ECON 6200: Econometrics II

The Linear Model: Large Sample Analysis

We next change perspective and look at large sample (asymptotic) theory.

Results will be weaker in that they only hold approximately, in senses formalized by stochastic convergence notions.

Assumptions will be considerably weaker as well. We begin with:

1. (X, Y) are i.i.d.
2. $\mathbb{E}Y^2 < \infty$, $\mathbb{E}\|X\|^2 < \infty$.
3. $\mathbb{E}(XX')$ is positive definite.

The Linear Model: Large Sample Analysis

These assumptions suffice for the projection coefficient

$$b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$$

to be well-defined.

These assumptions suffice for the projection coefficient

$$b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$$

to be well-defined.

Indeed, we have:

Theorem

Under the above assumptions, we have:

$$\hat{\beta} \equiv (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \xrightarrow{p} b^*.$$

The Linear Model: Large Sample Analysis

These assumptions suffice for the projection coefficient

$$b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$$

to be well-defined.

Indeed, we have:

Theorem

Under the above assumptions, we have:

$$\hat{\beta} \equiv (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \xrightarrow{p} b^*.$$

Proof

- $\mathbb{E}_n XX' \xrightarrow{p} \mathbb{E}XX'$ and $\mathbb{E}_n XY \xrightarrow{p} \mathbb{E}XY$ by WLLN.
- By CMT and assumptions, it follows that $\mathbb{E}_n XX'$ is nonsingular with probability approaching 1 and $(\mathbb{E}_n XX')^{-1} \xrightarrow{P} (\mathbb{E}XX')^{-1}$.
- The claim then follows from Slutsky's Theorem.

Restating the Result for the Structural Linear Model

Up to here, we did not use any assumptions that set the structural linear model apart from the BLP interpretation.

Indeed, the asymptotics apply to both!

However, it is useful to clarify exact assumption on ε needed if the linear model $Y = X'\beta + \varepsilon$ is maintained:

$$\mathbb{E}X\varepsilon = 0 \qquad (\text{"predetermination"}).$$

- The assumption is considerably weaker than what we used for unbiasedness: We only require uncorrelatedness of ε and contemporaneous regressors.
- This ensures that $b^* = \beta$ (and therefore consistency of OLS for β): $b^* = (\mathbb{E}(XX'))^{-1}\mathbb{E}XY = (\mathbb{E}(XX'))^{-1}\mathbb{E}X(X'\beta + \varepsilon) = \beta + (\mathbb{E}(XX'))^{-1}\mathbb{E}X\varepsilon = \beta$.
- Alternatively, can show consistency "from scratch" using this assumption (cf. Hayashi).

The Linear Model: Asymptotic Distribution

To look at the asymptotic distribution, write

$$\begin{aligned} \hat{\beta} &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY = (\mathbb{E}_n XX')^{-1}\mathbb{E}_n X(Xb^* + e) \\ \implies \hat{\beta} - b^* &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n Xe, \end{aligned}$$

where b^* is the population projection and e is the projection error.

The Linear Model: Asymptotic Distribution

To look at the asymptotic distribution, write

$$\begin{aligned} \hat{\beta} &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY = (\mathbb{E}_n XX')^{-1}\mathbb{E}_n X(Xb^* + e) \\ \implies \hat{\beta} - b^* &= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n Xe, \end{aligned}$$

where b^* is the population projection and e is the projection error.

- This provides another perspective on consistency:
  By uncorrelatedness of (X, e), $\mathbb{E}_n Xe \xrightarrow{p} 0$.
- But it also suggests an asymptotic result. It is compelling that
  $\sqrt{n}(\mathbb{E}_n Xe) \xrightarrow{d} N\big(0, \Omega\big) = N\big(0, \mathbb{E}(XX'e^2)\big)$,
  where the last equality just defines Ω.
- And if that were the case, it would somewhat easily yield
  $\sqrt{n}(\hat{\beta} - b^*) \xrightarrow{d} N\big(0, (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1}\big)$.
- This derivation is "morally true."
  We just need to ensure that all terms exist.

The Linear Model: Asymptotic Distribution

We slightly strengthen assumptions to:

1. (X, Y) are i.i.d.
2. $\mathbb{E}Y^4 < \infty$, $\mathbb{E}\|X\|^4 < \infty$.
3. $\mathbb{E}(XX')$ is positive definite.

