Advanced Econometrics I

Jürgen Meinecke

Lecture 4 of 12

Research School of Economics, Australian National University

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

Why do we care about the distribution of $\hat{\beta}^{\text{OLS}}$?

Knowing the distribution helps us understand *precision* of the estimate

In addition, people use the distribution to construct statistical tests

I prefer to focus on precision and ignore statistical testing

For the sake of illustration, let's tentatively assume that $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \sim \text{N}(0, \Omega)$

The point here is that we assume that the normal distribution is *exact*, not just an asymptotic approximation

Proposition

Let r be some K dimensional nonstochastic vector. Then

$$\sqrt{N}(r'\hat{\beta}^{OLS} - r'\beta^*) \sim N(0, r'\Omega r)$$

Corollary

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\Omega r/N}} \sim N(0,1)$$

You can pick r to consider any linear combination of the elements of β^* that you are interested in

Most times people use $r = e_k$ where e_k is the k-th unit vector taking the value 1 in position k and the value zero elsewhere

That way you are grabbing the kth element of a vector, or the (k,k) element of a matrix

- $\beta_k^* = e_k' \beta^* = \beta^{*\prime} e_k$
- $\omega_{kk} = e_k' \Omega e_k$

Therefore

$$\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\omega_{kk}/N}} = \frac{e_k' \hat{\beta}^{\text{OLS}} - e_k' \beta^*}{\sqrt{e_k' \Omega e_k / N}} \sim \text{N}(0,1)$$

The OLS estimator is a point estimator

It is unlikely that $\hat{\beta}^{\text{OLS}} = \beta^*$
(in fact, that event has probability zero)

Instead of a point estimator, should we consider an interval estimator?

Considerations:

- the smallest interval we would consider is $\hat{\beta}^{\text{OLS}}$ itself
- by having a proper interval, we can make sure that β^* is covered with a probability larger than zero (unlike for point estimates)
- the largest interval covers the whole real line and guarantees a 100% coverage probability (not very informative though)
- there's a tension between two goals:
  high coverage probability vs narrow (informative) interval

Idea: accept a coverage probability that is a little less than 100%, say 95%, and hope to obtain an informative interval

Because $\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\omega_{kk}/N}} \sim \text{N}(0,1)$,
the obvious interval that comes to mind is $\left[\hat{\beta}_k^{\text{OLS}} \pm c \cdot \sqrt{\omega_{kk}/N}\right]$

This is symmetric around the point estimate because of the symmetry of the normal distribution

A clever choice of c will ensure a 95% coverage probability:

$$P\left(\beta_k^* \in \left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]\right) = 0.95$$

Careful! Don't read this literally as
"the probability that β_k^* is in interval"

That's incorrect! It makes it sound as if β_k^* is a random variable

The random object is the interval $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]$

So the way to read the above statement is
"the probability that the interval covers β_k^*"

Many people do not understand what a confidence interval can tell them and what it cannot tell them

It means:
Prior to repeatedly estimating $\hat{\beta}_k^{\text{OLS}}$ in separate random experiments, the probability is 95% that the random interval $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]$ contains β_k^*

A frequentist's thought experiment: if I were given 100 random samples of size N then about 95 of them will yield confidence intervals that contain β_k^* (but I don't know which ones)

Common misconceptions regarding confidence intervals

The following statements are all false

- The specific 95% confidence interval presented by a study has a 95% chance of containing the coefficient
- The true coefficient β_k^* has a 95% probability of falling inside the confidence interval
- A coefficient outside the 95% confidence interval is refuted by the data

The first two in particular are believed by many people

Google: *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*, by Greenland at al, a worthwhile read

A few slides ago we tentatively assumed $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \sim \text{N}(0, \Omega)$

Now let's generalize by going back to $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \xrightarrow{\text{d}} \text{N}(0, \Omega)$

It's easy to adjust earlier results accordingly, basically by replacing '$\sim$' with '$\xrightarrow{\text{d}}$'

