Advanced Econometrics I

Jürgen Meinecke

Lecture 2 of 12

Research School of Economics, Australian National University

Roadmap

Projections (rinse and repeat)

- Linear Projections in L_2 — General Case

Ordinary Least Squares Estimation

Given a bunch of random variables $X_1, \ldots, X_K, Y$, we wanted to express Y as a linear combination in $X_1, \ldots, X_K$

A fancy way of saying the same thing:
We want to project Y onto the subspace spanned by $X_1, \ldots, X_K$

That projection is labeled $\mathbb{P}_{\text{sp}(X_1,\ldots,X_K)} Y$ or $\hat{Y}$

Instead of $\mathbb{P}_{\text{sp}(X_1,\ldots,X_K)}$, we may simply write $\mathbb{P}_X$,
where $X := (X_1, \ldots, X_K)'$

(Aside: the X_i can enter non-linearly, for example $X_2 := X_1^2$)

Viewing $X_1, \ldots, X_K, Y$ as elements of a Hilbert space, we learned the generic characterization using the inner product:

Using the orthonormal basis $\tilde{X}_1, \ldots, \tilde{X}_K$
(such that $\text{sp}(\tilde{X}_1, \ldots, \tilde{X}_K) = \text{sp}(X)$)

$$\begin{aligned}
\widehat{Y} = \mathbb{P}_X Y &= \sum_{i=1}^{K} \langle \tilde{X}_i, Y \rangle \tilde{X}_i \\
&= \sum_{i=1}^{K} \mathrm{E}(\tilde{X}_i \cdot Y) \tilde{X}_i \\
&= \sum_{i=1}^{K} \beta_i^* X_i
\end{aligned}$$

For example, when $X_1 = 1$ (constant term) and $K = 2$, we saw

$$\beta_2^* = \frac{\text{Cov}(X_2, Y)}{\text{Var}\ (X_2)}$$

$$\beta_1^* = \mathrm{E}Y - \beta_2^* \mathrm{E}X_2$$

For general K, we use matrices to express $\beta^* := (\beta_1, \ldots, \beta_K)'$

Let $X := (X_1, X_2, \ldots, X_K)'$ be a $K \times 1$ vector

$$\widehat{Y} = \mathbb{P}_X Y = \sum_{i=1}^{K} \beta_i^* X_i = X'\beta^*,$$

where $\beta^* := \left(\mathrm{E}(XX')\right)^{-1} \mathrm{E}(XY)$ is a $K \times 1$ vector

Aside: the last equality above is justified by the following result from linear algebra:

$\sum_{i=1}^{K} x_i y_i = x'y = y'x,$

for generic vectors x and y:

$$x := \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}, \qquad y := \begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix},$$

or written compactly as $x := (x_1, \ldots, x_K)'$ and $y := (y_1, \ldots, y_K)'$

When $X_1 = 1$, β^* can be expressed via covariances

Corollary

When $X = (1, X_2, \dots, X_K)'$, then the projection coefficients are

$$(\beta_2^*, \dots, \beta_K^*)' = \Sigma_{XX}^{-1}\Sigma_{XY}$$

$$\beta_1^* = EY - \beta_2^* EX_2 - \cdots - \beta_K^* EX_K,$$

where

$$\Sigma_{XX} := \begin{bmatrix} \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2K} \\ \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K2} & \sigma_{K3} & \dots & \sigma_K^2 \end{bmatrix}, \Sigma_{XY} := \begin{bmatrix} Cov(X_2, Y) \\ Cov(X_3, Y) \\ \vdots \\ Cov(X_K, Y) \end{bmatrix}$$

Σ_{XX} is matrix that collects variances of X on the diagonal and covariances on the off-diagonal

Σ_{XY} is vector that collects covariances between X and Y

Linear projection representation of Y:

$$\begin{aligned} Y &= \mathbb{P}_X Y + (Y - \mathbb{P}_X Y) \\ &=: \mathbb{P}_X Y + u \\ &= X'\beta^* + u, \end{aligned}$$

