The Finite Sample Properties of the Least Squares Estimator / Basic Hypothesis Testing

Greene Ch 4, Kennedy Ch. 2
R script *mod1s3*

To assess the quality and appropriateness of econometric estimators, we are always interested in their statistical properties. For most estimators, these can only be derived in a "large sample" context, i.e. by imagining the sample size to go to infinity. The statistical attributes of an estimator are then called "*asymptotic properties*". However, for the CLRM and the OLS estimator, we can derive statistical properties for any sample size, i.e. its "*small sample*" properties (Naturally, we can also derive its asymptotic properties. We'll do this later in this course)

Even better, we can derive some statistical properties for $\mathbf{b_{OLS}}$ without referring to a specific distribution of the error term.

Unbiasedness

The OLS estimator is unbiased conditional on **X**, and unconditionally. The latter implies that it is unbiased for every sample.

To show unbiasedness, we can first write the OLS estimator as follows:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \tag{1}$$

We then take the expectation of **b**, conditional and unconditional on **X**:

$$\begin{aligned} E\left(\mathbf{b}\,|\,\mathbf{X}\right) &= \boldsymbol{\beta} + E\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}E\left(\mathbf{X}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} \\ E\left(\mathbf{b}\right) &= E_X\left(E_\varepsilon\left(\mathbf{b}\,|\,\mathbf{X}\right)\right) = E_X\left(\boldsymbol{\beta}\right) = \boldsymbol{\beta} \end{aligned} \tag{2}$$

Note the use of Assumption 3 in the first line: $E\left(\mathbf{X}'\boldsymbol{\varepsilon}\right) = \mathbf{0}$. By doing so, we can derive the statistical expectation of **b** without referring to **b**'s actual distribution – very convenient. Otherwise, we would have to solve $E\left(\mathbf{b}\,|\,\mathbf{X}\right) = \int_{\mathbf{b}} \mathbf{b} * f\left(\mathbf{b}\,|\,\mathbf{X}\right) d\mathbf{b}$, the generic expression for an expectation, for some explicit multi-variate density $f\left(\mathbf{b}\,|\,\mathbf{X}\right)$.

See **R** script *mod1s3* for an illustration of "unbiasedness".

Variance and Gauss Markov Theorem (GMT)

Treating **X** as given, we can derive the *sampling variance* (or, more appropriately, Variance-Covariance Matrix) of **b** as follows:

$$V(\mathbf{b}\mid\mathbf{X}) = E\left((\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'\mid\mathbf{X}\right) = E\left((\mathbf{X'X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X'X})^{-1}\mid\mathbf{X}\right) =$$
$$= (\mathbf{X'X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mid\mathbf{X})\mathbf{X}(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1} \tag{3}$$

As shown in Greene, p.59/60, this variance is never larger than the variance of any other linear estimator of $\boldsymbol{\beta}$. Thus, we say that **b** is the "best (i.e. minimum variance) linear unbiased estimator", or **b** is "BLUE". The *Gauss-Markov Theorem* (Greene p. 60) formalizes this notion.

As is clear from (3) the unconditional variance of **b** cannot be exactly determined without specifying the distribution of **X** (which we usually don't know). By the decomposition-of-variance theorem (Greene B-69):

$$V(\mathbf{b}\mid\mathbf{X}) = E_{\mathbf{X}}\left(V(\mathbf{b}\mid\mathbf{X})\right) + V_{\mathbf{X}}\left(E(\mathbf{b}\mid\mathbf{X})\right) \tag{4}$$

The last expectation term is a constant ($\boldsymbol{\beta}$), so its variance equals zero. This leaves us with

$$V(\mathbf{b}) = E_{\mathbf{X}}\left(V(\mathbf{b}\mid\mathbf{X})\right) = \sigma^2 E_{\mathbf{X}}\left((\mathbf{X}'\mathbf{X})^{-1}\right) \tag{5}$$

Naturally, this will depend on the distribution of **X** in the population. However, the GMT extends to the unconditional case, i.e. holds for ANY **X.**