Proposition

Let r be some K dimensional nonstochastic vector. Then

$$\sqrt{N}(r'\hat{\beta}^{OLS} - r'\beta^*) \xrightarrow{d} N(0, r'\Omega r)$$

Corollary

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\Omega r/N}} \xrightarrow{d} N(0,1)$$

You may replace Ω by $\hat{\Omega} = \Omega + o_p(1)$:

Proposition

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\hat{\Omega}r/N}} \xrightarrow{d} N(0,1)$$

Grabbing one element from that vector: $\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\hat{\omega}_{kk}/N}} \xrightarrow{\text{d}} \text{N}(0,1)$

where $\hat{\omega}_{kk} := e_k'\hat{\Omega}e_k$
(this is a number that we can compute from the sample data)

The confidence interval for β_k^* therefore is $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\hat{\omega}_{kk}/N}\right]$

Terminology:
the term $\sqrt{\hat{\omega}_{kk}/N}$ is also called the **asymptotic standard error** of $\hat{\beta}_k^{\text{OLS}}$

Aside: by convention, an estimator of the standard deviation of an estimator is called a *standard error*

Definition (Asymptotic Standard Error of $\hat{\beta}^{\text{OLS}}$)

Let Ω/N be the asymptotic variance of $\hat{\beta}^{\text{OLS}}$. The **asymptotic standard errors** of the OLS estimator $\hat{\beta}^{\text{OLS}}$ and $\hat{\beta}_k^{\text{OLS}}$ are

$$\text{se}(\hat{\beta}^{\text{OLS}}) = \sqrt{\text{diag}\, \hat{\Omega}/N}$$

$$\text{se}(\hat{\beta}_k^{\text{OLS}}) = e_k' \cdot \text{se}(\hat{\beta}^{\text{OLS}}),$$

where $\hat{\Omega}/N$ is the estimator of the asymptotic variance of $\hat{\beta}^{\text{OLS}}$

We obtain this result regarding the asymptotic coverage probability:

Proposition

$$\lim_{N\to\infty} P\left(\beta_k^* \in [\hat{\beta}_k^{OLS} \pm 1.96 \cdot se(\hat{\beta}_k^{OLS})]\right) = 0.95$$

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

You study $Y_i = X_i\beta^* + u_i$ where $\mathrm{E}\left(u_i X_i\right) = 0$

For simplicity X_i is a scalar

For some reason you are interested in the value of β^*

In particular, you want to know $\beta^* \stackrel{?}{=} 0$

You remember that OLS delivers a consistent estimator

You obtain $\hat{\beta}^{\mathrm{OLS}} = 0.18$

What do you do?

You consider two states of nature:

- $\beta^* = 0$
- $\beta^* \neq 0$

These are mutually exclusive and exhaustive

You can look at them as *hypotheses*

Definition (Statistical Hypothesis)

A **statistical hypothesis** is a statement about a population parameter.

One is the null hypothesis, and one the alternative hypothesis (of course denoted by H_0 and H_1)

You would like to know which one is *true* (if there is such a thing)

To determine which hypothesis is true, you propose:

$$\textit{if } \hat{\beta}^{OLS} = 0 \textit{ then } \beta^* = 0, \textit{ else } \beta^* \neq 0$$

According to this decision rule, you decide that $\beta^* \neq 0$ (because $0.18 \neq 0$)

You have just conducted a *hypothesis test*

Definition

A **statistical hypothesis test** is a decision rule that specifies

(i) for which sample values H_0 is considered true;

(ii) for which sample values H_1 is considered true.

