Chapter 6

Asymptotic Properties of Least Squares

We have learned some basic asymptotic theory in the previous chapter. We apply these results to study asymptotic properties of the OLS estimator $\widehat{\beta} = (X'X)^{-1} X'Y$, which is of key interest in our course. We will show (i) $\widehat{\beta}$ is a consistent estimator of the linear projection coefficient β; (ii) $\widehat{\beta}$ is asymptotically normal; (iii) the asymptotic normality allows asymptotic inference of β; (iv) under what condition the variance components in the test statistic can be consistently estimated so that the testing procedure is make feasible.

6.1 Consistency

Consistency is the most basic requirement for estimators in large sample. Intuitively, it says that when the sample size is arbitrarily large, a desirable estimator should be arbitrarily close (in the sense of convergence in probability) to the population quantity of interest. Otherwise, if an estimator still deviates from the object of interest under infinite sample size, it is hard to persuade other researchers to use such an estimator unless compelling justification is provided.

Definition 6.1 (Consistency)**.** For a generic estimator $\widehat{\theta}$, we say $\widehat{\theta}$ is *consistent* for θ if $\widehat{\theta} \xrightarrow{p} \theta$, where θ is some non-random object.

In OLS, we say $\widehat{\beta}$ is *consistent* if $\widehat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$, where β is the linear projection coefficient of the population model $y_i = x_i'\beta + e_i$ with $E[x_i e_i] = 0$. To verify consistency, we write

$$\widehat{\beta} - \beta = (X'X)^{-1} X'e = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i e_i. \tag{6.1}$$

For simplicity, in this chapter we discuss the iid setting only. The first term, by LLN,

$$\widehat{Q} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q := E\left[x_i x_i'\right].$$

Here $\widehat{Q}$ is the sample mean of $x_i x_i'$ and Q is the population mean of $x_i x_i'$.

The second term, again by LLN,

$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i \xrightarrow{p} 0.$$

The continuous mapping theorem immediately implies

$$\widehat{\beta} - \beta \xrightarrow{p} Q^{-1} \times 0 = 0.$$

The OLS estimator $\widehat{\beta}$ is a consistent estimator of β.

Remark 6.1. No matter whether $(y_i, x_i)_{i=1}^n$ is an iid, or inid, or dependent sample, consistency holds as long as the convergence in probability holds for the above two expressions and Q is an invertible matrix.

6.2 Asymptotic Distribution

In finite sample, $\widehat{\beta}$ is a random variable. We have shown the distribution of $\widehat{\beta}$ under normality before. Without the restrictive normality assumption, how can we characterize the randomness of the OLS estimator?

We know from the previous section that $\hat{\beta} - \beta \xrightarrow{p} 0$ degenerates to a constant. To study its distribution, we must scale it up by a proper multiplier so that in the limit it neither degenerates nor explodes. The suitable

scaling factor is $\sqrt{n}$, as in a CLT.

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}x_{i}x_{i}'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_{i}e_{i}.$$

Since $E\left[x_{i}e_{i}\right]=0$, we apply a CLT to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_{i}e_{i}\xrightarrow{d}N\left(0,\Sigma\right)$$

where $\Sigma=E\left[x_{i}x_{i}'e_{i}^{2}\right]$. By the continuous mapping theorem,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}Q^{-1}\times N\left(0,\Sigma\right)\sim N\left(0,\Omega\right) \tag{6.2}$$

where $\Omega=Q^{-1}\Sigma Q^{-1}$ is called the *asymptotic variance*. This result is the *asymptotic normality* of the OLS estimator.

The asymptotic variance $\Omega=Q^{-1}\Sigma Q^{-1}$ is called of the *sandwich form*. It can be simplified under conditional homoskedasticity $E\left[e_{i}^{2}|x_{i}\right]=\sigma^{2}$ for all i, which gives

$$\Sigma=E\left[x_{i}x_{i}'e_{i}^{2}\right]=E\left[x_{i}x_{i}'E\left[e_{i}^{2}|X\right]\right]=\sigma^{2}E\left[x_{i}x_{i}'\right]=\sigma^{2}Q.$$