The element Y of a Hilbert space can be reached by adding two elements:

- an element $X'\beta^*$ from the subspace $\mathrm{sp}(X)$;
- an element u that is orthogonal to $\mathrm{sp}(X)$

Proof that u is orthogonal to $\mathrm{sp}(X)$, that is, $\mathrm{E}(Xu) = 0$:

$$\begin{aligned} \mathrm{E}(X(Y - \mathbb{P}_X Y)) &= \mathrm{E}(X(Y - X' \cdot \mathrm{E}(XX')^{-1}\mathrm{E}(XY))) \\ &= \mathrm{E}(XY - XX' \cdot \mathrm{E}(XX')^{-1}\mathrm{E}(XY)) \\ &= \mathrm{E}(XY) - \mathrm{E}(XX')\mathrm{E}(XX')^{-1}\mathrm{E}(XY) \\ &= 0 \end{aligned}$$

Using the linear projection representation

$$Y = X'\beta^* + u$$

Once you learn that $E(Xu) = 0$ you know that β^* must be the projection coefficient

You have learned that it exists and is unique

It is important to understand that the definition of the linear projection model is not restrictive

In particular, $\mathrm{E}(uX) = 0$ is not an assumption, it is *definitional*

To drive home this point, suppose I claim

$$Y = X'\theta + w$$

Next I tell you that $\mathrm{E}(wX) = 0$

You therefore conclude that $\theta = \beta^* = \left(\mathrm{E}(XX')\right)^{-1} \mathrm{E}(XY)$

In summary

Definition (Linear Projection Model)

Given

(i) $X_1, \ldots, X_K, Y \in L_2$

(ii) $\mathrm{E}(XX') > 0$ (positive definite)
(aka, *no perfect multicollinearity*)

Then the **linear projection model** is given by

$$Y = X'\beta^* + u,$$

where $\mathrm{E}(uX) = 0$ and $\beta^* = \left(\mathrm{E}(XX')\right)^{-1} \mathrm{E}(XY)$.

We accept and understand now that the unique projection coefficient exists

Let's say we're interested in knowing the value of β^*

We just learned that $(\beta_2^*, \ldots, \beta_K^*)' = \Sigma_{XX}^{-1}\Sigma_{XY}$

Do we know the objects on the rhs?

These are **population** variances and covariances

We don't know these, therefore we don't know β^*

How else could we quantify β^*?

Roadmap

Projections (rinse and repeat)

Ordinary Least Squares Estimation

- The Problem of Estimation
- Definition of the OLS Estimator
- Basic Asymptotic Theory (part 1 of 2)
- Large Sample Properties of the OLS Estimator

Let's indulge ourselves and take a short detour to think about estimation in an abstract way

This subsection is based on Stachurski *A Primer in Econometric Theory* chapters 8.1 and 8.2

We're dealing with a random variable Z with distribution P

We're interested in a *feature* of P

Definition (Feature)

Let $Z \in L_2$ and $P \in \mathscr{P}$ where $\mathscr{P}$ is a class of distributions on Z.
A **feature** of P is an object of the form $\gamma(P)$ for some $\gamma : \mathscr{P} \to S$.

Here S is an arbitrarily flexible space (usually $\mathbb{R}$)

Examples of features: means, moments, variances, covariances

For some reason we are interested in $\gamma(P)$

If we knew P then we may be able to derive $\gamma(P)$

Example: P is standard normal and $\gamma(P) = \int Z dP = 0$
(mean of the standard normal distribution)

But we typically don't know P

If all we're interested in is $\gamma(P)$ then we may not need to know P
(unless the feature we're interested in is P itself)

Instead, we use a random sample to make an inference about a feature of P

Definition (Random Sample)

The random variables $Z_1, \ldots, Z_N$ are called a **random sample of size** N from the population P if $Z_1, \ldots, Z_N$ are mutually independent and all have probability distribution P.

