the intuitive estimator is not necessarily unbiased!

We have

$$\sigma^2 = \mathrm{E}((\boldsymbol{\varepsilon} - \mathrm{E}(\boldsymbol{\varepsilon}))'(\boldsymbol{\varepsilon} - \mathrm{E}(\boldsymbol{\varepsilon}))) = \mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}),$$

so intuitively we might want to simply take $\mathbf{e}$ as an estimate for $\boldsymbol{\varepsilon}$ and compute $\frac{\sum e_i^2}{N}$.

But $\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) \neq \mathrm{E}(\mathbf{e}'\mathbf{e})$. Consider the squared error for a particular observation i:

$$\begin{aligned} e_i^2 &= (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 \\ &= (\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2. \end{aligned}$$

Let's take expectations:

$$\begin{aligned} \mathrm{E}(e_i^2) &= \mathrm{E}\left((\varepsilon_i + \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\right) \\ &= \mathrm{E}\left(\varepsilon_i^2 + (\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 + 2\varepsilon_i(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\right) \\ &= \mathrm{E}(\varepsilon_i^2) + \mathrm{E}[(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2] \end{aligned}$$

We know that $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, but that's not enough.

So, let's find an unbiased estimator. Note that:

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{I}_N - \underbrace{\mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{k \times k}\mathbf{X}'}_{N \times N})\mathbf{y} \end{aligned}$$

$\mathbf{M}$ is symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{MM} = \mathbf{M}$).

Given

$$\mathbf{I}_N = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}}_{N \times N},$$

let's show that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is idempotent:

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Show that $(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ is idempotent:

$$\begin{aligned} & (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \\ = \; & \mathbf{I}_N - \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{N \times N}\mathbf{I}_N - \mathbf{I}_N\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ = \; & \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

Note also $\mathbf{MX} = 0$. That is, $(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X}$.

So $\mathbf{e} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{My}$, and hence

$$\begin{aligned} \Rightarrow \underbrace{\mathbf{e}'\mathbf{e}}_{1 \times 1} &= \mathbf{y}'\mathbf{M}'\mathbf{My} && \text{(transpose rule)} \\ &= \mathbf{y}'\mathbf{My} && \text{(because } \mathbf{M} \text{ is symmetric, idempotent)} \\ &= \mathbf{y}'\mathbf{e} = \mathbf{e}'\mathbf{y} \end{aligned}$$

Note, $\mathbf{e}'\mathbf{e}$ is the sum of squared standard errors.

Now, let's relate $\mathbf{e}$ to $\boldsymbol{\varepsilon}$, since we want to estimate the variance of true error $\boldsymbol{\varepsilon}$.

$$\begin{aligned} \mathbf{e} &= \mathbf{My} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) && \text{(we use } \boldsymbol{\beta} \text{ so we can use } \boldsymbol{\varepsilon}) \\ &= \mathbf{MX}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} \\ &= \mathbf{M}\boldsymbol{\varepsilon} && (\mathbf{MX}\boldsymbol{\beta} = 0 \text{ because } \mathbf{MX} = 0) \end{aligned}$$

So, $\mathrm{E}(\mathbf{e}'\mathbf{e}) = \mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})$.

We'll simplify the RHS using the trace operator, which gives the sum of the diagonal elements. E.g.,

$$\mathrm{tr}\begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} = 4$$

This is useful because of the following properties:

1. $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CBA})$
2. $\mathrm{tr}(c\mathbf{A}) = c\mathrm{tr}(\mathbf{A})$
3. $\mathrm{tr}(\mathbf{A} - \mathbf{B}) = \mathrm{tr}(\mathbf{A}) - tr(\mathbf{B})$

The trace of a 1×1 matrix is equal to the matrix itself, and $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ is 1×1. Hence:

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) &= \mathrm{E}(\mathrm{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})) \\
&= \mathrm{E}(\mathrm{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \quad \text{(because of property 1)} \\
&= \mathrm{tr}(\mathbf{M}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \\
&= \mathrm{tr}(\mathbf{M}\sigma^2\mathbf{I}_N) \\
&= \sigma^2\mathrm{tr}(\mathbf{M}) \quad \text{(because of property 2)}
\end{aligned}$$

Hence, $\mathrm{E}(\mathbf{e}'\mathbf{e}) = \sigma^2\mathrm{tr}(\mathbf{M})$. We're getting closer to σ^2! What is $\mathrm{tr}(\mathbf{M})$?

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}) &= \mathrm{tr}(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \mathrm{tr}(\mathbf{I}_N) - \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \quad \text{(because if property 3)} \\
&= N - \mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \quad \text{(because if property 1)} \\
&= N - \mathrm{tr}(\mathbf{I}_k) \\
&= N - k
\end{aligned}$$

So $\mathrm{E}(\mathbf{e}'\mathbf{e}) = \sigma^2(N-k)$, and hence

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$$

is an unbiased estimator for σ^2 because

$$\mathrm{E}(s^2) = \mathrm{E}\left(\frac{\mathbf{e}'\mathbf{e}}{N-k}\right) = \frac{\mathrm{E}(\mathbf{e}'\mathbf{e})}{N-k} = \sigma^2$$

Hence,

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

Quant I: Lecture 13

Bernd Beber
New York University

Fall 2012

Efficiency of OLS coefficients

We can show that no other unbiased linear estimator $\tilde{\boldsymbol{\beta}}$ exists for which $\text{Var}(\tilde{\boldsymbol{\beta}}) < \text{Var}(\hat{\boldsymbol{\beta}})$. The notion of $\tilde{\boldsymbol{\beta}}$ as a linear estimator refers to the fact that the estimator is linear in $\mathbf{y}$, *not* the first Gauss-Markov assumption that $\mathbf{y}$ is linear in $\boldsymbol{\beta}$. We can relax the former assumption, even as we maintain the latter one.

Since we are considering only estimators that are linear in $\mathbf{y}$, any $\tilde{\boldsymbol{\beta}}$ can be computed as:

$$\underbrace{\mathbf{C}\mathbf{y}}_{(k \times N)(N \times 1)}$$

If $\tilde{\boldsymbol{\beta}}$ is unbiased, we have

$$\begin{aligned} \text{E}(\tilde{\boldsymbol{\beta}}) &= \boldsymbol{\beta} \\ \Rightarrow \text{E}(\mathbf{C}\mathbf{y}) &= \boldsymbol{\beta} \\ \Rightarrow \text{E}(\mathbf{C}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) &= \boldsymbol{\beta} \\ \Rightarrow \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C} \cdot \text{E}(\boldsymbol{\varepsilon}) &= \boldsymbol{\beta}, \end{aligned}$$

which implies $\mathbf{C}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ and hence $\mathbf{C}\mathbf{X} = \mathbf{I}_k$.