The hypothesis test

$$\textit{if } \hat{\beta}^{OLS} = 0 \textit{ then } \beta^* = 0, \textit{ else } \beta^* \neq 0$$

is not good because you will almost certainly conclude that $\beta^* \neq 0$

This test is extremely conservative

You understand that $\hat{\beta}^{\text{OLS}}$ could be nonzero even if $\beta^* = 0$

The estimator $\hat{\beta}^{\text{OLS}}$ is subject to sampling error

As sensible test should reflect this possibility of sampling error, and therefore the variance of $\hat{\beta}^{\text{OLS}}$ should play a role too

If we are unable to quantify the exact variance of $\hat{\beta}^{\text{OLS}}$, the asymptotic variance will be good enough

The most common statistic to combine information of the point estimate and its variance is the t-statistic

Definition (t-Statistic)

Let $\hat{\theta}$ be an estimator and $\text{se}(\hat{\theta})$ be its asymptotic standard error. Then

$$t_{\hat{\theta}}(\theta) := \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})}$$

is the **t-statistic** or **t-ratio** for θ.

It has the shape of the *standardized* estimator $\hat{\theta}$

Let's say we have two competing estimators, labelled $\hat{\beta}^{\text{OLS}}$ and $\hat{\beta}^{\text{IV}}$ and we want to test if $\beta^* = 24$

Then we would look at $t_{\hat{\beta}^{\text{OLS}}}(24)$ and $t_{\hat{\beta}^{\text{IV}}}(24)$

It should be clear that because $\hat{\beta}^{\text{OLS}} = \beta^* + \text{o}_p(1)$

$$t_{\hat{\beta}^{\text{OLS}}}(\beta^*) \xrightarrow{\text{d}} \text{N}(0,1)$$

Software packages such as Stata have the terrible habit of reporting $t_{\hat{\beta}^{\text{OLS}}}(0)$ as part of a standard regression output

$t_{\hat{\beta}^{\text{OLS}}}(0)$ facilitates a hypothesis test of the null $\beta^* = 0$ against the alternative $\beta^* \neq 0$, the critical value is simply ± 1.96

It is not clear that the null $\beta^* = 0$ is interesting at all

There is an awful practice in applied econometrics to focus on the value of t-statistics, or, equivalently, on *significance stars*

The vast majority of researchers present their estimation tables with *STATA significance stars*

- $|t| > 1.64$ receives one *star*
- $|t| > 1.96$ receives two *stars*
- $|t| > 2.58$ receives three *stars*

It's like the Michelin restaurant guide: the more *stars*, the better!

For example, if the return to schooling is estimated to equal 0.14 and it is statistically significant at the 95% level, then the table will say 0.14^{**}

Many applied papers limit the discussion of their results only to those coefficient estimates with *stars* attached, that is, only to those who are *statistically significant*

Results that don't have any *stars* are often ignored

Our average Monday seminar follows this pattern

Sadly, PhD students copy this terrible practice

I have had countless conversations with PhD students whose goal it is to obtain *stars* in their tables

Because the opinion is: NO STARS, NO PAPER!

The research objective becomes: obtain *stars*

But often times stars are out of reach

Try do your estimations without *stars* or t-statistics

They are simplistic or reductionist

They seem to apply a binary world:
results are either statistically significant or irrelevant

(Also, they encourage *star*-hacking:
the strong incentive to obtain stars)

So what should you be doing?

What ought to be best practice?
(But admittedly and unfortunately isn't)

Report standard errors and confidence intervals

They offer a notion of *precision of estimates*

Also, never ever say this:
"*The estimate is highly significance*"
(or variations thereof)
It demonstrates that you don't understand what you are doing

(Also: don't use STATA)

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

We have extensively studied projections of $Y \in L_2$ on the space spanned by $X_1, \dots, X_K \in L_2$

This linear projection has the following minimization problem representation:

$$\begin{aligned}\widehat{Y} &:= \underset{Z \in \mathrm{sp}(X_1,\dots,X_K)}{\operatorname{argmin}} \|Y - Z\| \\ &= \underset{b_1,\dots,b_K}{\operatorname{argmin}} \|Y - (b_1 X_1 + \cdots + b_K X_K)\|\end{aligned}$$

But why limit ourselves to the subspace $\mathrm{sp}(X_1, \dots, X_K)$?

