Lecture 11 Multiple Regression Analysis - OLS Asymptotics

Recap

Let's recall our population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U.$$

We made several assumptions about it.

1. Linear CEF
    1. Linear model
    2. Error term is mean-independent of predictors
2. Random Sampling
3. No Perfect Collinearity
4. Homoskedasticity
5. Normality

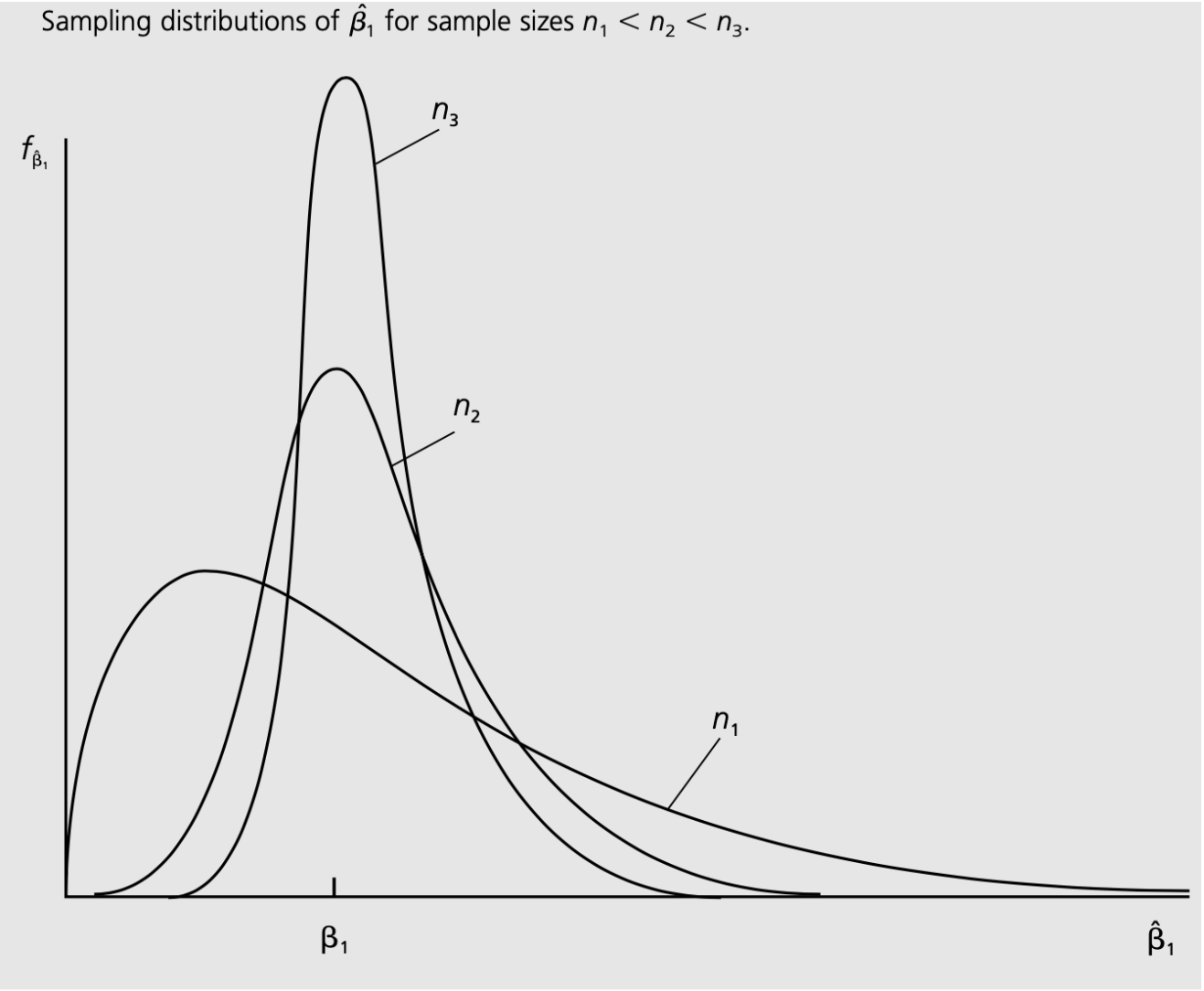
We call the first four (or five if you break down the first assumption into two separate ones) assumptions the *Gauss-Markov* assumptions. We call the full set of assumptions the *classical linear model* assumptions. We showed that under the first three assumptions (the Gauss-Markov assumptions without homoskedasticity) the OLS estimator is *unbiased*. We showed that under the full set of Gauss-Markov assumptions the OLS estimator is the best linear unbiased estimator (the *Gauss-Markov theorem*) (has the smallest variance). Once we added the normality assumption, we derived the *exact sampling distribution* of the OLS estimator (normal distribution), as well as the exact distributions of the various test statistics, which opened the way for hypotheses testing.