Use this to calculate the variance

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\beta}}) &= \text{E}((\tilde{\boldsymbol{\beta}} - \text{E}(\tilde{\boldsymbol{\beta}}))(\tilde{\boldsymbol{\beta}} - \text{E}(\tilde{\boldsymbol{\beta}}))') \\ &= \text{E}((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})') \\ &= \text{E}((\mathbf{C}\mathbf{y} - \boldsymbol{\beta})(\mathbf{C}\mathbf{y} - \boldsymbol{\beta})') \\ &= \text{E}((\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} - \boldsymbol{\beta})(\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} - \boldsymbol{\beta})') \\ &= \text{E}((\mathbf{C}\boldsymbol{\varepsilon})(\mathbf{C}\boldsymbol{\varepsilon})') \\ &= \mathbf{C}\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{C}' \qquad \text{(transpose rule)} \\ &= \sigma^2 \underbrace{\mathbf{C}\mathbf{C}'}_{k \times k} \end{aligned}$$

How does this compare to $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$?

From $\text{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{C}\mathbf{C}'$, we can write

$$\sigma^2(\underbrace{\mathbf{D}}_{k \times N} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{k \times N})(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'$$

where $\mathbf{D}$ is the matrix that we add to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in order to obtain any arbitrary $\mathbf{C}$.

Recall that $\mathbf{C}\mathbf{X} = \mathbf{I}_k$, so

$$\begin{aligned}(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} &= \mathbf{I}_k \\ \Rightarrow \mathbf{D}\mathbf{X} + \mathbf{I}_k &= \mathbf{I}_k \\ \Rightarrow \mathbf{D}\mathbf{X} &= 0.\end{aligned}$$

Hence,

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D}' + ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')') \qquad \text{(because } (\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}') \\ &= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}') \\ &= \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})') \qquad \text{(because } \mathbf{D}\mathbf{X} = 0) \\ &= \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}) \\ &= \text{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}'\end{aligned}$$

Diagonal elements of $\text{Var}(\tilde{\boldsymbol{\beta}})$ and $\text{Var}(\hat{\boldsymbol{\beta}})$ are variance estimates. That is,

$$\text{Var}(\tilde{\beta}_i) = (\text{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}')_{ii}.$$

So, now we just need to show that the diagonal elements of $\sigma^2\mathbf{D}\mathbf{D}'$ are non-negative. But each element of $\mathbf{D}\mathbf{D}'$ on the diagonal is made up of d_{ii}^2, which cannot be negative. And if $\mathbf{D} = 0$, then $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so $\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y} = \hat{\boldsymbol{\beta}}$. Therefore, OLS is BLUE!

Sampling distribution of OLS coefficients

Recall the solving for $\hat{\boldsymbol{\beta}}$ yielded

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.\end{aligned}$$

That is, we can think of $\hat{\boldsymbol{\beta}}$ as a linear function of $\boldsymbol{\varepsilon}$, with $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ as a coefficient on this variable. (Remember that $\boldsymbol{\beta}$ is constant.)

Can we then say something about the distribution of $\hat{\boldsymbol{\beta}}$? Yes, if we are willing to make an assumption about the distribution of $\boldsymbol{\varepsilon}$. In particular, remember that a linear transformation of a normally distributed random variable is also normally distributed.

So, let's assume $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(0, \sigma^2\mathbf{I}_N)$, i.e. a normally distributed error term.

From the matrix version of the rule $\text{Var}(b+aY) = a^2\text{Var}(Y)$, we have that for any vector of random variables

$$\underbrace{\mathbf{y}}_{(N\times 1)} \sim N(\underbrace{\boldsymbol{\mu}}_{(N\times 1)}, \underbrace{\boldsymbol{\Sigma}}_{(N\times N)}),$$

$$\underbrace{\mathbf{b}}_{(N\times 1)} + \underbrace{\mathbf{A}\mathbf{y}}_{(N\times N)(N\times 1)} \sim N(\mathbf{b}+\mathbf{A}\boldsymbol{\mu}, \underbrace{\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'}_{(N\times N)}).$$

So, we have

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')')$$

which simplifies to

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{-1})') \qquad \text{(transpose rule)}$$

and hence

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

If $\mathbf{X}$ is reasonably non-stochastic, $\hat{\boldsymbol{\beta}}$ are distributed multivariate normal. Each element is also distributed normal:

$$\hat{\beta}_i|\mathbf{X} \sim N(\beta_i, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ii})$$

This is the case under the classical linear model (CLM) assumptions, which are the Gauss-Markov assumptions plus the assumption of a normally distributed error term.

Under CLM, the OLS estimator is the minimum variance unbiased estimator—in general, not just relative to other linear estimators! For example, we can estimate coefficients through maximum likelihood, which is not linear in $\mathbf{y}$, but we won't gain efficiency.

Confidence intervals

Now we can move to inference. As in the univariate case, let's construct a confidence interval.

Recall what we had derived for the sampling distribution of the sample average under similar conditions:

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

By standardizing, we obtained

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and used a standard normal distribution for hypothesis testing.

Given confidence level α, we computed confidence intervals of the form

$$\left[\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Now, we have

$$\hat{\beta}_i|\mathbf{X} \sim N(\beta_i, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ii}),$$

and if we assume that $\mathbf{X}$ is fixed in repeated samples,

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ii}).$$

So,

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim N(0,1)$$

Can we use this for a hypothesis test? Why does `Stata` report regressions with t-statistics, not z-statistics?

We don't know σ! We have to estimate it, which means it varies across samples. In particular, we showed that

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$$

is an unbiased estimator for σ^2, where a new $\mathbf{e}$ is drawn for each sample.

Claim. $\frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim t_{N-k}$

Proof. Recall that $\frac{Z}{\sqrt{W/\nu}} \sim t_\nu$ if $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$, with independent Z and W and degrees of freedom ν.

We also know that $Z = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim N(0,1)$, so let's divide by a random variable that's distributed chi-squared.

In particular, suppose $\frac{(N-k)s^2}{\sigma^2} \sim \chi^2_{N-k}$. Then

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}}\sqrt{\frac{\sigma^2(N-k)}{(N-k)s^2}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}}$$

is distributed t_{N-k}.

All we have left to show is that $\frac{(N-k)s^2}{\sigma^2} \sim \chi^2_{N-k}$.