How about this more flexible problem:

$$\underset{g \in G}{\operatorname{argmin}} \|Y - g(X_1, \dots, X_K)\|,$$

where G is the space of functions from $\mathbb{R}^K \to \mathbb{R}$ with $g(X_1, \dots, X_K) \in L_2$

Clearly, the latter minimization problem contains the former

It plays an important role and gets a familiar label:

Definition (Conditional Expectation Function)

Let $Y, X_1, \dots, X_K \in L_2$. Let G be the space of functions from $\mathbb{R}^K \to \mathbb{R}$ with $g(X_1, \dots, X_K) \in L_2$.

Then the **conditional expectation function** is defined by

$$\mathrm{E}(Y|X_1, \dots, X_K) := \underset{g \in G}{\operatorname{argmin}} \|Y - g(X_1, \dots, X_K)\|.$$

Accordingly, the conditional expectation function is defined as a *projection* of Y on the space of functions G

From the projection theorem we realize that $\mathrm{E}(Y|X_1, \dots, X_K)$

- exists
- is unique

We typically have available random variables Y_i and random vectors X_i with $\dim(X_i) = K \times 1$

Define $\mu(X_i) := \mathsf{E}(Y_i|X_i)$

Notice that $\mu(X_i) \in L_2$

We could define $e_i := Y_i - \mu(X_i)$

This is referred to as the CEF error (or simply error term)

This implies the following representation: $Y_i = \mu(X_i) + e_i$

By definition

$$\begin{aligned}\mathsf{E}(e_i|X_i) &= \mathsf{E}(Y_i - \mu(X_i)|X_i) \\ &= \mathsf{E}(Y_i|X_i) - \mathsf{E}(\mu(X_i)|X_i) \\ &= \mu(X_i) - \mu(X_i) \\ &= 0\end{aligned}$$

By the law of iterated expectations: $\mathrm{E}(e_i) = \mathrm{E}(\mathrm{E}(e_i|X_i)) = 0$

That is, the conditional mean equals the unconditional mean

This is called **conditional mean independence**

Similar to the case of the linear projection model, the statement

$$Y_i = g(X_i) + e_i, \qquad \mathrm{E}(e_i|X_i) = 0$$

is not restrictive at all

It tells you that the function $g(X_i)$ must be the CEF $\mu(X_i)$

The CEF has a very important property

Pick arbitrary $h \in G$, then with $Y_i = \mu(X_i) + e_i$,

$$
\begin{aligned}
&\mathsf{E}\left((Y_i - h(X_i))^2\right) \\
&= \mathsf{E}\left(e_i + (\mu(X_i) - h(X_i))\right)^2 \\
&= \mathsf{E}(e_i^2) + 2 \cdot \mathsf{E}\left(e_i(\mu(X_i) - h(X_i))\right) + \mathsf{E}\left((\mu(X_i) - h(X_i))^2\right) \\
&= \mathsf{E}(e_i^2) + \mathsf{E}\left((\mu(X_i) - h(X_i))^2\right) \\
&= \mathsf{E}\left((Y_i - \mu(X_i))^2\right) + \mathsf{E}\left((\mu(X_i) - h(X_i))^2\right) \\
&\geq \mathsf{E}\left((Y_i - \mu(X_i))^2\right)
\end{aligned}
$$

notice, the third equality is an application of LIE:
$\mathsf{E}(e_i(\mu(X_i) - h(X_i))) = \mathsf{E}((\mu(X_i) - h(X_i)) \cdot \mathsf{E}(e_i|X_i)) = 0$

The CEF $\mu(X_i)$ leads to minimal mean square error, so it's the best predictor using MSE as criterion

When trying to explain Y_i using X_i

- $\mu(X_i)$ is the best predictor of Y_i
  in contrast:
- $X_i'\beta^*$ is the best linear predictor of Y_i

Let's turn to the practical implication next

If an oracle offered us either $X_i'\beta^*$ or $\mu(X_i)$ which one would you prefer to have?