Estimating the variance of the OLS Estimator

Note that the formula for the sampling variance of **b** includes the variance of the error terms, σ^2, which is usually unknown. We thus have to estimate it. A natural candidate might be the "sample variance" of the residuals, i.e. **e**'**e**/n (we continue to assume that the regression has an intercept s.t. the sum and thus the mean of residuals is zero). Let's check if this estimator is unbiased:

$$\mathbf{e} = \mathbf{My} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon} \qquad \text{since} \quad \mathbf{MX} = \mathbf{0}$$
$$\Rightarrow \mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \tag{6}$$

We now seek the expectation of the last term (conditional on **X**, which I will tacitly assume throughout). Note that $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ is a scalar, thus its value is equal to its trace (= sum of diagonal elements of a matrix). We can thus use the popular "*trace trick*" to derive the expectation:

$$E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) = E\left(tr(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})\right) = E\left(tr(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right) = tr\left(\mathbf{M}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right) = \sigma^2 tr(\mathbf{M}) \tag{7}$$

The trace of M is (*n*-*k*) (see Greene p. 51), so we get

$$E(\mathbf{e}'\mathbf{e}/n) = \left(\frac{n-k}{n}\right)\sigma^2 \tag{8}$$

So this estimator is slightly biased towards zero, although the bias is small for large samples (and small *k*). However, we can do better by using

$$s^2 = \mathbf{e}'\mathbf{e}/(n-k) \tag{9}$$

as estimator for σ^2. You can easily verify that this estimator is unbiased. The square root of this estimator, "*s*", is often called the "*standard error of the regression*". We can now derive the estimated variance of b as

$$\hat{V}(\mathbf{b}) = s^2(\mathbf{X}'\mathbf{X})^{-1} \tag{10}$$

This is a *k* by *k* symmetric matrix. The diagonal of this matrix gives the estimated variance for each element of **b.** The square root of this element-specific estimated variance is known as the *standard error* for a given coefficient. You've seen them in regression print-outs. Now you know where they're coming from.

The normality assumption and hypothesis testing

So far, we were able to show that **b** is unbiased and has variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (estimated via $s^2(\mathbf{X}'\mathbf{X})^{-1}$). However, we can't say anything about the full distribution of **b** without invoking Assumption 5: $\boldsymbol{\varepsilon} \sim n(\mathbf{0}, \sigma^2\mathbf{I})$. By properties of the normal distribution, it follows that

$$\mathbf{b} \sim n\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right) \tag{11}$$

Details: We know that $\varepsilon \sim n(0, \sigma^2)$ and that we can express **b** as $\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ (see (1)). Thus, **b** can be considered a linear function of $\boldsymbol{\varepsilon}$. By the properties of the normal (or multivariate normal) density, a random variable (or vector) that is a linear function of a normally distributed variable (vector) will also be normally distributed. Note that we didn't need this assumption to derive the moments (i.e. mean, variance) of **b**. An alternative way to derive the variance for **b** would be:

$$\begin{aligned}
&V(\mathbf{b}) = V\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right) = V\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right) && \text{since} \quad V(\boldsymbol{\beta}) = 0 \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} && \text{by} \quad \mathrm{V}(\mathbf{Az}) = \mathbf{A}V(\mathbf{z})\mathbf{A}' \\
& && \text{(e.g. Greene B-87, p. 1040)} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned} \tag{12}$$

We can now use this additional piece of information (or better: "assumption") to derive hypothesis tests for the elements of **b**. Let b_k be the k^{th} element of **b** and S^{kk} be the k^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Then we have:

$$b_k \sim n(\beta_k, \sigma^2 S^{kk}) \Rightarrow (b_k - \beta_k) \sim n(0, \sigma^2 S^{kk}) \Rightarrow \frac{(b_k - \beta_k)}{\sqrt{\sigma^2 S^{kk}}} = z_k \sim n(0,1) \tag{13}$$

If we knew σ^2 we could use the standard normal ("z") distribution to draw statistical inferences about β_k.