In this case, $\Omega=Q^{-1}\Sigma Q^{-1}=\sigma^{2}Q^{-1}$, and thus

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\sigma^{2}Q^{-1}\right). \tag{6.3}$$

Remark 6.2. If we are interested in the k-th parameter β_k, then the joint distribution in ((6.2)) implies

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right) &= \sqrt{n}\eta_k'\left(\widehat{\beta} - \beta\right) \\ &\xrightarrow{d} N\left(0, \sigma^2\eta_k' Q^{-1}\eta_k\right) \sim N\left(0, \sigma^2[Q^{-1}]_{kk}\right), \end{aligned} \tag{6.4}$$

where $\eta_k = (0, \ldots, 0, 1, 0 \ldots, 0)'$ is the selector of the k-th element.

Remark 6.3. If $\Omega^{-1/2}$ is multiplied on both sides of (6.2), we have

$$\Omega^{-1/2}\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, I_K\right). \tag{6.5}$$

We say the asymptotic distribution in (6.5), $N\left(0, I_K\right)$, is *pivotal* because it does not involve any unknown parameter. In contrast, the asymptotic distribution in (6.2) is not pivotal because Ω is unknown in $N\left(0, \Omega\right)$. If we are interested in the k-th parameter β_k, we can write (6.5) into the pivotal form as

$$\frac{\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right)}{\sqrt{\sigma^2[Q^{-1}]_{kk}}} \xrightarrow{d} N\left(0, 1\right). \tag{6.6}$$

6.3 Asymptotic Inference

Up to now we have derived the asymptotic distribution of $\widehat{\beta}$. However, (6.2) or (6.5) will be useful for statistical inference only if Ω is known. In reality Ω is mostly unknown, and therefore we will need to estimate it to

make statistical inference feasible. Suppose $\tilde{\Omega}$ is any consistent estimator for Ω in that $\tilde{\Omega} \xrightarrow{p} \Omega$. When we replace Ω in (6.5) with $\tilde{\Omega}$, we have

$$\tilde{\Omega}^{-1/2}\sqrt{n}\left(\widehat{\beta}-\beta\right)=\tilde{\Omega}^{-1/2}\Omega^{1/2}\times\Omega^{-1/2}\sqrt{n}\left(\widehat{\beta}-\beta\right).$$

Because Ω is positive definite, we have the first factor $\tilde{\Omega}^{-1/2}\Omega^{1/2} \xrightarrow{p} I_K$ by the continuous mapping theorem. The second factor is asymptotic normal by (6.5). Thus Slutsky's theorem implies

$$\tilde{\Omega}^{-1/2}\sqrt{n}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N\left(0, I_K\right) \tag{6.7}$$

and (6.7) is a feasible statistic for asymptotic inference.

The next question is how to consistently estimate $\Omega = Q^{-1}\Sigma Q^{-1}$, or equivalent how to come up with an $\tilde{\Omega}$. We have had $\widehat{Q} \xrightarrow{p} Q$. If we have a consistent estimator $\tilde{\Sigma}$ for Σ, then we can plug in these consistent estimators to form $\tilde{\Omega} = \widehat{Q}^{-1}\tilde{\Sigma}\widehat{Q}^{-1}$. The tricky question is how to consistently estimate $\Sigma = E\left[x_i x_i' e_i^2\right]$. We cannot use the sample mean of $x_i x_i' e_i^2$ to estimate Σ because e_i is unobservable. Under homoskedasticity $\Omega = Q^{-1}\Sigma Q^{-1} = \sigma^2 Q^{-1}$, and similarly we cannot use the sample mean of e_i^2 to estimate σ^2.

Example 6.1. Heteroskedasticity is ubiquitous in econometrics. A regression example that naturally generates conditional heteroskedasticity is the *linear probability model* $y_i = x_i'\beta + e_i$, where $y_i \in \{0, 1\}$ is a binary de-

pendent variable. Assume CEF as $E[y_i|x_i] = x_i'\beta$, so we can use OLS to consistently estimate β. The conditional variance

$$\text{var}[e_i|x_i] = \text{var}[y_i|x_i] = E[y_i|x_i](1 - E[y_i|x_i]) = x_i'\beta(1 - x_i'\beta)$$

explicitly depends on x_i. In other words, the conditional variance varies with x_i.