These properties of the OLS estimator are often called *finite sample* or *small sample* or *exact* properties. In this lecture, we will study the *asymptotic* (or *large sample*) properties of the OLS estimator and test statistics. The asymptotic properties are not defined for a particular sample size, they are defined as the sample size grows without bound. The main takeaway from this analysis will be that even without the normality assumption, the test statistics that we discussed (t and F) will approximately follow their corresponding distributions.

Consistency

Recall that under our assumptions, the OLS estimator is unbiased. Unbiasedness is a nice property to have, although it cannot always be achieved. On the other hand, there is another property that an estimator *must* have to be useful. This property is called *consistency*. The famous econometrician Clive W.J. Granger once remarked: "If you can't get it right as n goes to infinity, you shouldn't be in this business." The idea is that if an estimator is not consistent, it is unlikely to be useful. But what is consistency?

Suppose we have an OLS estimator, $\hat{\beta}_j$ of some population coefficient β_j. For each sample size n, $\hat{\beta}_j$ has a probability distribution. Because $\hat{\beta}_j$, is unbiased under our assumptions, this distribution has mean β_j. *Consistency* means that the distribution of $\hat{\beta}_j$ becomes more and more tightly distributed around β_j as the sample size grows. As n goes to infinity, the distribution of $\hat{\beta}_j$ collapses to the single point β_j. In other words, if an estimator is consistent, we can make our estimator arbitrarily close to β_j if we can collect as much data as we want. This idea is illustrated on the picture below.



Formally, consistency is defined as follows. Let $\hat{\theta}_n$, be an estimator of some population parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is a consistent estimator of θ, if for every $\epsilon > 0$

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

When $\hat{\theta}_n$ is consistent, we also say that θ is the probability limit of $\hat{\theta}_n$, written as $\operatorname{plim} \hat{\theta}_n = \theta$.

Unbiasedness and consistency are related. Unbiased estimators whose variances shrink to zero as the sample size grows are consistent. Formally, if $\hat{\theta}_n$ is an unbiased estimator of θ and $\operatorname{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\operatorname{plim} \hat{\theta}_n = \theta$.

It turns out that the OLS estimator is, in fact, consistent, which is summarized in the following result.

ⓘ Theorem: Consistency of OLS

Under assumptions 1-3, the OLS estimator $\hat{\beta}_j$ is consistent for β_j, for all $j = 0, \ldots, k$.

It is easy to show consistency using the relationship between unbiasedness and consistency. We know that the OLS estimator is unbiased. We also know that the variance of each $\hat{\beta}_j$ is

$$\text{Var}(\hat{\beta}_j \mid \mathbf{X}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}.$$

Notice that $SST_j \equiv \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = n\text{Var}(X_j)$. Therefore,

$$\text{Var}(\hat{\beta}_j \mid \mathbf{X}) = \frac{\sigma^2}{n\text{Var}(X_j)(1 - R_j^2)}.$$

Recall that the assumption of no perfect collinearity implies that $1 - R_j^2$ is non-zero. We also need to assume that $\text{Var}(X_j)$ is finite. The assumption of no perfect collinearity also guarantee that it is non-zero. It is now easy to see that as n grows, the variance shrinks to zero, which in turn implies that $\hat{\beta}_j$ is consistent for β_j.