Since $s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$, we have

$$\frac{(N-k)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2}.$$

Recall

$$\begin{aligned}
\mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\
&= \mathbf{M}\mathbf{y} \qquad (\text{“residual maker” } \mathbf{M}) \\
&= \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} \\
&= \mathbf{M}\boldsymbol{\varepsilon}
\end{aligned}$$

which gives us

$$\begin{aligned}
\frac{(N-k)s^2}{\sigma^2} &= \frac{\boldsymbol{\varepsilon}'\mathbf{M}\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} \qquad (\text{transpose rule and } \mathbf{M} = \mathbf{M}') \\
&= \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'\mathbf{M}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right). \qquad (\text{because } \mathbf{M}\mathbf{M} = \mathbf{M} \text{ and } \sigma \text{ constant})
\end{aligned}$$

But we know $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_N)$ and so $\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim N(0,1)$.

Then if $\mathbf{M} = \mathbf{I}_N$,

$$\underbrace{\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'}_{(1 \times N)} \underbrace{\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)}_{(N \times 1)} \sim \chi^2_N$$

because we're looking at a sum of N squared normally distributed variables. In fact, for any idempotent matrix $\mathbf{M}$,

$$\begin{aligned} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) &\sim \chi^2_{\mathrm{tr}(\mathbf{M})} \\ &\sim \chi^2_{N-k}. \end{aligned}$$

So $\frac{(N-k)s^2}{\sigma^2} \sim \chi^2_{N-k}$, and hence

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim t_{N-k}$$

□

By default, `Stata` reports a t-statistic for $H_0 : \beta_i = 0$.

Sampling distribution in large samples

Again, we required the classic linear model assumptions (CLM) for this, i.e. the five Gauss-Markov assumptions and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_N)$.

(For bias computations, as on the homework, we don't need normally distributed error.)

Does that mean we have to abandon this approach unless we have normal errors? Fortunately, no! The Central Limit Theorem comes to the rescue and yields

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \overset{a}{\sim} N(0,1)$$

where s is a consistent estimator for σ_i.

We say that $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed, i.e. for sufficiently large samples. This is approximate instead of exact inference.

Since it is approximate, the standardized sampling distribution is now normal, not t. But since the t approaches the normal distribution for large N, we often do exactly the same thing with normal or non-normal errors.

How large is sufficiently large? Some will say 30, but it depends on the error distribution.

This application of the Central Limit Theorem does not do away with Gauss-Markov assumptions $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$ and $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_N$.

For example, a bounded range of the dependent variable implies bounds for residuals.

In this case, $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$ may be violated (if $\mathbf{X}$ does not predict $\mathbf{y}$).

Or $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_N$ may be violated (if $\mathbf{X}$ does predict $\mathbf{y}$).

Testing a hypothesis about the equality of coefficients

Suppose the true model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

and

$$H_0 : \beta_2 = \beta_3 \rightarrow \beta_2 - \beta_3 = 0.$$

Test this hypothesis with the estimator $(\hat{\beta}_2 - \hat{\beta}_3)$ and t-statistic

$$t = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{\text{se}(\hat{\beta}_2 - \hat{\beta}_3)}$$

What is the standard error? $\hat{\beta}_2$ and $\hat{\beta}_3$ are random variables, so apply the rule

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y),$$

which yields

$$\text{se}(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{\text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2\text{Cov}(\hat{\beta}_2, \hat{\beta}_3)}.$$

How do we get these variances and covariance? They are elements of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$!

Examples:

- Education in years in junior college as opposed to 4-year university on income;
- Effect of having one as opposed to two older siblings on educational outcomes;
- Effect of $1000 worth of in-person as opposed to phone canvassing on turnout.

We could also solve this problem by rearranging terms in the model we estimate.

For example, to estimate $\beta_2 - \beta_3$, define parameter $\theta = \beta_2 - \beta_3$, which yields $\beta_2 = \theta + \beta_3$.

So

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \boldsymbol{\varepsilon}$$

becomes

$$\begin{aligned}\mathbf{y} &= \beta_0 + \beta_1\mathbf{x}_1 + (\theta + \beta_3)\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \boldsymbol{\varepsilon} \\ &= \beta_0 + \beta_1\mathbf{x}_1 + \theta\mathbf{x}_2 + \beta_3(\mathbf{x}_2 + \mathbf{x}_3) + \boldsymbol{\varepsilon}\end{aligned}$$

We then generate the variable $\mathbf{z} = \mathbf{x}_2 + \mathbf{x}_3$ and run

```
reg y x1 x2 z
```

where the coefficient on `x2` is our quantity of interest.

Testing a hypothesis about any single linear combination of coefficients

If our null hypothesis sets a linear combination of $\boldsymbol{\beta}$ equal to a particular value, we can express this hypothesis in terms of a restriction on the coefficient vector.

For example, is the effect of \$2000 of in-person canvassing the same as \$1000 of mailings plus a \$1000 of phone canvassing?

Suppose the true model is

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is vote turnout, $\mathbf{x}_1$ in-person contact, $\mathbf{x}_2$ phone canvassing, and $\mathbf{x}_3$ direct mail, where all the regressors are in \$1000 of expenditures.

We could then test restriction

$$\begin{bmatrix}0 & 2 & -1 & -1\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3\end{bmatrix} = 0$$

or $\mathbf{r}'\boldsymbol{\beta} = q$, where $\mathbf{r} = \begin{bmatrix}0 \\ 2 \\ -1 \\ -1\end{bmatrix}$ and $q = 0$. We calculate the test statistic as $t = \frac{\hat{q}-q}{\text{se}(\hat{q})}$, where $\hat{q} = \mathbf{r}'\hat{\boldsymbol{\beta}}$.

In order to compute the t-statistic, we still need $\text{se}(\hat{q})$. But $\hat{q}$ is a linear function of $\hat{\boldsymbol{\beta}}$, so since $\text{Var}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$, we have $\text{Var}(\hat{q}) = \mathbf{r}'(s^2(\mathbf{X}'\mathbf{X})^{-1})\mathbf{r}$ by applying the matrix version of $\text{Var}(aX) = a^2\text{Var}(X)$.

So, the test statistic is

$$t = \frac{\hat{q} - q}{\text{se}(\hat{q})} = \frac{\mathbf{r}'\hat{\boldsymbol{\beta}} - q}{\sqrt{\mathbf{r}'(s^2(\mathbf{X}'\mathbf{X})^{-1})\mathbf{r}}}$$

Testing a hypothesis about multiple linear restrictions

Following the same principle as above, we can construct an $m \times k$ restriction matrix (instead of a restriction vector of length k), where m is the number of linear restrictions we wish to impose. See e.g. Fox, pp. 202–203.