Naturally, one may attempt to use the OLS residual $\widehat{e}_i = \widehat{y}_i - x_i'\widehat{\beta}$ to replace the regression error e_i, so that we would have the plug-in estimators $\widehat{\Omega} = \widehat{\sigma}^2\widehat{Q}^{-1}$ for homoskedasticity, where $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/(n-K)$ or $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/n$, and $\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$ for heteroskedasticity, where $\widehat{\Sigma} = n^{-1}\sum_i x_i x_i'\widehat{e}_i^2$.

Remark 6.4. If we choose $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/(n-K)$ and replace σ^2 in (6.6), then the resulting statistic $T_k = \frac{\sqrt{n}(\widehat{\beta}_k - \beta_k)}{\sqrt{\widehat{\sigma}^2[\widehat{Q}^{-1}]_{kk}}}$ is exactly the t-statistic in the finite sample analysis. Recall that under the classical normal-error assumption, the t-statistics follows exact finite sample t-distribution with degrees of freedom $n-K$. In asymptotic analysis, we allow e_i to be any distribution if $E[e_i^2|x_i] < \infty$ (We impose this assumption for simplicity. It can be further relaxed in inid cases.) The asymptotic normality allows us to conduct asymptotic statistical inference. For the same t-statistic, we must draw the critical values from the normal distribution, because

$$T_k = \frac{\sqrt{\sigma^2[Q^{-1}]_{kk}}}{\sqrt{\widehat{\sigma}^2[\widehat{Q}^{-1}]_{kk}}} \cdot \frac{\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right)}{\sqrt{\sigma^2[Q^{-1}]_{kk}}} \xrightarrow{d} 1 \times N(0,1) \sim N(0,1)$$

by Slutsky's theorem if $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

The next section will give sufficient conditions for $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $\widehat{\Sigma} \xrightarrow{p} \Sigma$.

6.4 Consistency of Feasible Variance Estimator

We first show under what conditions all elements of $\Sigma = E\left[x_i x_i' e_i^2\right]$ are finite. That is, $\|\Sigma\|_\infty < \infty$, where $\|\cdot\|_\infty$ is the value of the largest element in absolute value of a matrix or vector. Let $z_i = x_i e_i$, so $\Sigma = E\left[z_i z_i'\right]$.

Definition 6.2 (norm and inner product of random variables)**.** For a generic random variable u_i with finite variance, define its L_2*-norm* as $\sqrt{E\left[u_i^2\right]}$. Given another generic random variable v_i with finite variance, define the *inner product* of u_i and v_i as $E\left[u_i v_i\right]$.

Fact 6.1 (Cauchy-Schwarz inequality for random variables)**.** $|E\left[u_i v_i\right]| \leq \sqrt{E\left[u_i^2\right] E\left[v_i^2\right]}$.

Because of the Cauchy-Schwarz inequality (cross moments are no larger than variance)

$$\|\Sigma\|_\infty = \max_{k \in [K]} E\left[z_{ik}^2\right],$$

where $[K] := \{1, 2, \ldots, K\}$. For each k,

$$E\left[z_{ik}^2\right] = E\left[x_{ik}^2 e_i^2\right] \leq \left(E\left[x_{ik}^4\right] E\left[e_i^4\right]\right)^{1/2}$$

where the last inequality again follows by the Cauchy-Schwarz inequality. It implies that the *sufficient conditions* for finite variance are

$$\max_k E\left[x_{ik}^4\right] < \infty \text{ and } E\left[e_i^4\right] < \infty. \tag{6.8}$$

We will maintain these conditions in the following derivation.

6.4.1 Homoskedasticity

For the estimation of variance, if the error is homoskedastic,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2 &= \frac{1}{n}\sum_{i=1}^n \left(e_i + x_i'\left(\widehat{\beta}-\beta\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n e_i^2 + \left(\frac{2}{n}\sum_{i=1}^n e_i x_i\right)'\left(\widehat{\beta}-\beta\right) + \frac{1}{n}\sum_{i=1}^n \left(\widehat{\beta}-\beta\right)' x_i x_i'\left(\widehat{\beta}-\beta\right).\end{aligned} \tag{6.9}$$

Definition 6.3 (Norm and inner product for vectors)**.** For a generic m-vector u, define its *L_2-norm* as $\|u\|_2 = \sqrt{u'u}$. Given another generic m-vector v, define the *inner product* of u and v as $\langle u, v\rangle = u'v$.