The joint distribution of $Z_1, \ldots, Z_N$ is P^N by independence

We sometimes say that $Z_1, \ldots, Z_N$ are iid copies of Z

We sometimes say that $Z_1, \ldots, Z_N$ are iid random variables

By the way: Z_i could be vectors or matrices too

Definition (Statistic)

A **statistic** is any function $g : \mathbb{R}^N \to \mathbb{R}$ that maps the sample data somewhere.

The definition of a statistic is deliberately broad

It is a function that maps the sample data somewhere

Where to? Depends on the feature $\gamma(P)$ you're interested in

There are countless examples

Illustration: let $K = 1$ (i.e., univariate)

$$\begin{aligned}
\text{sample mean:} &\quad g(Z_1, \ldots, Z_N) = \sum_{i=1}^N Z_i / N =: \bar{Z}_N \\
\text{sample variance:} &\quad g(Z_1, \ldots, Z_N) = \sum_{i=1}^N (Z_i - \bar{Z}_N)^2 / N \\
\text{sample min:} &\quad g(Z_1, \ldots, Z_N) = \min\{Z_1, \ldots, Z_N\} \\
\text{answer to everything:} &\quad g(Z_1, \ldots, Z_N) = 42
\end{aligned}$$

A statistic becomes an estimator when linked to a feature $\gamma(P)$

Definition (Estimator)

An **estimator** $\hat{\gamma}$ is a statistic used to infer some feature $\gamma(P)$ of an unknown distribution P.

In other words: an estimator is a statistic *with a purpose*

Earlier example: P is the standard normal distribution
(but let's pretend we don't know this, as is usually the case)

So $Z \sim \mathrm{N}(0, 1)$

And we're interested in $\mathrm{E}Z$ so we set $\gamma(P) = \mathrm{E}Z = \int Z dP$

We have available a random sample $\{Z_1, \ldots, Z_N\}$

Each $Z_i \sim \mathrm{N}(0, 1)$, but we don't know this

But we do know: all Z_i are iid

So they must all have the same mean $\mathrm{E}Z_i$

What would be an estimator for $\mathrm{E}Z$?

Aside: there are infinitely many

What would be a *good* estimator for $\mathrm{E}Z_i$?
(perhaps not so many anymore)

Analogy Principle

A good way to create estimators is the **analogy principle**

Goldberger explains the main idea of it:

the analogy principle of estimation...proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population (Goldberger, 1968, as cited in Manski, 1988)

That is very unspecific, of course

Manski (1988) wrote an entire book on analog estimation and explains the analogy principle precisely and comprehensively

But we can illustrate it using our earlier framework

Definition (Empirical Distribution)

The **empirical distribution** P_N of the sample $\{Z_1, \ldots, Z_N\}$ is the discrete distribution that puts equal probability $1/N$ on each sample point $Z_i, i = 1, \ldots, N$.

Definition (Analogy Principle)

To estimate $\gamma(P)$ use $\hat{\gamma} := \gamma(P_N)$.

How do we use this in our example?

We wanted to estimate $\gamma(P) := \int ZdP$

According to the analogy principle, we should use $\int ZdP_N$

By definition, the empirical distribution is discrete therefore

$$\int ZdP_N = \sum_{i=1}^{N} Z_i/N =: \bar{Z}_N$$

This is, of course, the sample average and we use the conventional notation $\bar{Z}_N$

The analogy principle results in the estimator $\hat{\gamma} = \sum_{i=1}^{N} Z_i/N$

How can we use the analogy principle to estimate β^*?

Roadmap

Recall linear projection representation

$$Y = X'\beta^* + u,$$

where $X := (X_1, \dots, X_K)'$, and $X_1, \dots, X_K, Y \in L_2$

We saw that $\mathrm{E}(uX) = 0$ implied $\beta^* = \left(\mathrm{E}(XX')\right)^{-1} \mathrm{E}(XY)$

In other words: β^* is the projection coefficient

We want to estimate β^* using a random sample

The use of a random sample necessitates some changes of notation...