Similarly, if an oracle offered us either a good estimator of $X_i'\beta^*$ or a good estimator of $\mu(X_i)$ which one would you prefer to have?

This raises the question:
What is our overall objective anyway?

Why are we running regressions?

I'm not sure that any group of econometricians (or economists) could agree on a common objective

I'll dip my toe into the waters...

...many econometricians are interested in **causal effects**!

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

What is a causal effect?

As good econometricians, we must have some kind of economic thinking that informs us about the relationship between variables

In other words, we have in mind some kind of *structural economic model*

Definition (Structural Model)

Let $X_1, \dots, X_M$ be an exhaustive list of variables that explain Y.
A **structural model** is given by

$$Y = h(X_1, \dots, X_M),$$

where $h(\cdot)$ is a well behaved function.

Given a structural model, we are now in a position to define precisely what we mean by *causal effect*

Definition (Causal Effect)

The **causal effect** of X_i on the outcome variable Y in the structural model $Y = h(X_1, \dots, X_M)$ is

$$C_i(X_1, \dots, X_M) := \frac{\partial h(X_1, \dots, X_M)}{\partial X_i} \qquad i \in \{1, \dots, M\}$$

Alternatively, we could call this a *structural effect*

Notice that the definition links the term 'causal effect' to the idea of the 'structural model'

Causality therefore is within a particular model

This definition applies for continuous X_i

It's clear how one would tweak this to allow discrete X_i

Can you know C_i?

Three problems:

- you don't know h
- you don't know all the variables $X_1, \ldots, X_M$ that should enter the rhs,
  (the known unknowns, and the unknown unknowns)
  and even if you did
- you only observe iid copies of $(X_1, \ldots, X_K)'$ for $K < M$

How can we help ourselves?

You observe iid copies of Y and $X := (X_1, \dots, X_K)'$

You don't observe $e := (X_{K+1}, \dots, X_M)'$

From your point of view, the structural model becomes $Y = h(X, e)$

(Aside: in the current subsection, the symbol X defines a $K \times 1$-vector, it will revert back to an $N \times K$-matrix subsequently)

Knowing about the best predictor property of μ, you consider

$$\begin{aligned}\mu(X) &= \mathrm{E}\left(h(X_1,\ldots,X_M)|X_1,\ldots,X_K\right)\\ &= \mathrm{E}\left(h(X,e)|X\right)\\ &= \int h(X,e)\cdot f(e|X)de\end{aligned}$$

and its partial derivative, for $i = 1,\ldots,K$,

$$\begin{aligned}\partial\mu(X)/\partial X_i &= \frac{\partial\int h(X,e)\cdot f(e|X)de}{\partial X_i}\\ &= \int\left(\frac{\partial h(X,e)}{\partial X_i}f(e|X) + h(X,e)\frac{\partial f(e|X)}{\partial X_i}\right)de\\ &= \int\left(C_i(X,e)\cdot f(e|X) + h(X,e)\frac{\partial f(e|X)}{\partial X_i}\right)de\end{aligned}$$

The first term gets its own name...

Definition (Average Causal Effect)

The **average causal effect** of X_i on the outcome variable Y in the structural model $Y = h(X_1, \dots, X_M)$ is

$$ACE_i(X) := \int C_i(X, e) \cdot f(e|X) de.$$

It follows that

$$\partial \mu(X) / \partial X_i = ACE_i(X) + \int \left(h(X, e) \frac{\partial f(e|X)}{\partial X_i} \right) de$$

What do we do with the second term on the rhs?