We can also show consistency of the OLS more directly. For example, recall that in the simple linear regression the slope coefficient is

$$\hat{\beta}_1 = \frac{\widehat{Cov}(Y, X)}{\widehat{Var}(X)}.$$

Using the law of large numbers, we can re-write this expression as

$$\text{plim}\ \hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}.$$

✎ The law of large numbers

Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}\ \bar{Y}_n = \mu$$

Now substitute for Y

$$\text{plim}\ \hat{\beta}_1 = \frac{\text{Cov}(\beta_0 + \beta_1 X + U, X)}{\text{Var}(X)}.$$

We have that $\text{Cov}(\beta_0, X) = 0$, $\text{Cov}(\beta_1 X, X) = \beta_1 \text{Var}(X)$, therefore

$$\text{plim}\ \hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)}.$$

Since $\text{Cov}(X, U) = 0$ by our first assumption, $\text{plim}\ \hat{\beta}_1 = \beta_1$.

Inconsistency of the OLS

Just as failure of the first assumption causes bias in the OLS estimators, correlation between the error term and any of the predictors generally causes *all* of the OLS estimators to be inconsistent. The bias will persist even as the sample size grows.

In the simple regression case, we can obtain the inconsistency of the OLS from the derivations we made before. The inconsistency (or sometimes called the *asymptotic bias*) is

$$\text{plim}\ \hat{\beta}_1 - \beta_1 = \frac{\text{Cov}(X, U)}{\text{Var}(X)}.$$

Because $\text{Var}(X) > 0$, the inconsistency in $\hat{\beta}_1$ is positive if X and U are positively correlated, and the inconsistency is negative if X and U are negatively correlated. If the covariance between X and U is small relative to the variance in X, the inconsistency can be negligible. However, we cannot estimate how big the covariance is because U is unobserved.

We can use this formula to derive the asymptotic analogue of the omitted variables bias. For example, suppose the true population model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + V.$$

Now suppose that we instead estimate a model without X_2

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

where $U \equiv V + \beta_2 X_2$.

Then the inconsistency in the OLS estimator of $\hat{\beta}_1$ from the incorrect model is

$$\begin{aligned}
\text{plim}\ \tilde{\beta}_1 &= \beta_1 + \frac{\text{Cov}(X_1, U)}{\text{Var}(X_1)} = \beta_1 + \frac{\text{Cov}(X_1, V + \beta_2 X_2)}{\text{Var}(X_1)} \\
&= \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)} \\
&= \beta_1 + \beta_2 \rho_2,
\end{aligned}$$

where $\rho_2 \equiv \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$.

Thus, for practical purposes, we can view the inconsistency as being the same as the bias. The difference is that the inconsistency is expressed in terms of the population variance of X_1 and the population covariance between X_1 and X_2, while the bias is based on their sample counterparts.

If X_1 and X_2 are uncorrelated (in the population), then $\rho_2 = 0$, and $\tilde{\beta}_1$ is a consistent estimator of β_1 (although not necessarily unbiased). If X_2 has a positive partial effect on Y ($\beta_2 > 0$), and X_1 and X_2 are positively correlated ($\rho_2 > 0$) , then the inconsistency is positive,

and so on. If the covariance between X_1 and X_2 is small relative to the variance of X_1 the inconsistency can be small.

Asymptotic normality

Consistency of an estimator is an important property, but consistency by itself does not allow us to perform statistical inference. For testing, we need the sampling distribution of the OLS estimators. Under the classical linear model assumptions, we showed that the sampling distribution of the OLS is normal. We then used this result to derive the distributions of the t and F statistics.