The Linear Model: Asymptotic Distribution

We slightly strengthen assumptions to:

1. (X, Y) are i.i.d.
2. $\mathbb{E}Y^4 < \infty$, $\mathbb{E}\|X\|^4 < \infty$.
3. $\mathbb{E}(XX')$ is positive definite.

We can then argue that Ω is finite. For any one of its components:

$$\begin{aligned}|\mathbb{E}(X_k X_\ell e^2)| &\leq \mathbb{E}|X_k X_\ell e^2| = \mathbb{E}\big(|X_k||X_\ell||e^2|\big) \\ &\leq \left(\mathbb{E}X_k^2 X_\ell^2\right)^{1/2}\left(\mathbb{E}e^4\right)^{1/2} \leq \left(\mathbb{E}X_k^4\right)^{1/4}\left(\mathbb{E}X_\ell^4\right)^{1/4}\left(\mathbb{E}e^4\right)^{1/2} < \infty,\end{aligned}$$

where we repeatedly used Cauchy-Schwarz.

The Linear Model: Asymptotic Distribution

We slightly strengthen assumptions to:

1. (X, Y) are i.i.d.
2. $\mathbb{E}Y^4 < \infty$, $\mathbb{E}\|X\|^4 < \infty$.
3. $\mathbb{E}(XX')$ is positive definite.

We can then argue that Ω is finite. For any one of its components:

$$\begin{aligned}|\mathbb{E}(X_k X_\ell e^2)| &\leq \mathbb{E}|X_k X_\ell e^2| = \mathbb{E}\big(|X_k||X_\ell||e^2|\big) \\ &\leq \big(\mathbb{E}X_k^2 X_\ell^2\big)^{1/2}\big(\mathbb{E}e^4\big)^{1/2} \leq \big(\mathbb{E}X_k^4\big)^{1/4}\big(\mathbb{E}X_\ell^4\big)^{1/4}\big(\mathbb{E}e^4\big)^{1/2} < \infty,\end{aligned}$$

where we repeatedly used Cauchy-Schwarz. We wrap up by writing

$$\sqrt{n}(\hat{\beta} - b^*) \xrightarrow{d} (\mathbb{E}XX')^{-1}N\big(0, \Omega\big) = N\big(0, (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1}\big),$$

where we used Slutsky and the fact that Normality passes through linear mappings.

The Linear Model: Asymptotic Distribution

Theorem

Assume:

1. (X, Y) are i.i.d.
2. $\mathbb{E}Y^4 < \infty$, $\mathbb{E}\|X\|^4 < \infty$.
3. $\mathbb{E}(XX')$ is positive definite.

Then

$$\begin{aligned} \sqrt{n}(\hat{\beta} - b^*) &\overset{d}{\to} N\big(0, \mathrm{avar}(\hat{\beta})\big), \\ \mathrm{avar}(\hat{\beta}) &= (\mathbb{E}XX')^{-1}\Omega(\mathbb{E}XX')^{-1} \\ [&= \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\Omega\boldsymbol{Q}_{\boldsymbol{XX}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}\Omega\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}] \end{aligned}$$

Here, avar stands for asymptotic variance, i.e., the variance of the limiting distribution. In general, this need not be the asymptotic limit of an estimator's (scaled) variance, although under current assumptions it is.

The Linear Model: Asymptotic Distribution

The theorem provides the **joint** asymptotic distribution of estimates.

The information contained in joint normality is relevant for:

- Inference on linear combinations of coefficients, e.g. their sum or difference. (This can also be achieved by reparameterization, but that can be impractical.)
- Joint inference, e.g. confidence ellipsoids, on several coefficients.
- Inference on a known, differentiable function of β through the Delta method. (Omitted because conceptually straightforward, but important in practice! See the textbook for a worked-out example.)
- Conservative inference on a known, nondifferentiable function of β through projection (i.e., operate the function on every b in the confidence ellipsoid). In structured cases, one may be able to improve on this one. Ask me if the question arises in your work.

Reminder: Pointwise vs Uniform Asymptotics

Results in this course hold pointwise as $n \to \infty$ for given parameter values, not (even as $n \to \infty$) uniformly over parameter values.

How big of a problem is this?

With additional effort, most results from this course are available uniformly over "nice" cases.

However, some "less nice" cases are empirically relevant:

- Estimators that can be corner solutions in the problem defining them,
- Estimation of maxima.
- Rare events.
- Post-model selection estimation and inference.