We'll focus here on the most common test involving multiple linear restrictions, namely whether a group of variables has no effect on the dependent variable. (What would the restriction matrix look like in this case?)

In essence, we'll ask whether there is a significant difference between the variance that we can explain when we include this group of variables and the variance that we can explain without it. This is the intuition behind the likelihood-ratio principle.

Suppose our unrestricted model u is

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_{k-1} \mathbf{x}_{k-1} + \boldsymbol{\varepsilon}$$

and our null hypothesis is

$$H_0 : \beta_{k-J} = \beta_{k-J+1} = \ldots = \beta_{k-1} = 0$$

where J is the number of restrictions.

Then the restricted model r is

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_{k-J-1} \mathbf{x}_{k-J-1} + \boldsymbol{\varepsilon}$$

We can then define the F-statistic,

$$F \equiv \frac{(SSR_r - SSR_u)/J}{SSR_u/(N-k)},$$

which follows an F-distribution with J numerator degrees of freedom and $N-k$ denominator degrees of freedom. That is, $F \sim F_{J,N-k}$.

Since $SSR_r \geq SSR_u$, this statistic is always non-negative. It is the ratio of two independent chi-square random variables, divided by their respective degrees of freedom. (Recall that a chi-square random variable is the sum of the squares of independent standard normal random variables, which is what an SSR is.)

If H_0 is rejected, we say that the excluded variables are jointly significant. If we fail to reject H_0, we say that the variables are jointly insignificant.

Quant I: Lecture 14

Bernd Beber
New York University

Fall 2012

Summary of hypothesis tests

In general, we standardize our point estimate to get a test statistic that follows some known distribution.

$$\text{test statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error of estimate}}$$

- Hypothesis about a single coefficient

  Test statistic:

$$\begin{aligned} t &= \frac{\hat{\beta}_i - \beta_i}{\text{s.e.}(\hat{\beta}_i)} \\ &= \frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \\ &= \frac{\hat{\beta}_i - \beta_i}{\frac{\mathbf{e}'\mathbf{e}}{N-k}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \end{aligned}$$

  P-value: The p-value (in a two-tailed test) gives the probability of observing a point estimate as far away from β_i as $\hat{\beta}_i$, given that β_i is in fact true. That is, $\Pr(|T| > |t|)$ for random variable T distributed t with $N - k$ degrees of freedom.

  To get the p-value, plug your test statistic into the relevant cumulative distribution function:

$$2 \cdot \text{c.d.f.}_{t_{N-k}}(-|t|)$$

  In `Stata`, you can compute this quantity manually using `display 2 * ttail(df, t)`

- Hypothesis about a single linear combination of coefficients

  Test statistic:

$$t = \frac{\mathbf{r}'\hat{\boldsymbol{\beta}} - q}{\sqrt{\mathbf{r}'(s^2(\mathbf{X}'\mathbf{X})^{-1})\mathbf{r}}}$$

  where $\mathbf{r}$ is a $k \times 1$ restriction vector and q is a scalar.

  P-value: Again, we have

$$2 \cdot \text{c.d.f.}_{t_{N-k}}(-|t|)$$

  In `Stata`, use post-estimation command `lincom`

- Hypothesis about multiple linear restrictions

  Test statistic:

$$F = \frac{(\mathbf{r}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{r}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}']^{-1}(\mathbf{r}\hat{\boldsymbol{\beta}} - \mathbf{q})}{ms^2}$$

  where $\mathbf{r}$ is now a $m \times k$ restriction matrix and $\mathbf{q}$ is a $m \times 1$ vector.

  P-value:

$$\text{c.d.f.}_{F_{m,N-k}}(F)$$

  In `Stata`, use post-estimation command `test`

  For the particular case of nested models, we can compute the test statistic as

$$\begin{aligned} F &= \frac{(SSR_r - SSR_u)/m}{SSR_u/(N-k)} \\ &= \frac{(\mathbf{e}_r'\mathbf{e}_r - \mathbf{e}_u'\mathbf{e}_u)/m}{\mathbf{e}_u'\mathbf{e}_u/(N-k)} \end{aligned}$$

  where SSR_r is the error sum of squares in the restricted model and SSR_u is the error sum of squares in the unrestricted model.

  Since $\frac{SSR}{SST} = 1 - R^2$, we can also write

$$F = \frac{(R_u^2 - R_r^2)/m}{(1 - R_u^2)/(N-k)}$$

To test the model as a whole, compute

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)}$$

where the restricted model is $y_i = \beta_0 + \varepsilon_i$, so the number of restrictions is $k-1$.

Make sure the restricted model is nested! But can estimate for example the unrestricted model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \boldsymbol{\varepsilon}$$

and test the hypothesis that $\beta_1 = \beta_2$ and $\beta_3 = 1$ by estimating the restricted model

$$\mathbf{y} - \mathbf{x}_3 = \beta_0 + \beta_{12}(\mathbf{x}_1 + \mathbf{x}_2) + \boldsymbol{\varepsilon}.$$

Substantive versus statistical significance

Effect size is only one factor in determining the magnitude of the test statistic! A sufficiently large sample can produce statistically significant findings even for very small effects.

In the case of OLS, the coefficient is informative about a regressor's substantive impact. But that's not generally true for complicated models.

Goodness of fit

Recall from lecture 10 that

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

By construction, R^2 usually increases and never decreases when we add additional regressors to the model.

This makes it difficult to compare R^2 as a measure of the goodness of fit across models. An alternative is the adjusted R^2 statistic:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{SSR/(N-k)}{SST/(N-1)} \\ &= 1 - \frac{N-1}{N-k}(1-R^2) \end{aligned}$$

(Remember that we use k to refer to the size of data matrix X including a constant.)

Partialling out

Frisch-Waugh-Lovell Theorem (FWL): In the linear least squares regression of $\mathbf{y}$ on two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$, the subvector $\hat{\boldsymbol{\beta}}_2$ is the set of coefficients obtained when the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone are regressed on the residuals obtained when each column of $\mathbf{X}_2$ is regressed on $\mathbf{X}_1$.

If $\mathbf{X}_2$ contains just one variable:

- Regress $\mathbf{y}$ on $\mathbf{X}_1$ and get residuals $\mathbf{e}^*$.
- Regress $\mathbf{X}_2$ on $\mathbf{X}_1$ and get residuals $\mathbf{e}^{**}$.
- Regress $\mathbf{e}^*$ on $\mathbf{e}^{**}$ to find $\hat{\beta}_2$.