The exact normality of the OLS estimator hinges crucially on the normality of the of the error. But normality is a strong assumption. What happens if it fails? The the OLS will not be normally distributed, which means that the t statistics will not have t distributions and the F statistics will not have F distributions. This is a potentially serious problem because our inference relies on obtaining the critical values of p values from those specific distributions. Notice, however, that normality will not affect the unbiasedness or consistency of OLS, nor does it affect the conclusion that OLS is the best linear unbiased estimator.

Recall that the normality assumption is equivalent to saying that the distribution of an outcome given predictors is normal. Since the outcome is observed (unlike the error term), in a particular application, it is much easier to think about whether the distribution of the outcome is likely to be normal. A normally distributed random variable is symmetrically distributed about its mean, it can take on any positive or negative value (but with zero probability), and more than 95% of the area under the distribution is within two standard deviations. Many outcome variables do not have these properties. Does this mean that we cannot use the t and F tests for models with these outcome variables?

ⓘ Theorem: Asymptotic normality of OLS

Under the Gauss-Markov assumptions,

1. $\sqrt{n}(\hat{\beta}_j - \beta_j) \dot{\sim} \mathcal{N}(0, \sigma^2/a_j^2)$, where $a_j^2 \equiv \text{plim} \sum_{i=1}^n \hat{u}_{ij}^2/n = \text{plim}\, SSR_j/n$, and $\hat{u}_{ij}$ are the residuals from the regression of X_j on all other predictors
2. $\hat{\sigma}^2$ is a consistent estimator of the $\text{Var}(U)$
3. For each j,

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \dot{\sim} \mathcal{N}(0, 1).$$

Asymptotic normality

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of random variables, such that for all numbers z

$$\mathbb{P}(Z_n \leqslant z) \to \Phi(z) \text{ as } n \to \infty,$$

where $\Phi(z)$ is the standard normal CDF. Then Z_n is said to have an *asymptotic standard normal distribution*. We could write this as

$$Z_n \dot{\sim} \mathcal{N}(0,1)$$

The fact that the expected value of $\sqrt{n}(\hat{\beta}_j - \beta_j)$ is zero follows from the unbiasedness of the OLS. The asymptotic variance follows from the formula for the variance of $\hat{\beta}_j$:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)} = \frac{\sigma^2}{SST_j \frac{SSR_j}{SST_j}} = \frac{\sigma^2}{SSR_j}.$$

Then

$$\text{Var}(\sqrt{n}(\hat{\beta}_j - \beta_j)) = n\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SSR_j/n}.$$

The proof of the asymptotic normality is somewhat involved and is omitted. It is based on the *central limit theorem*. The theorem above is useful because the normality assumption has been dropped. The only restriction on the distribution of the error is that it has finite variance.

The central limit theorem

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample with mean μ and variance σ^2, then

$$Z_n \equiv \frac{\bar{Y}_n - \mu}{\sigma/n}$$

has an asymptotic standard normal distribution.

Notice how the standard normal distribution appears in the theorem, as opposed to the t_{n-k-1} distribution as before. This is because the distribution is only approximate. By contrast before the distribution of the ratio in was exactly t_{n-k-1} for any sample size. From a practical perspective, this difference is irrelevant. In fact, it is just as legitimate to write

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \dot{\sim} t_{n-k-1}$$

since t_{n-k-1} approaches the standard normal distribution as the degrees of freedom gets large.

This result tells us that t testing and the construction of confidence intervals are carried out exactly as under the classical linear model assumptions. If the sample size is not very large, then the t distribution can be a poor approximation to the distribution of the t statistics when U is not normally distributed. Unfortunately, there are no general prescriptions on how big the sample size must be before the approximation is good enough. It is important to realize that the theorem above does require the homoskedasticity assumption. Without it,

the usual t statistics and confidence intervals are invalid no matter how large the sample size is.

The asymptotic normality of the OLS estimators also implies that the F statistics have approximate F distributions in large samples. Thus, for testing exclusion restrictions or other multiple hypotheses, nothing changes from what we have done before.