In all of these examples, a pointwise perspective can be seriously misleading.

Keep in mind that this course develops applicable results for "nice" cases (and hopefully conceptual insight beyond)...

Estimating the Asymptotic Variance

Theorem:

Under our current assumption,

$$\widehat{\mathsf{avar}}_{HC0} \equiv (\mathbb{E}_n XX')^{-1}\hat{\Omega}(\mathbb{E}_n XX')^{-1} \xrightarrow{p} \mathsf{avar}(\hat{\beta})$$

and similarly for $HC1$ etc. That is, all of these are consistent.

Estimating the Asymptotic Variance

Theorem:

Under our current assumption,

$$\widehat{\text{avar}}_{HC0} \equiv (\mathbb{E}_n XX')^{-1}\hat{\Omega}(\mathbb{E}_n XX')^{-1} \xrightarrow{p} \text{avar}(\hat{\beta})$$

and similarly for $HC1$ etc. That is, all of these are consistent.

Proof Sketch

The bottleneck is consistency of $\hat{\Omega}$. Write

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2}_{\xrightarrow{p}\Omega} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i' (\hat{e}_i^2 - e_i^2)}_{\xrightarrow{p} 0},$$

The r.h. claim can be tediously shown through repeated use of the Cauchy-Schwarz and Hölder inequalities.

Simplification under Homoskedasticity

If we assume that $\mathbb{E}(e^2 \mid X) = \sigma^2$, then we have simplification

$$\text{avar}(\hat{\beta}) = (\mathbb{E}XX')^{-1}\sigma^2,$$

and showing consistency of

$$\widehat{\text{avar}}(\hat{\beta}) \equiv (\mathbb{E}_n XX')^{-1} s^2$$

is not hard and also technically requires weaker assumptions (namely, 2^{nd} moments of X and Y suffice).

Recall that this assumption makes sense only for a structural linear model (and then still is restrictive).

The Linear Model: Inference

Let $r(\cdot) : \mathbb{R}^k \to \mathbb{R}$ a continuously differentiable scalar-valued function with gradient $\nabla r(\cdot) = \boldsymbol{R}(\cdot)$.

The easiest example is that $r(\cdot)$ extracts a component of β.

To keep algebraic expressions tight, define $\theta = r(\beta)$ and $\hat{\theta} = r(\hat{\beta})$.

By the Delta method, standard convergence results, and Slutsky, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\overset{d}{\to} N\big(0, \text{avar}(\hat{\theta})\big) \\
\text{avar}(\hat{\theta}) &= \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\widehat{\text{avar}}(\hat{\theta}) \equiv \boldsymbol{R}(\hat{\beta})\widehat{\text{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})' &\overset{p}{\to} \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\implies t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \equiv \frac{\sqrt{n}(\hat{\theta} - \theta)}{\big(\boldsymbol{R}(\hat{\beta})\widehat{\text{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})'\big)^{1/2}} &\overset{d}{\to} N(0, 1).
\end{aligned}$$

This is the asymptotic t-statistic.

For the derivation to be true, we need that $\text{avar}(\hat{\theta})$ is finite.

A sufficient condition is that $\boldsymbol{R}(\beta) \neq 0$ and $\text{avar}(\hat{\beta})$ has full rank.

The Linear Model: Inference

Let $r(\cdot) : \mathbb{R}^k \to \mathbb{R}$ a continuously differentiable scalar-valued function with gradient $\nabla r(\cdot) = \boldsymbol{R}(\cdot)$.

The easiest example is that $r(\cdot)$ extracts a component of β.

To keep algebraic expressions tight, define $\theta = r(\beta)$ and $\hat{\theta} = r(\hat{\beta})$.

By the Delta method, standard convergence results, and Slutsky, we have

$$
\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} N\big(0, \mathrm{avar}(\hat{\theta})\big) \\
\mathrm{avar}(\hat{\theta}) &= \boldsymbol{R}(\beta)\,\mathrm{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\widehat{\mathrm{avar}}(\hat{\theta}) \equiv \boldsymbol{R}(\hat{\beta})\widehat{\mathrm{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})' &\xrightarrow{p} \boldsymbol{R}(\beta)\,\mathrm{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\Longrightarrow \; t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \equiv \frac{\sqrt{n}(\hat{\theta} - \theta)}{\big(\boldsymbol{R}(\hat{\beta})\widehat{\mathrm{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})'\big)^{1/2}} &\xrightarrow{d} N(0, 1).
\end{aligned}
$$

Dividing by the standard error is called studentization.