This is useful for displaying bivariate relationships in a multivariate context: Plot $\mathbf{M}_{\mathbf{X}_1}\mathbf{y}$ against $\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$, not $\mathbf{y}$ against $\mathbf{X}_2$.

Indicator variables

Suppose we want to know whether campaign contributions vary with an individual's region of residence. Let's say we divide the country into four regions (Northeast, South, Midwest, West) and estimate

$$y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \varepsilon_i$$

where d_{ji} is an indicator (or dummy) variable which takes the value of 1 if i is from region j and 0 otherwise.

Why are we only including three instead of four indicator variables?

What is the interpretation of the intercept? What is the interpretation of the coefficients on the indicator variables and their t-statistics?

Polynomial regression

Suppose we are interested in testing for a curvilinear relationship between age and productivity, and we estimate

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \varepsilon_i$$

where x_{1i} measures age and x_{2i} is some other regressor.

What is the marginal effect of age? What is the correct standard error?

Cubic terms are also sometimes used, but rarely do we see higher-order polynomials. The flexibility of the curve you're fitting depends on the order of the polynomial: A polynomial

of order p can have $p-1$ changes in curvature. Polynomials impose substantial structure on the data; alternatives to capture nonlinear relationships include local regression estimators and smoothing splines.

In order to minimize correlation between a regressor and its higher-order terms, it can be helpful to standardize the regressor before creating the polynomial terms.

You'll almost always want to include all lower-order terms. The reason is that the model fit will not necessarily be invariant to the scale of a regressor for which lower-order terms are omitted.

Interaction effects

Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \varepsilon_i$$

where the marginal effect of X_1 is given by

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2.$$

What is the correct standard error of the interaction effect?

It's not s.e.($\hat{\beta}_1$) or s.e.($\hat{\beta}_3$), which are reported by `Stata` as part of its default output.

We need

$$\begin{aligned} s_{\frac{\partial y}{\partial X_1}} &= \sqrt{\text{Var}(\hat{\beta}_1 + \hat{\beta}_3 X_2)} \\ &= \sqrt{\text{Var}(\hat{\beta}_1) + X_2^2 \text{Var}(\hat{\beta}_3) - 2X_2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_3)} \end{aligned}$$

because $\text{Cov}(X, aY) = a\text{Cov}(X, Y)$ and so $\text{Var}(X+aY) = \text{Var}(X)+a^2\text{Var}(Y)-2a\text{Cov}(X, Y)$.

Just like the marginal effect, the variance is not constant across values for X_2.

In `Stata`, use the `margins` command.

Typically, we ask at what values for X_2 the variable X_1 has a significant effect on Y. That is, we perform hypotheses tests across a range of values of X_2, where

$$\frac{\hat{\beta}_1 + \hat{\beta}_3 X_2}{\sqrt{\text{Var}(\hat{\beta}_1 + \hat{\beta}_3 X_2)}} \sim t_{N-k}$$

Finally, make sure to include all relevant constitutive terms. Why?

Specification error: Omitted variable bias

So far, we've assumed a correct specification of the regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

What if the model is misspecified? Suppose the correct model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are matrices with k_1 and k_2 columns. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, we have

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}
\end{aligned}$$

With $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$, taking expectations yields

$$E(\tilde{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.$$

So, $\tilde{\boldsymbol{\beta}}_1$ will be a biased estimator of $\boldsymbol{\beta}_1$ if both $\mathbf{X}_1'\mathbf{X}_2 \neq 0$ and $\boldsymbol{\beta}_2 \neq 0$.

The term $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is $k_1 \times k_2$, where each column contains the coefficients from the OLS regression of the corresponding columns from $\mathbf{X}_2$ on $\mathbf{X}_1$. Note the similarity to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$!

What is the direction of bias?

If $\mathbf{X}_1$ and $\mathbf{X}_2$ contain just one variable each, the sign of bias is the sign of the product of the correlation of $\mathbf{X}_1$ and $\mathbf{X}_2$ and the effect of $\mathbf{X}_2$ on $\mathbf{y}$.

But in practice, $\mathbf{X}_1$ often includes many variables and we worry about a single variable $\mathbf{X}_2$. In this case, the bias on a particular element of $\tilde{\boldsymbol{\beta}}_1$ has the sign of the effect of $\mathbf{X}_2$ on $\mathbf{y}$ times the effect of the corresponding column of $\mathbf{X}_1$ on $\mathbf{X}_2$ controlling for the other columns of $\mathbf{X}_1$.

Example (from Greene):

What is the effect of gas prices on consumption? Market data suggests that the bivariate relationship is positive!

The confounder here is income per capita. What is the direction of bias for the estimated effect of price on consumption? Positive, because income is positively correlated with both consumption and price.

What if the model also included car prices? The direction of bias then has the sign of the effect of income on consumption (which is positive) times the coefficient from the regression of income on gas price controlling for car prices.

Quant I: Lecture 15

Bernd Beber
New York University

Fall 2012

Inclusion of irrelevant variables

Suppose the correct model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta_1} + \boldsymbol{\varepsilon}$$

but we estimate the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta_1} + \mathbf{X}_2\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}.$$

Do we estimate coefficients with bias, as in the case of the omission of relevant regressors?

No, because the estimated model is not incorrect! It just fails to impose the restriction $\boldsymbol{\beta_2} = 0$. Knowing that $\boldsymbol{\beta_2} = 0$, the two expressions are identical, and $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathbf{1}}) = \boldsymbol{\beta_1}$ and $\mathrm{E}(s^2) = \sigma^2$.

So what is the problem? Why not "overfit"?

The reason is that the inclusion of additional regressors (almost always) leads to a loss of precision. (This does not imply that the sparsest model will necessarily produce the smallest attained significance level on your regressor of interest. Why?)

Suppose we include only $\mathbf{X}_1$ in the regression. Then

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

Let's compare this to the variance-covariance matrix we get from a regression that includes both $\mathbf{X}_1$ and $\mathbf{X}_2$, and let's call the coefficients on $\mathbf{X}_i$ from that regression $\tilde{\boldsymbol{\beta}}_{\boldsymbol{i}}$. In that case, $\mathrm{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}})$ is the upper-left block of $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}}$ is the stacked vector of coefficients $\tilde{\boldsymbol{\beta}}_{\mathbf{1}}$ and $\tilde{\boldsymbol{\beta}}_{\mathbf{2}}$. That is,

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}}) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$$

where

$$\mathbf{M}_2 = \mathbf{I}_N - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'.$$

Hence,

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1)^{-1}.$$

It's more convenient to compare $\text{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}})$ and $\text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})$ by comparing their inverses, which we can do because $\mathbf{A} > \mathbf{B}$ implies $\mathbf{B}^{-1} > \mathbf{A}^{-1}$, just as it does for scalars.