Fact 6.2 (Cauchy-Schwarz inequality for vectors)**.** $|\langle u, v\rangle| \leq \|u\|_2 \|v\|_2$*, or equivalently* $|u'v| \leq \sqrt{(u'u)(v'v)}$.

Notice $\frac{1}{n}\sum_{i=1}^{n} e_i x_i \xrightarrow{p} E[e_i x_i] = 0$, the second term of (6.9) is

$$\begin{aligned}\left|\left(\frac{2}{n}\sum_{i=1}^{n} e_i x_i\right)'\left(\widehat{\beta}-\beta\right)\right| &\leq 2\left\|\frac{1}{n}\sum_{i=1}^{n} x_i e_i\right\|_2 \left\|\widehat{\beta}-\beta\right\|_2 \\ &= o_p(1)\, o_p(1) = o_p(1) \end{aligned} \tag{6.10}$$

by Cauchy-Schwarz inequality.

Fact 6.3 (Quadratic inequality)**.** *For a generic $m \times m$ symmetric positive semi-definite matrix A and a generic m vector u, we have the*

$$\|u\|_2^2 \lambda_{\min}(A) \leq u'Au \leq \|u\|_2^2 \lambda_{\max}(A).$$

The third term of (6.9) is bounded by

$$\begin{aligned}\left(\widehat{\beta}-\beta\right)\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'\right)\left(\widehat{\beta}-\beta\right) &\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) \\ &\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \operatorname{trace}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) \\ &\leq \left\|\widehat{\beta}-\beta\right\|_2^2 K \max_k\left\{\frac{1}{n}\sum_{i=1}^{n} x_{ik}^2\right\} \\ &= o_p(1)\, O_p(1) = o_p(1), \end{aligned} \tag{6.11}$$

where the stochastic order follows by

$$\frac{1}{n}\sum_{i=1}^{n} x_{ik}^2 \xrightarrow{p} E\left[x_{ik}^2\right] < \infty$$

in view of the condition (6.8).

(6.10) and (6.11) implies that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}e_i^2 + o_p(1) + o_p(1) = \frac{1}{n}\sum_{i=1}^{n}e_i^2 + o_p(1) \xrightarrow{p} \sigma_e^2.$$

(See Appendix for the operations of small op and big Op.)

6.4.2 Heteroskedasticity

The basic strategy of proof is similar for the general case of heteroskedasticity, though each step is more complicated.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\widehat{e}_i^2 &= \frac{1}{n}\sum_{i=1}^{n}x_i x_i'\left(e_i + x_i'\left(\widehat{\beta}-\beta\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i x_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n}x_i x_i'\cdot e_i x_i'\left(\widehat{\beta}-\beta\right) + \frac{1}{n}\sum_{i=1}^{n}x_i x_i'\left(\left(\widehat{\beta}-\beta\right)' x_i\right)^2.\end{aligned} \tag{6.12}$$

Definition 6.4 (L_p-norm for vectors)**.** For a generic m-vector u, its L_p-norm (for $p \geq 1$) is defined as $\|u\|_p = \left(|u_1|^p + \cdots + |u_m|^p\right)^{1/p}$.

Fact 6.4 (Holder's inequality)**.** *For two generic m-vectors u and v,*

$$\left|u'v\right| \leq \|u\|_p \|q\|_q$$

for any $p, q \in [1, \infty)$ *and* $1/p + 1/q = 1$.

Remark 6.5. Cauchy-Schwarz inequality is a special case of Holder's in-

equality when $p = q = 2$.

