Regression #4: Properties of OLS Estimator (Part 2)

Econ 671

Purdue University

Introduction

- In this lecture, we continue investigating properties associated with the OLS estimator.

- Our focus now turns to a derivation of the *asymptotic normality* of the estimator as well as a proof of a well-known efficiency property, known as the *Gauss-Markov Theorem*.

Asymptotic Normality

To begin, let us consider the regression model when the error terms are normally distributed:

$$y_i = x_i\beta + \epsilon_i, \qquad \epsilon | X \sim \mathcal{N}(0, \sigma^2 I_n).$$

In this case, the sampling distribution of $\hat{\beta}$ (given X) is *immediate:*

-

Asymptotic Normality

Since

$$\epsilon | X \sim \mathcal{N}\left(0, \sigma^2 I_n\right),$$

it follows that

-

Thus, the sampling distribution follows a *normal* distribution, with the mean and covariance matrix derived in the previous lecture.

Asymptotic Normality

- In many cases, however, we do not want to assume that the errors are normally distributed.

- If we replace the Gaussian assumption with something different, however, it can prove to be quite difficult to determine the exact (finite sample) sampling distribution of the OLS estimator.

- Instead, we can look for a large sample approximation that works for a variety of different cases. The approximation will be exact as $n \rightarrow \infty$, and we will take it as a reasonable approximation in data sets of moderate or small sizes.

Asymptotic Normality

With a minor abuse of the theorem itself, we first introduce the Lindberg-Levy CLT:

Theorem

Asymptotic Normality

It remains for us to figure out how to apply this result to (approximately) characterize the sampling distribution of the OLS estimator.

To this end, let us write:

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

Rearranging terms and multiplying both sides by a $\sqrt{n}$, we can write

-

and we note from our very first lecture that

-

Asymptotic Normality

Thus, the term $(\sqrt{n})^{-1}X'\epsilon$ can be written as:

-

In this last form, we can see that this term is simply a sample average of $k \times 1$ vectors $x_i'\epsilon_i$, scaled by $\sqrt{n}$. Such quantities fall under the "jurisdiction" of the Lindberg-Levy CLT.

Asymptotic Normality

Specifically, we can apply this CLT once we characterize the mean and covariance matrix of the terms appearing within the summation. To this end, note:

1.

   and

2.

Hence, we can apply the Lindberg-Levy CLT to give:

-

Asymptotic Normality

As for the other key term appearing in our expression for $\sqrt{n}(\hat{\beta} - \beta)$, we note:

-

so that

-

OK, so let's review:

-

Asymptotic Normality

Based on earlier derivations, the right hand side (Slutsky) must converge in distribution to:

-

or

-

We can then write:

Asymptotic Normality

In practice, we replace the unknown population quantity

$$\left[E_x(x_i'x_i)\right]^{-1}$$

with a consistent estimate:

$$\left[\frac{1}{n}\sum_i x_i'x_i\right]^{-1} = \left[\frac{1}{n}X'X\right]^{-1} \xrightarrow{p} \left[E_x(x_i'x_i)\right]^{-1}.$$

Thus,

-

We can also get an asymptotic result for the quadratic form:

-

Asymptotic Normality

- Note that we did not assume normality to get this result; provided the assumptions of the regression model are satisfied, the sampling distribution of $\hat{\beta}$ will be *approximately* normally distributed.

- This result will form the basis for testing hypotheses regarding β, as we will discuss in the following lectures.

Gauss-Markov Theorem

We now move on to discuss an important result, related to the *efficiency* of the OLS estimator, known as the *Gauss-Markov Theorem.*

This theorem states:

-

Gauss-Markov Theorem

We will first prove this result for *any* linear combination of the elements of β.

That is, suppose we seek to estimate

-

where c is an arbitrary $k \times 1$ selector vector. For example,

-

would select the intercept parameter. We seek to show that the OLS estimator of μ,

-

has a variance at least as small as any other linear, unbiased estimator of μ.

Gauss-Markov Theorem

To establish this result, let us first consider *any* other linear, unbiased estimator of μ. Call this estimator h. *Linearity* implies that h can be written in the form:

-

for some $n \times 1$ vector a.

We note that

-

Gauss-Markov Theorem

For *unbiasedness* to hold, it must be the case that

-

or (since this must apply for any β and c):

-

Now,

-

Gauss-Markov Theorem

The variance of our candidate estimator is:

-

Comparing these, we obtain:

-

Clearly, this is greater than or equal to zero, right?

Gauss-Markov Theorem

Actually it is. To see this, note:

-

The last line follows as the product represents a sum of squares.

Gauss-Markov Theorem

Does this result hold unconditionally?

-

Gauss-Markov Theorem

We will now prove this in a more general way, by directly comparing the covariance *matrices* between the two estimators.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ. We would say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$ if the difference between the covariance matrices is *negative semidefinite*. That is, for any $k \times 1$ vector $x \neq 0$,

$$x' \left(\text{Var}(\hat{\theta}_1) - \text{Var}(\hat{\theta}_2) \right) x \leq 0.$$

This implies that element-by-element (and in terms of linear combinations), that $\hat{\theta}_1$ is preferable to $\hat{\theta}_2$.

Gauss-Markov Theorem

Consider any other linear estimator of β,

-

where A^* is $k \times n$ and nonstochastic, given X. In terms of unbiasedness,

-

so that unbiasedness implies

-

Write

-

where D is arbitrary. We then note:

-

Gauss-Markov Theorem

Similarly,

-

The condition that $A^*X = I_k$ must mean that $DX = 0$. This makes all the cross terms in the above vanish since, for example,

-

Therefore,

-

Gauss-Markov Theorem

Let us now consider the variance of our candidate estimator:

-

Taking this further,

-

The matrix DD' is postive semidefinite, since $x'DD'x$ is again a sum of squares. This difference is strictly positive unless $D = 0$, in which case $\tilde{\beta} = \hat{\beta}$.

We conclude that $\hat{\beta}$ is more efficient than $\tilde{\beta}$.