Copy and paste from last slide:

$$\partial\mu(X)/\partial X_i = ACE_i(X) + \int \left(h(X,e) \frac{\partial f(e|X)}{\partial X_i} \right) de,$$

where $i = 1, \ldots, K$

Recall that $f(e|X) = f(e|X_1, \ldots, X_K)$

What would happen if the distribution of e conditional on $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_K)$ was independent of X_i?

Then $f(e|X) = f(e|X_1, \ldots, X_K) = f(e|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_K)$

It follows that under *conditional independence* $\partial f(e|X)/\partial X_i = 0$ and therefore $\partial\mu(X)/\partial X_i = ACE_i(X)$

In words: small deviations of the conditional expectation function identify average causal effects

(This is an important result, but please be aware that we had to pay a high price: we imposed conditional independence)

Exploring some middle ground:
what if we knew that the structural function h was linear?

That is, if $h(X, e) = X'\beta + e$, then

$$\begin{aligned}\mu(X) &= \mathrm{E}\left(h(X, e)|X\right) \\ &= \mathrm{E}\left(X'\beta + e|X\right) \\ &= X'\beta + \mathrm{E}\left(e|X\right)\end{aligned}$$

Studying the partial derivative:

$$\partial\mu(X)/\partial X_i = \partial X'\beta/\partial X_i + \partial\mathrm{E}\left(e|X\right)/\partial X_i = \beta_i + \partial\mathrm{E}\left(e|X\right)/\partial X_i$$

What do we do with the second term on the rhs?

Copy and paste from previous slide:

$$\partial \mu(X)/\partial X_i = \partial X'\beta/\partial X_i + \partial \mathrm{E}\left(e|X\right)/\partial X_i = \beta_i + \partial \mathrm{E}\left(e|X\right)/\partial X_i$$

Conditional ***mean*** *independence*: $\mathrm{E}(e|X) = \mathrm{E}(e) = 0$

This implies $\frac{\partial \mathrm{E}(e|X)}{\partial X_i} = 0$

It follows that $\partial \mu(X)/\partial X_i = \beta_i$

So the average causal effect is very simple

Conditional mean independence is a weaker restriction than conditional independence

But we needed to accept the bargain that the structural function is linear

This leads to the linear regression model...

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

Definition (Linear Regression Model)

$$Y_i = X_i'\beta + e_i, \qquad \mathrm{E}(e_i|X_i) = 0$$

and $\mathrm{E}Y_i^2 < \infty$ and $\mathrm{E}\left\|X_i\right\|^2 < \infty$.

The equality $E(e_i|X_i) = 0$ is called **conditional mean independence (CMI)** or **exogeneity** condition

It implies $\mathrm{E}(e_i X_i) = 0$ because: $\mathrm{E}(e_i X_i) = \mathrm{E}(X_i \cdot \mathrm{E}(e_i|X_i)) = 0$

But not the other way round, eg: $e_i = X_i^2$ with $X_i \sim \mathrm{N}(0,1)$
(then $\mathrm{E}(e_i X_i) = \mathrm{E}(X_i^3) = 0$ but clearly $\mathrm{E}(e_i|X_i) = \mathrm{E}(X_i^2|X_i) = X_i^2 \neq 0$)

In the linear regression model, because $\mathrm{E}(e_i X_i) = 0$, we have
$\beta = \beta^* = \mathrm{E}(X_i X_i')^{-1} \mathrm{E}(X_i Y_i)$
(provided $\mathrm{E}(X_i X_i')$ is invertible)

In a way, the linear regression model merely says that the projection coefficient is also the structural coefficient of interest

Luckily we know a good estimator for β^*: the OLS estimator

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

We've seen that $\mathrm{E}(e_i|X_i) = 0$ is stronger (or more restrictive) than $\mathrm{E}(e_i X_i) = 0$

So it shouldn't be too surprising that the CMI condtion enables us to do new things

We can study the *finite sample properties* of the OLS estimator:

- $\mathrm{E}(\hat{\beta}^{\mathrm{OLS}})$
- Var $(\hat{\beta}^{\mathrm{OLS}})$

Last week we studied *approximate* mean and variance via the asymptotic distribution

Today we are able to derive the *exact* mean and variance (without the need to resort to the CLT)

The CMI condition makes this possible

Expected value

Rewrite

$$\begin{aligned}\hat{\beta}^{\text{OLS}} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + e) \\ &= \beta + (X'X)^{-1}X'e\end{aligned}$$

(implicit definitions of vectors/matrices are the obvious ones)

So that $\hat{\beta}^{\text{OLS}} - \beta = (X'X)^{-1}X'e$

For some reason, people like when $\text{E}\hat{\beta}^{\text{OLS}} = \beta$

If that were true then $\hat{\beta}^{\text{OLS}}$ is **unbiased**

We could make this happen by claiming $\text{E}((X'X)^{-1}X'e) = 0$

But can we claim this?

Yes we can!

This follows readily using the law of iterated expectation:

$$\begin{aligned}\mathsf{E}\left((X'X)^{-1}X'e\right) &= \mathsf{E}\left(\mathsf{E}\left((X'X)^{-1}X'e|X\right)\right)\\ &= \mathsf{E}\left((X'X)^{-1}X'\mathsf{E}(e|X)\right)\\ &= 0\end{aligned}$$

The CMI condition makes the OLS estimator unbiased

Conditional variance

Under CMI, $\mathrm{E}\hat{\beta}^{\mathrm{OLS}} = \beta$, and we get

$$\begin{aligned}
\mathrm{Var}\,(\hat{\beta}^{\mathrm{OLS}}|X) &= \mathrm{E}\left((\hat{\beta}^{\mathrm{OLS}} - \beta)(\hat{\beta}^{\mathrm{OLS}} - \beta)'|X\right) \\
&= \mathrm{E}\left((X'X)^{-1}X'ee'X(X'X)^{-1}|X\right) \\
&= (X'X)^{-1}X' \cdot \mathrm{E}(ee'|X) \cdot X(X'X)^{-1}
\end{aligned}$$

To make this look simpler, assume homoskedasticity:
$\mathrm{E}(ee'|X) = \sigma_e^2 I_N$

It means that the error variances are not functions of X

It follows that $\mathrm{Var}\,(\hat{\beta}^{\mathrm{OLS}}|X) = \sigma_e^2(X'X)^{-1}$

Unconditional variance

Lemma (Variance Decomposition)

For $P, Q \in L_2$, Var P = EVar $(P|Q)$ + Var E$(P|Q)$

Therefore

$$\text{Var } \hat{\beta}^{\text{OLS}} = \text{EVar } (\hat{\beta}^{\text{OLS}}|X) + \text{Var E}(\hat{\beta}^{\text{OLS}}|X)$$

But the second term is zero (why?), thus

$$\begin{aligned}\text{Var } \hat{\beta}^{\text{OLS}} &= \text{E}\left(\sigma_e^2 (X'X)^{-1}\right) \\ &= \sigma_e^2 \cdot \text{E}((X'X)^{-1})\end{aligned}$$

This is the unconditional variance under CMI and homoskedasticity

Roadmap

- Ordinary Least Squares Estimation
  - Standard Errors and Confidence Intervals
  - Hypotheses Tests
  - Conditional Expectation Function
  - Causal Effects
  - Linear Regression Model
  - Finite Sample Properties of the OLS Estimator
  - Gauss Markov Theorem

Gauss Markov Theorem

Theorem (Gauss Markov Theorem)

In the linear regression model with homoskedastic errors, amongst all linear estimators that are conditionally unbiased, $\hat{\beta}^{OLS}$ has the lowest variance.

Some people say that OLS is BLUE:
best linear unbiased estimator

That's not very precise, what does 'best' refer to?