This gives us

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})^{-1} - \text{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}})^{-1} &= \frac{1}{\sigma^2}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 \\
&= \frac{1}{\sigma^2}(\mathbf{X}_2'\mathbf{X}_1)'(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1.
\end{aligned}$$

Since this expression is nonnegative definite, the inverse of $\text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})$ is always at least as large as the inverse of $\text{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}})$, and so $\text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})$ is never larger than $\text{Var}(\tilde{\boldsymbol{\beta}}_{\mathbf{1}})$.

In other words, we estimate the coefficients on $\mathbf{X}_1$ less efficiently if we include $\mathbf{X}_2$ in the model, unless the regressors in $\mathbf{X}_2$ are unrelated to the regressors in $\mathbf{X}_1$. The more related they are, the bigger the potential loss of precision.

As a rule of thumb, we gain efficiency when the variance of regressors goes up and their covariance goes down. We can also see this by noting that

$$\begin{aligned}
\text{Var}(\hat{\beta}_j) &= \frac{\sigma^2}{(1 - R_k^2)\sum_i^N (x_{ij} - \bar{x}_j)^2} \\
&= \frac{\sigma^2}{(1 - R_k^2)\text{Var}(\mathbf{x}_j)(N-1)}
\end{aligned}$$

where R_k^2 is the R-squared from a regression of $\mathbf{x}_j$ on all other variables (Fox, p. 308).

This illustrates four ways in which we can lose precision when estimating a coefficient:

- A smaller sample,
- more variability in the error term,
- less variation in our regressor of interest,
- greater correlation between our regressor of interest and other variables.

So, we have a problem: We stand to lose efficiency when we try to prevent omitted variable bias!

Implications for model building?

- Move from simple to more general models?
  1. But the criteria to decide on the inclusion of a variable will be tainted by bias.
  2. Bias is generally considered a bigger problem than inefficiency. We typically worry more about Type I than Type II errors.
- Move from general to simple models?

  Perhaps preferable, and feasible with modern computers.
  1. But small datasets might not allow estimation of general models.
  2. A stepwise procedure poses again the risk of bias in inclusion decisions, since we only get the full benefits of generality at the initial step.
  3. Large models with a fixed rejection probability will generate statistically significant results by chance. For example, we would expect 1 in 20 tests to be significant at the 5% level. So even though large models can be unnecessarily inefficient, it's often when faced with large models that we worry about misleadingly significant results.

     However, there are a variety of corrections we can apply, with the Bonferroni correction being the most conservative: Use significance level $\alpha_k = \frac{\alpha}{k}$ where k is the number of tests and α is the significance level for a single test.

Some practical suggestions:

- Have a plan before you start working with the data. Even better, share or even register your plan!
- Avoid data mining. If you want to make correlative or even causal claims with respect to specific regressors, focus on a small number of well-motivated explanatory factors. Then build each model around one regressor of interest. Don't offer a substantive interpretation of the coefficients on the other covariates.
- Do present different specifications. Be aware if data availability varies across regressors and so your sample varies across specifications.
- Make sure small changes in the composition of your sample don't change your results, for example by bootstrapping or jackknifing.
- Use time period and location indicators or other relevant group indicators to account for heterogeneity across groups of observations.

- Avoid highly correlated regressors. Standardize regressors before you use them to generate squared or cubic terms (or certain other interactions). Don't include alternative measures of your variable of interest in the same model.
- In general don't include "post-treatment" variables. Figure out the counterfactual that's implied by your model.
- Make sure the implied counterfactual is in fact represented in your data. Note that you can but perhaps shouldn't fit a regression line through the data even if that's not the case. If appropriate, use techniques that ensure common support, such as matching.
- Try not to make claims about values that your regressors can take on but for which you don't have any data.
- If your goal is prediction, assess fit with a measure that penalizes large models, such as adjusted R-squared. Two other alternatives with an even greater penalty for lost degrees of freedom are the Akaike Information Criterion, which is given by

$$\text{AIC}(k) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{N}\right) + \frac{2k}{N},$$

  and the Bayesian Information Criterion, which is given by

$$\text{BIC}(k) = \ln\left(\frac{\mathbf{e}'\mathbf{e}}{N}\right) + \frac{k \ln N}{N}.$$

- Try cross-validation, especially with large data sets: Divide your data, build a model with one part of the data, then apply to the other part.

In the end, there is no clear-cut way to figure out the "correct" model specification. At the same time, we assume that we know exactly what the true data-generating process is whenever we run a regression! A better approach might be to acknowledge that we are uncertain about the model and to try and incorporate this uncertainty into our analysis, something that you may be able to accomplish through Bayesian model averaging, although this remains very computationally intensive. Other ways to automate the process of model selection exist as well, e.g. BART.

Know your data

Even though you should devise a model when or as if you didn't know your data, make sure you understand any peculiarities you might have in your data once you're working with it!

- Outlier: an observation with a large residuals/far from the regression line.

- Leverage point: a point far from $\bar{x}$, the average value for a given regressor, because the regression line tilts around $\bar{x}$ (when a constant is included).

- Influence point: an observation that actually changes the regression line depending on whether it is included.

  In `Stata`, diagnose outliers with `rvfplot` and `avplot`. Use `lvrzplot` to plot leverage points against residuals.

- Missingness:
  - If data is missing completely at random (MCAR), it leads to inefficiency but not bias.
  - If data is missing at random (MAR), then it is a function only of observables, and we can prevent bias through appropriate modeling choices.
  - If missingness is non-ignorable, we need to worry about bias and inefficiency.

  If data is MAR, use multiple imputation:
  - Predict values of missing observations from a model for variables with missingness.
  - Repeat multiple times.
  - Run model of interest on each imputed dataset.
  - Average across estimations.

  In `Stata`, use the extensive `mi` suite of commands.

Quant I: Lecture 16

Bernd Beber
New York University

Fall 2012

Last time we talked about model misspecification as a possible violation of the Gauss-Markov assumptions. Now, let's focus on the accurate specification of the error structure.

We have made two assumptions under Gauss-Markov:

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_N \quad \text{(spherical errors)}$$

and

$$\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = 0 \quad \text{(exogeneity)}.$$

Non-spherical errors

We assumed

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_N = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

but suppose $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\boldsymbol{\Omega}$.