Quick Recap of a hypothesis test

For example, if we wanted to test the null hypothesis H_0: $\beta_k \le a$ against the alternative hypothesis H_a: $\beta_k > a$ where *a* is some specific value, we could proceed as follows: First, we argue that if we reject that $\beta_k = a$ in favor of $\beta_k > a$, we would certainly also reject $\beta_k \le a$. Then we argue: If $\beta_k = a$, this implies that that $E(b_k) = a$. Thus, the further away b_k is from *a*, the less likely we have $\beta_k = a$. More specifically, the further to the right b_k is from *a*, the less likely we have $\beta_k \le a$. This can be re-stated as "the further $z_k = \frac{b_k - a}{\sqrt{\sigma^2 S^{kk}}}$ from 0, the less likely we have $\beta_k = a$", and " the further to the right z_k is from 0, the less likely we have $\beta_k \le a$". The analogous reasoning holds for the left tail of the distribution if the inequality sign is reversed in the null.

So what is "too far to the right" to reject the null? That's where conventional threshold values come in. These are different for different distributions, and also change with the underlying level of confidence we'll have in our test result. For hypothesis tests, though, we prefer to work with (1-level of confidence), which we call the level of significance (usually denoted as α). Common levels of significance are 0.1, 0.05, and 0.01. The smaller the level of significance, the higher the level of confidence. Each level of significance comes with a "critical value" in the underlying distribution.

For a test of H_0: $\beta_k = a$ (called a "two-tailed test"), we need to consider both tails, and two (symmetric) critical values. This splits the level of significance in half for each tail, and pushes the critical values further from the mean.

For the z-distribution:

α	critical value, one tailed (absolute values)	critical value, two-tailed (absolute values)
0.1	1.28	1.645
0.05	1.645	1.96
0.01	2.33	2.575

Examples:

H_0	b_k	$\sqrt{\sigma^2 S^{kk}}$	z_k	α	z_{crit}	decision
$\beta_k \ge 2$	0.8	1.2	-1	0.05	1.645	do not reject
$\beta_k \le 2$	4.4	1.2	2	0.05	1.645	reject
$\beta_k = 2$	4.4	1.2	2	0.05	1.96	reject

There are two ways to draw test conclusions: compare the computed test value (here z_k) to the critical value (can be looked up in tables, usually known for common distributions), OR compare the *p*-value to α. The *p*-value is the area under the tail of the distribution further out from the critical value.

Thus: Reject if $|z_k| > z_{crit}$. Equivalently, reject if $p < \alpha$. STATA reports *p*-values for all coefficients with its regression output. In Matlab, you can compute *p*-values for a given derived statistic via built-in cumulative distribution functions, such as `tcdf(.)`, `chi2cdf(.)`, etc. In ***R***, use `pnorm()`, `tnorm(),pchisq(),` etc.

The t-test

That said, we still have the problem of not knowing σ. This means that we can't work with z-values, z-distributions, and z-tables to perform hypothesis tests involving $\boldsymbol{\beta}$. However, we can work with

$$t_k = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \tag{14}$$

which follows a standard *t*-distribution with (*n*-*k*) degrees of freedom (the *t*-distribution has 3 parameters: mean, variance, and D.o.F. The standard *t*-distribution has mean 0 and variance 1). We can thus proceed with hypothesis testing as above, replacing σ^2 with s^2, and *z* with *t*. The critical values for the conventional levels of α can be looked up in a *t*-table. As shown in detail in Greene section 5.4.1., the assumption of normality of $\boldsymbol{\varepsilon}$ is critical in deriving this *t*-value.

Using the t-distribution, we can also compute a $100(1-\alpha)\%$ confidence interval for a given β_k as:

$$C.I._{(1-\alpha)} = \left\{ b_k - t_{\alpha/2} s_{b_k} \le \beta_k \le b_k + t_{\alpha/2} s_{b_k} \right\} \qquad \text{where}$$
$$s_{b_k} = \sqrt{s^2 S^{kk}} \tag{15}$$

At this point, we have discussed all elements of a standard regression output. We will return to the topic of hypothesis testing later in this course.