The random sample offers us N iid copies of the random vector X and the random variable Y

Letting i run from 1 to N, I define $X_i = (X_{i1}, \dots, X_{iK})'$

This gives us a collection if K-vectors, $X_1, \dots, X_N$

Notice that this effectively overwrites earlier notation:

- previously: X_1 denoted the first random variable in a collection of K random variables that span the subspace for the projection of Y;
- from now on: X_1 is a K-dimensional column vector collecting iid copies of the K random variables that hitherto were labelled $X_1, \dots, X_K$

Similarly, $Y_1, \dots, Y_N$ are iid copies of Y

With this new notation, the random sample is $(X_1, Y_1), \dots, (X_N, Y_N)$

Or simpler: $(X_i, Y_i), i = 1, \dots, N$ is a random sample

These are iid copies of the ordered pair (X, Y)

Given the random sample (X_i, Y_i), $i = 1, \dots, N$ we can write the linear projection representation as

$$Y_i = X_i'\beta^* + u_i,$$

Because $\mathrm{E}(uX) = 0$ we have $\mathrm{E}(u_i X_i) = 0$

Combining findings from last lecture and assignment 1:

$$\begin{aligned}
\beta^* &= \underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \, \mathrm{E}\left((Y - X'b)^2 \right) \\
&= \underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \, \mathrm{E}\left((Y_i - X_i'b)^2 \right) && (1) \\
&= \mathrm{E}(X_i X_i')^{-1} \mathrm{E}(X_i Y_i) && (2)
\end{aligned}$$

Equations (1) and (2) motivate two succinct analog estimators for β^*:

(1) the ordinary least squares estimator;

(2) the method of moments estimator

Let's look at both

If we define β^* like so:

$$\beta^* := \underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \operatorname{E}\left(\left(Y_i - X_i' b \right)^2 \right),$$

then the analogy principle suggests the estimator

$$\underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - X_i' b \right)^2$$

This seems very sensible and deserves a famous definition

Definition (Ordinary Least Squares (OLS) Estimator)

The ordinary least squares estimator is

$$\hat{\beta}^{\text{OLS}} := \underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - X_i' b \right)^2$$

It is obvious how this estimator obtained its name

When you solve this you get

$$\hat{\beta}^{\text{OLS}} = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i Y_i \right)$$

Most people, when writing vectors, use the default column notation, meaning that if I tell you that X_i is a K-dimensional vector, you automatically know it is a $K \times 1$ vector

The second way of defining an estimator for β^*, via:

$$\beta^* = \mathsf{E}(X_i X_i')^{-1} \mathsf{E}(X_i Y_i)$$

The analogy principle suggests the estimator

$$\left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} X_i Y_i$$

This also seems very sensible and deserves a familiar name:

Definition (Method of Moments (MM) Estimator)

Applying the analogy principle results in

$$\hat{\beta}^{\mathsf{MM}} = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} X_i Y_i$$

You immediately see that $\hat{\beta}^{\mathsf{OLS}} = \hat{\beta}^{\mathsf{MM}}$

I'll simply refer to it as the OLS estimator

The OLS estimator does have a compact matrix representation

Recall that $X_i := (X_{i1}, X_{i2}, \ldots, X_{iK})'$ is the K-dimensional column vector that collects the K 'regressors' for observation i

Collecting all N of these vectors in an $N \times K$ matrix:

$$X := (X_1, X_2, \ldots, X_N)' = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ X_{21} & X_{22} & \cdots & X_{2K} \\ X_{31} & X_{32} & \cdots & X_{3K} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{NK} \end{pmatrix}$$

Notice again that this effectively overwrites earlier notation:

- previously: X denoted the K-dimensional column vector that collected the K random variables representing the regressors;
- from now on: X is an $N \times K$-dimensional matrix collecting in each row the K regressors for each observation $i = 1, \ldots, N$

Similarly, from now on $Y := (Y_1, Y_2, \ldots, Y_N)'$ is the $N \times 1$ vector collecting all Y_i

The new matrix X and the new vector Y let us replace sums:
$\sum X_i X_i' = X'X$ and $\sum X_i Y_i = X'Y$

It follows that $\hat{\beta}^{\text{OLS}}$ has a nice and short matrix representation:

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

Now let's turn to the question: How good is $\hat{\beta}^{\text{OLS}}$?