1. We may have heteroskedastic (as opposed to homoskedastic) errors, where diagonal element $\sigma_i^2 \neq \sigma_j^2$ for some i, j. This is a problem especially with cross-sectional data.
2. We may have autocorrelated (as opposed to uncorrelated) errors, so that we have non-zero off-diagonal elements. This is a problem in particular with time series data.

Example: Data with a binary outcome variable.

One problem is heteroskedasticity!

Formally, if $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, then

$$\begin{aligned} \varepsilon_i &= 1 - \mathbf{x}_i'\boldsymbol{\beta} \text{ for } y_i = 1 \text{ and} \\ \varepsilon_i &= -\mathbf{x}_i'\boldsymbol{\beta} \text{ for } y_i = 0. \end{aligned}$$

So,

$$\begin{aligned} \text{Var}(\varepsilon_i) &= \text{E}(\Pr(y_i = 1)(1 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \Pr(y_i = 0)(-\mathbf{x}_i'\boldsymbol{\beta})^2) \quad (\text{given E}(\varepsilon_i) = 0) \\ &= \mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta})^2 + (1 - \mathbf{x}_i'\boldsymbol{\beta})(-\mathbf{x}_i'\boldsymbol{\beta})^2 \\ &= (1 - \mathbf{x}_i'\boldsymbol{\beta})[\mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta}) + (-\mathbf{x}_i'\boldsymbol{\beta})^2] \\ &= (1 - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i'\boldsymbol{\beta} \end{aligned}$$

which is analogous to the variance of a Bernoulli-distributed random variable with parameter π, where $\text{Var}(Y) = \pi(1 - \pi)$.

Hence, $\text{Var}(\varepsilon_i)$ is maximized when $\mathbf{x}_i'\boldsymbol{\beta} = 1 - \mathbf{x}_i'\boldsymbol{\beta}$ or $\mathbf{x}_i'\boldsymbol{\beta} = \frac{1}{2}$, and it varies with $\mathbf{x}_i$ and hence i.

What if $\mathbf{x}_i'\boldsymbol{\beta}$ falls outside of $[0, 1]$? Then the fitted values are not interpretable as probabilities, another limitation of the linear probability model and a violation of the Gauss-Markov specification assumption. That is, we also suffer from misspecification, but for now we'll focus on problems with the error structure.

What is the problem with non-spherical errors? Not bias in coefficient estimates! As before $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ and so $\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ because $\text{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$.

But there are two problems with $\text{Var}(\hat{\boldsymbol{\beta}})$:

1. The estimate may be biased, usually downward,
2. The estimate may no longer be best (i.e. most efficient).

Recall that

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{E}((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})') \\ &= \text{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where $\boldsymbol{\Omega}$ is now causing problems.

How do we detect problems?

Take advantage of the fact that since $\hat{\boldsymbol{\beta}}$ is unbiased, residuals $\mathbf{e}$ are unbiased!

1. Visual inspection of residuals
2. Test statistics:
   - For null of homoskedasticity, Breusch-Pagan test:

     Regress
     $$e_i^2 = \delta_0 + \mathbf{x}_{1i}\delta_1 + \mathbf{x}_{2i}\delta_2 + \ldots + \nu$$
     and test
     $$H_0 : \delta_1 = \delta_2 = \ldots = \delta_{k-1} = 0$$
     using an F-test. In `Stata`, use post-estimation command `estat hettest`.

     An alternative is the White test, which specifies a different functional form (involving interaction terms, fewer degrees of freedom).
   - For null of no correlation, Durbin-Watson test for first-order correlation, Breusch-Godfrey test for higher-order serial correlation. In `Stata`, type `estat dwatson` and `estat bgodfrey` after time-series estimation.

What are the solutions?

Two approaches:

1. Compute robust OLS standard errors, i.e. correct for the bias in $\text{Var}(\hat{\boldsymbol{\beta}}_{OLS})$. This usually increases standard errors.

Suppose $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$.

We sometimes write $\sigma_i^2 = \sigma^2\omega_i$, which is an arbitrary scaling but permits $\sum_{i=1}^N \omega_i = N$.

Recall that we require

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

which in the absence of autocorrelation simplifies to

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^N \sigma_i^2 \underbrace{\mathbf{x}_i}_{k\times 1}\underbrace{\mathbf{x}'_i}_{1\times k}\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

But $\sum_{i=1}^N \varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'$ converges in probability to $\sum_{i=1}^N \sigma_i^2\mathbf{x}_i\mathbf{x}_i'$, and we do have a consistent estimator for ε—the residual e, which is consistent because $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent. (Note that we're making an asymptotic claim here.)

Hence

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})}_{asym.} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^N e_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1},$$

known as White standard errors (or alternatively as White-Huber standard errors, robust standard errors, heteroskedasticity-consistent standard errors, etc.). In `Stata`, use the `robust` option.

2. Compute a generalized least squares (GLS) model, which models $\boldsymbol{\Omega}$ to recover efficiency. This will usually decrease your standard errors. In particular, we'll compute a weighted least squares (WLS) estimator used for heteroskedastic data, which is one type of GLS estimator.

   Suppose we know $\boldsymbol{\Omega}$. Then we could simply divide each observation's error by its unit-specific ω_i to obtain:

   $$\mathrm{E}(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}) = \sigma^2\mathbf{I}_N.$$

   In a linear model, dividing ε_i means we have to divide everything else too, so that

   $$\frac{y_i}{\omega_i} = \frac{\beta_0}{\omega_i} + \frac{\beta_1 x_{i1}}{\omega_i} + \cdots + \frac{\varepsilon_i}{\omega_i}$$

with new estimator

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

which is BLUE (since Gauss-Markov assumptions are met for the transformed model).

For example, correct for known heteroskedasticity in a particular covariate $\mathbf{x}$ by dividing by it (if $\sigma_i^2 = \sigma^2\mathbf{x}_i^2$) or its square root (if $\sigma_i^2 = \sigma^2\mathbf{x}_i$).

What if we don't know $\boldsymbol{\Omega}$? We can estimate a feasible GLS model:

- Estimate consistent coefficients, e.g. $\hat{\boldsymbol{\beta}}_{OLS}$
- Compute residuals, $e_i = (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})$
- Estimate $e_i^2 = \mathbf{z}_i'\boldsymbol{\delta} + \nu_i$, and obtain $\boldsymbol{\Omega}(\hat{\boldsymbol{\delta}})$
- Compute $\hat{\boldsymbol{\beta}}_{GLS}$ with $\boldsymbol{\Omega}(\hat{\boldsymbol{\delta}})$

This is the 2-step GLS procedure; you can also implement FGLS by way of maximum likelihood. In `Stata`, use `xtgls`.