Answer: minimal variance

Sketch of proof

- let $\tilde{\beta} := C'Y$ be any other linear unbiased estimator
  (where C is a $N \times K$-dimensional matrix based on X)
- similar to above, its conditional variance is $\text{Var}\,(\tilde{\beta}|X) = \sigma_e^2 C'C$
- define $D' := C' - (X'X)^{-1}X'$
- unbiasedness implies $C'X = I_K$
  because: $\text{E}(C'Y|X) = \text{E}(C'(X\beta + e)|X) = C'X\beta = \beta$
- therefore $D'X = (C' - (X'X)^{-1}X')X = 0$
- then
  $C'C = (D' + (X'X)^{-1}X')(D' + (X'X)^{-1}X')' = D'D + (X'X)^{-1}$
- it follows

$$\begin{aligned}\text{Var}\,(\tilde{\beta}|X) &= \sigma_e^2 C'C = \sigma_e^2 \left((X'X)^{-1} + D'D\right) \\ &\geq \sigma_e^2 (X'X)^{-1} \\ &= \text{Var}\,(\hat{\beta}^{\text{OLS}}|X),\end{aligned}$$

  inequality because $D'D$ is nonnegative definite (next slide)

Definition (Nonnegative definiteness)

A symmetric matrix P is nonnegative definite if $q'Pq \geq 0$ for all vectors q.

Lemma

$D'D$ is nonnegative definite.

Sketch of proof

D is an $N \times K$ matrix, so $D'D$ is $K \times K$

Take $K \times 1$ vector q, then $q'D'Dq = (q'D')(Dq) = (Dq)'(Dq) \geq 0$

The rhs has the form $r'r$ where r is $N \times 1$

$r'r$ is obviously nonnegative

A constructive proof of Gauss Markov Theorem
(from Amemiya's textbook)

Let $\tilde{\beta} := C'Y$ be any other linear unbiased estimator,
where C is a $N \times K$-dimensional matrix based on X

Under homoskedasticity, Var $(\tilde{\beta}|X) = \sigma_e^2 C'C$

Unbiasedness implies $C'X = I$
because: $\mathrm{E}(C'Y|X) = \mathrm{E}(C'(X\beta + e)|X) = C'X\beta = \beta$

Let's come up with some way of minimizing the variance given the 'constraint' that the estimator is unbiased

We set up a scalar minimization problem

For that, let p be an arbitrary K-vector
(its purpose is to pick any particular linear combination of $\tilde{\beta}$

Instead of considering the vector $\tilde{\beta}$ we look at the scalar $p'\tilde{\beta}$

If $\tilde{\beta} = C'Y$ is an estimator of β,
then $p'C'Y$ is the corresponding estimator of $p'\beta$

Define $c := Cp$, so that $p'C'Y = c'Y$
and notice that unbiasedness implies $X'c = X'Cp = p$

Also, Var $(p'\tilde{\beta}|X) =$ Var $(c'Y|X) = \sigma_e^2 c'c$

Our minimization problem is:

Minimize Var $(c'Y|X)$ subject to the unbiasedness constraint $X'c = p$

The corresponding Lagrangian is $L(c, \lambda) = c'c - 2\lambda'(X'c - p)$

Next: take derivative with respect to c and solve

Remember the following useful matrix derivative rules:

$$\frac{\partial Az}{\partial z} = A' \qquad \frac{\partial (z'Az)}{\partial z} = (A + A')z$$

Therefore,

$$\frac{\partial (c'c - 2\lambda'(X'c - p))}{\partial c} = 2c - 2X\lambda$$

Setting equal to zero results in $c = X\lambda$ or, alternatively, $X'c = X'X\lambda$

Using the constraint $X'c = p$ gives $X'X\lambda = p$ or, alternatively, $\lambda = (X'X)^{-1}p$

And finally, $c = X\lambda = X(X'X)^{-1}p$, so that the minimum variance unbiased estimator of $p'\beta$ turns out to be $p'(X'X)^{-1}X'Y$