What is goodness?

In the next few weeks we'll consider things such as

- bias
- variance (small sample and large sample)
- consistency
- distribution (large sample)

Roadmap

Definition (Convergence in Probability)

A sequence of random variables $Z_1, Z_2, \ldots$ **converges in probability** to a random variable Z if for all $\epsilon > 0$,

$$\lim_{N \to \infty} P(|Z_N - Z| > \epsilon) = 0.$$

We say that Z is the **probability limit** of Z_N and write $Z_N \xrightarrow{p} Z$.

Often times the probability limit Z is a degenerate random variable that takes on a constant value everywhere

Definition (Bounded in Probability)

A sequence of random variables $Z_1, Z_2, \ldots$ is **bounded in probability** if for all $\epsilon > 0$, there exists $b_\epsilon \in \mathbb{R}$ and an integer N_ϵ such that

$$P(|Z_N| \geq b_\epsilon) < \epsilon \text{ for all } N \geq N_\epsilon.$$

Here's some new notation:

- for sequences that are bounded in probability we write $Z_N = \mathrm{O}_p(1)$;
- for sequences that converge to zero in probability we write $Z_N = \mathrm{o}_p(1)$.

The 'order' (Bachmann-Landau) notation is quite handy

Here some useful rules how to work with the new notation:

Lemma

If $Z_N = c + o_p(1)$ then $Z_N = O_p(1)$ for a real constant c.

Proposition

Let $W_N = o_p(1)$, $X_N = o_p(1)$, $Y_N = O_p(1)$, and $Z_N = O_p(1)$.

$$W_N + X_N = o_p(1) \qquad W_N + Y_N = O_p(1) \qquad Y_N + Z_N = O_p(1)$$

$$W_N \cdot X_N = o_p(1) \qquad W_N \cdot Y_N = o_p(1) \qquad Y_N \cdot Z_N = O_p(1)$$

We've got a few more tricks up our sleeves

Theorem (Slutsky Theorem)

If $Z_N = c + o_p(1)$ *and* $g(\cdot)$ *is continuous at* c *then* $g(Z_N) = g(c) + o_p(1)$.

In short: $g(c + \mathrm{o}_p(1)) = g(c) + \mathrm{o}_p(1)$

That's a reason to like the plim, it passes through nonlinear functions (which is not true for expectation operators)

Corollary

$1/(c + o_p(1)) = 1/c + o_p(1)$ *whenever* $c \neq 0$.

All the definitions on the previous four slides also apply element by element to sequences of random vectors or matrices

Theorem (Weak Law of Large Numbers (WLLN))

Let $Z_1, Z_2, \ldots$ be independent and identically distributed random variables with $EZ_i = \mu_Z$ and Var $Z_i = \sigma_Z^2 < \infty$. Then

$$\frac{1}{N}\sum_{i=1}^{N} Z_i = \mu_Z + o_p(1).$$

Of course, $\frac{1}{N}\sum_{i=1}^{N} Z_i$ is the sample mean or sample average

WLLN in words:
sample mean converges in probability to population mean

Proving the WLLN is easy, using Chebyshev's inequality

Lemma (Chebyshev's Inequality)

Let Z be a random variable with $EZ^2 < \infty$ and let $g(\cdot)$ be a nonnegative function. Then for any $c > 0$

$$P\left(g(Z) \geq c\right) \leq \frac{E\left(g(Z)\right)}{c}.$$

Let $\bar{Z}_N := \frac{1}{N}\sum_{i=1}^{N} Z_i$

Here we're interested in bounding $\lim_{N\to\infty} P\left(|\bar{Z}_N - \mu_Z| > \epsilon\right)$

$$\begin{aligned} P\left(|\bar{Z}_N - \mu_Z| > \epsilon\right) &= P\left((\bar{Z}_N - \mu_Z)^2 > \epsilon^2\right) \\ &\leq \frac{\mathrm{E}(\bar{Z}_N - \mu_Z)^2}{\epsilon^2} = \frac{\mathrm{Var}\ \bar{Z}_N}{\epsilon^2} = \frac{\sigma_Z^2}{N \cdot \epsilon^2} \end{aligned}$$

which converges to zero as $N \to \infty$

We have used the fact $\mathrm{E}(\bar{Z}_N) = \mu_Z$ and Var $\bar{Z}_N = \sigma_Z^2/N$
(we remember this from undergrad metrics)