Which of the two solutions is better?

No right or wrong answer. But in practice, be conservative: If you claim to reject the null, compute robust standard errors under heteroskedasticity. If you claim not to reject the null, compute more efficient GLS estimates.

Huber-White standard errors are much more popular than GLS, also because we usually do not know the correct expression for $\boldsymbol{\Omega}$. (Note that the linear probability model is is an exception in this respect, because we know $\text{Var}(\varepsilon_i) = (1 - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i'\boldsymbol{\beta}$.)

Quant I: Lecture 17

Bernd Beber
New York University

Fall 2012

Latent choice models

We could model a binary outcome by way of an LPM and ignore possible Gauss-Markov violations, or adjust for heteroskedasticity and run GLS. But more commonly, we'll want to account for the possibility of both non-spherical errors and non-linearity, and we'll usually do so with a latent choice model:

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

where we observe

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

So,

$$\begin{aligned} \Pr(y_i = 1|\mathbf{x}_i) &= \Pr(y_i^* \geq 0|\mathbf{x}_i) \\ &= \Pr(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \geq 0) \\ &= \Pr(\varepsilon_i \geq -\mathbf{x}_i'\boldsymbol{\beta}) \end{aligned}$$

Let $F(\cdot)$ be the cdf of ε_i, such that $F(\varepsilon) = \Pr(\varepsilon_i \leq \varepsilon)$.

Then $\Pr(y_i = 1|\mathbf{x}_i) = 1 - F(-\mathbf{x}_i'\boldsymbol{\beta})$.

If $f(\cdot)$ is symmetric, we have

$$\Pr(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta}).$$

Popular link functions F:

- LPM:
$$\Pr(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
- Probit:
$$\Pr(y_i = 1|\mathbf{x}_i) = \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta})$$
- Logit:
$$\Pr(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

We then find $\hat{\boldsymbol{\beta}}$ through maximum likelihood:

Given independent y_i, we have

$$\begin{aligned}\Pr(Y_1 = y_1; Y_2 = y_2 \ldots Y_N = y_N) &= \prod_{y_i=0}(1 - F(\mathbf{x}_i'\boldsymbol{\beta})) \prod_{y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} F(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - F(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}\end{aligned}$$

Taking logs yields

$$\ln L = \sum_{i=1}^{N}[y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i)\ln(1 - F(\mathbf{x}_i'\boldsymbol{\beta}))]$$

We can minimize this expression numerically:

- Pick a starting value for $\boldsymbol{\beta}$, e.g. $\hat{\boldsymbol{\beta}}_{OLS}$,
- Compute the log likelihood,
- Move to a different $\hat{\boldsymbol{\beta}}$, recompute $\ln L$, and compare.

Estimated coefficients are not usually directly interpretable!

Non-exogeneity

This is the second type of non-Gauss-Markov error structure we frequently encounter, where $E(\boldsymbol{\varepsilon}|\mathbf{X}) \neq 0$.

This could be due to:

1. Measurement error:

   Suppose $\mathbf{y} = \beta_0 + \beta_1\mathbf{x} + \boldsymbol{\varepsilon}$ but $\mathbf{x}$ is measured as $\mathbf{z} = \mathbf{x} + \boldsymbol{\nu}$, with $\mathrm{E}(\boldsymbol{\nu}) = 0$ and $\mathrm{E}(\boldsymbol{\nu}\boldsymbol{\nu}') = \sigma^2_{\boldsymbol{\nu}}\mathbf{I}_N > 0$.

   This implies $\mathbf{x} = \mathbf{z} - \boldsymbol{\nu}$, so,

   $$\begin{aligned} \mathbf{y} &= \beta_0 + \beta_1(\mathbf{z} - \boldsymbol{\nu}) + \boldsymbol{\varepsilon} \\ &= \beta_0 + \beta_1\mathbf{z} + (\boldsymbol{\varepsilon} - \beta_1\boldsymbol{\nu}) \end{aligned}$$

   Now, $\boldsymbol{\nu}$ is in both $\mathbf{z}$ and the error term, so the error term is not zero in expectation conditional on the data.

   What's the direction of the bias? The coefficient on the endogenous regressor will be attenuated towards zero ($\mathbf{y}$ gets smaller but $\mathbf{z}$ larger as $\boldsymbol{\nu}$ increases), so measurement error is often a limited concern (if we claim to reject the null).

   Note this is true only if the measurement error isn't systematic, i.e. it's not the case that $\mathrm{E}(\boldsymbol{\nu}) \neq 0$.

2. Simultaneous equations:

   Suppose

   $$\begin{aligned} \mathbf{y} &= \beta_0 + \beta_1\mathbf{x} + \boldsymbol{\varepsilon}, \text{ and} \\ \mathbf{x} &= \alpha_0 + \alpha_1\mathbf{y} + \boldsymbol{\nu}. \end{aligned}$$

   Large values ε_i then lead to large y_i which lead to large x_i. But then $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are correlated, and $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{x}) \neq 0$, a violation of exogeneity.

3. Omitted variables:

   Suppose

   $$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$$

   and that $\mathbf{x}_2$ is not observed.

   Then the estimated model is

   $$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \boldsymbol{\nu}$$

   where $\boldsymbol{\nu} = \beta_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$ is the error term, which is correlated with $\mathbf{x}_1$ if $\mathbf{x}_2$ and $\mathbf{x}_1$ are correlated and $\beta_2 \neq 0$.

Solutions?

- Randomized control trials
- Regression discontinuity
- Instrumental variable estimation

Instrumental variable estimation:

Suppose we estimate

$$Y = \beta_0 + X\beta_1 + \varepsilon,$$

but suspect $\text{Cov}(X, \varepsilon) \neq 0$. For illustrative purposes, let's say we've mean-standardized all of our variables. (This means we can drop β_0. Why?)

Suppose further that we can identify some Z that predicts X but can be excluded from our model for Y.

We can then write

$$\text{Cov}(Y, Z) = \text{Cov}(X, Z)\beta_1 + \text{Cov}(\varepsilon, Z)$$

If $\text{Cov}(\varepsilon, Z) = 0$ and $\text{Cov}(X, Z) \neq 0$, we have

$$\hat{\beta}_{1,\text{IV}} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}$$

We require:

1. Instrument strength (Z predicts X)
2. Valid exclusion restriction (Z must not be included in the model for Y)