The second term of (6.12) is bounded by

$$\begin{aligned}
&\max_{k,k'} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ik} x_{ik'} \cdot e_i x_i' \left(\widehat{\beta} - \beta \right) \right| \\
&\leq \left\| \widehat{\beta} - \beta \right\|_2 \max_{k,k'} \left\| \frac{1}{n} \sum_{i=1}^{n} x_i e_i x_{ik} x_{ik'} \right\|_2 \\
&\leq \left\| \widehat{\beta} - \beta \right\|_2 \sqrt{K} \max_{k,k',k''} \left| \frac{1}{n} \sum_{i=1}^{n} e_i x_{ik} x_{ik'} x_{ik''} \right| \\
&\leq \left\| \widehat{\beta} - \beta \right\|_2 \sqrt{K} \left(\frac{1}{n} \sum_{i=1}^{n} e_i^4 \right)^{1/4} \max_{k,k',k''} \left(\frac{1}{n} \sum_{i=1}^{n} \left(x_{ik} x_{ik'} x_{ik''} \right)^{4/3} \right)^{3/4} \\
&\leq \left\| \widehat{\beta} - \beta \right\|_2 \sqrt{K} \left(\frac{1}{n} \sum_{i=1}^{n} e_i^4 \right)^{1/4} \max_{k} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ik}^4 \right)^{3/4} \\
&= o_p(1) O_p(1) O_p(1) = o_p(1)
\end{aligned}$$

where the third inequality hold by the Holder's inequality with $p = 4$ and $q = 4/3$, and the stochastic order is guaranteed if under suitable conditions

$$\frac{1}{n} \sum_{i=1}^{n} e_i^4 \xrightarrow{p} E\left[e_i^4 \right] < \infty \text{ and } \frac{1}{n} \sum_{i=1}^{n} x_{ik}^4 \xrightarrow{p} E\left[x_{ik}^4 \right] < \infty.$$

The third term of (6.12) is bounded by

$$\begin{aligned}
&\max_{k_1,k_2}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ik_1}x_{ik_2}\left(\left(\widehat{\beta}-\beta\right)'x_i\right)^2\right| \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \max_{k_1,k_2}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ik_1}x_{ik_2}\left(x_i x_i'\right)\right| \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \max_{k_1,k_2,k_3,k_4}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ik_1}x_{ik_2}x_{ik_3}x_{ik_4}\right| \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \max_{k_1,k_2}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ik_1}^2x_{ik_2}^2\right| \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2^2 \max_{k}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ik}^4\right| \\
&= o_p(1)\,O_p(1) = o_p(1).
\end{aligned}$$

where the third and the fourth inequalities follow by applying Cauchy Schwarz inequality.

6.5 Summary

One of the most important techniques in asymptotic theory is manipulating inequalities. These derivations of the variances look complicated at first glance, but is often encountered in proofs of theoretical results. After many years of torment, you will be accustomed to these routine calculations.

Historical notes: White (1980) drew attention of economic contexts

that violate the classical statistical assumptions in linear regressions. It seeded econometricians' care, or obsession, in variance estimation for statistical inference. The following decades has witnessed a plethora of proposals of variance estimation that deal with various deviation from the classical assumptions.

Further reading: In this chapter all vectors are of finite dimensional. Some results can be extended to allow infinite K when $K \to \infty$ at a much slower speed than n. Such asymptotic development will require multiple indices, and it goes beyond the simplest case of $n \to \infty$ that we learned here. Big data is accompanied by big model, in which the model itself is indexed by the sample size and can grow more sophisticated as n get bigger. In the proofs of my latest paper Shi et al. (2020), You will find loads of inequality operations of similar flavor to this chapter.

6.6 Appendix

We introduce the "big Op and small op" notation. They are the stochastic counterparts of the "big O and small o" notation in deterministic cases.

- Small op: $x_n = o_p(r_n)$ if $x_n / r_n \xrightarrow{p} 0$.
- Big Op: $x_n = O_p(r_n)$ if for any $\varepsilon > 0$, there exists a $c > 0$ such that $P(|x_n| / r_n > c) < \varepsilon$.

Some operations:

- $o_p(1) + o_p(1) = o_p(1)$;
- $o_p(1) + O_p(1) = O_p(1)$;
- $o_p(1) O_p(1) = o_p(1)$.

The big Op and small op notation allows us to keep using equalities in calculation while expressing the stochastic order of random objects.

Zhentao Shi. Oct 21, 2020.

Bibliography

Shi, Z., L. Su, and T. Xie (2020). High dimensional forecast combinations under latent structures. *arXiv 2010.09477*. 6.5

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817–838. 6.5