This takes us back to the analogy principle

Remember earlier:

We wanted to estimate the feature $\gamma(P) := \mathrm{E}Z = \int Z dP$

According to the analogy principle, we should use $\int Z dP_N$

This led to the estimator $\hat{\gamma} = \sum_{i=1}^{N} Z_i / N$

Immediately by the WLLN: $\hat{\gamma} \xrightarrow{\mathrm{p}} \gamma(P)$

Definition (Consistency of an Estimator)

An estimator $\hat{\gamma}$ for $\gamma := \gamma(P)$ is called **consistent** if $\hat{\gamma} \xrightarrow{\mathrm{p}} \gamma$.

Intuition: if the sample size is large, sample mean is almost equal to population mean

So there is some hope that the analogy principle leads to consistent estimators

Roadmap

Let's first show that the OLS estimator is consistent

$$\hat{\beta}^{\text{OLS}} := \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1} \sum_{i=1}^{N} X_i Y_i$$
$$= \beta^* + \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i u_i \right),$$

where we have used $Y_i = X_i'\beta^* + u_i$

Big picture to establish consistency:
want to show that second term on rhs is close to zero
(in a probabilistic sense)

Let's take a look

Copy and paste from previous slide:

$$\hat{\beta}^{\text{OLS}} = \beta^* + \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i u_i \right)$$

By WLLN

$$\frac{1}{N} \sum_{i=1}^{N} X_i X_i' = \mathsf{E}(X_i X_i') + \mathsf{o}_p(1)$$

and for its inverse:

$$\begin{aligned} \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} &= \left(\mathsf{E}(X_i X_i') + \mathsf{o}_p(1) \right)^{-1} \\ &= \mathsf{E}(X_i X_i')^{-1} + \mathsf{o}_p(1) = \mathsf{O}_p(1) \end{aligned}$$

using Slutsky's theorem, and a matrix version of the earlier Lemma that $c + \mathsf{o}_p(1) = \mathsf{O}_p(1)$, and assuming that $\mathsf{E}(X_i X_i')$ is positive definite (inverse exists)

Copy and paste again:

$$\hat{\beta}^{\mathsf{OLS}} = \beta^* + \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i u_i \right)$$

For the other factor on the rhs:

$$\frac{1}{N} \sum_{i=1}^{N} X_i u_i = \mathsf{E}(X_i u_i) + \mathsf{o}_p(1) = 0 + \mathsf{o}_p(1) = \mathsf{o}_p(1)$$

It follows

$$\begin{aligned} \hat{\beta}^{\mathsf{OLS}} &= \beta^* + \mathsf{O}_p(1) \cdot \mathsf{o}_p(1) \\ &= \beta^* + \mathsf{o}_p(1) \end{aligned}$$

In words: $\hat{\beta}^{\mathsf{OLS}}$ converges in probability to β^*

This means $\hat{\beta}^{\mathsf{OLS}}$ is a consistent estimator for the projection coefficient β^*

It illustrates the benefit of the analogy principle when it works

But what is the distribution of $\hat{\beta}^{\text{OLS}}$?

- that's a tricky one
- $\hat{\beta}^{\text{OLS}} = \beta^* + (X'X)^{-1}X'u$, what's the distribution of the second term on the rhs?
- short answer: we have no idea
- there's some suspicion that $\hat{\beta}^{\text{OLS}}$ may have an *exact* normal distribution if u is normally distributed
- but we don't know what the distribution of u is