It ensures that the statistic's asymptotic distribution does not depend on unknown parameters.

Any statistic with this property is called an asymptotic pivot.

The Linear Model: Inference

The previous result yields hypothesis tests and confidence intervals with asymptotic size respectively coverage.

Let $\Phi(\cdot)$ denote the standard normal c.d.f. and define the quantiles (criticsl values) $c_\alpha \equiv \Phi^{-1}(1-\alpha)$, then (under currently maintained assumptions):

$$\begin{aligned}
\Pr(|t(\theta)| \leq c_{\alpha/2}) &\overset{p}{\to} 1-\alpha \\
\Pr(\theta \in CI_\alpha(\theta)) &\overset{p}{\to} 1-\alpha \\
CI_\alpha(\theta) &\equiv \left[\hat{\theta} - c_{\alpha/2} \cdot SE(\hat{\theta}), \hat{\theta} + c_{\alpha/2} \cdot SE(\hat{\theta})\right].
\end{aligned}$$

One-sided confidence intervals are constructed similarly.

The Linear Model: Inference

If $r(\beta) = p'\beta$ for known vector p, then $\boldsymbol{R}(\cdot) = p$ and the t-statistic simplifies to

$$t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(p' \widehat{\text{avar}}(\hat{\beta}) p\right)^{1/2}}.$$

The Linear Model: Inference

If $r(\beta) = p'\beta$ for known vector p, then $\boldsymbol{R}(\cdot) = p$ and the t-statistic simplifies to

$$t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(p' \widehat{\text{avar}}(\hat{\beta}) p\right)^{1/2}}.$$

- If p is a basis vector, this further simplifies to the t-statistic for an individual coefficient that you have surely seen before.
- Choices like $p = (0, 1, -1, 0, \ldots, 0)$ allow to test hypotheses like equality of two coefficients.
  (Whereas differences of confidence intervals are not confidence intervals for differences!)
- The latter can also be achieved by reparametrization, but that is more a teaching tool than practically important.

The Linear Model: Inference

If $r(\beta) = p'\beta$ for known vector p, then $\boldsymbol{R}(\cdot) = p$ and the t-statistic simplifies to

$$t(\theta) \equiv \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\left(p' \widehat{\text{avar}}(\hat{\beta}) p\right)^{1/2}}.$$

- Under the causal linear model assumption, another interesting application is to $p = x$, in which case $\theta = E(Y \mid X = x)$.
  This is called a regression interval.
  Note that its standard error depends on x and will tend to be smaller for "central" values of x.
- The regression interval is not a forecast confidence interval!
  (A frequent mistake in reports on COVID-19 modelling!)
  For forecast intervals, i.e. forecasting Y_t from X_t, must take ε_t into account.
  Under a homoskedasticity assumption, this is achieved by adding s^2 to $\widehat{\text{avar}}(x'\hat{\beta})$.

Multiple Hypothesis Testing

We next generalize to $\theta = r(\beta)$, where $r(\cdot) : \mathbb{R}^k \to \mathbb{R}^q$ is vector valued.

Defining $\boldsymbol{R}(\cdot)$ as the **Jacobian** of r, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} N\big(0, \text{avar}(\hat{\theta})\big) \\
\text{avar}(\hat{\theta}) &= \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\widehat{\text{avar}}(\hat{\theta}) \equiv \boldsymbol{R}(\hat{\beta})\widehat{\text{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})' &\xrightarrow{p} \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\implies W(\theta) \equiv \sqrt{n}(\hat{\theta} - \theta)'(\widehat{\text{avar}}(\hat{\theta}))^{-1}\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} \chi^2_q.
\end{aligned}$$

This is the Wald statistic.

Multiple Hypothesis Testing

We next generalize to $\theta = r(\beta)$, where $r(\cdot) : \mathbb{R}^k \to \mathbb{R}^q$ is vector valued.

Defining $\boldsymbol{R}(\cdot)$ as the **Jacobian** of r, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} N\big(0, \mathrm{avar}(\hat{\theta})\big) \\
\mathrm{avar}(\hat{\theta}) &= \boldsymbol{R}(\beta)\,\mathrm{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\widehat{\mathrm{avar}}(\hat{\theta}) \equiv \boldsymbol{R}(\hat{\beta})\widehat{\mathrm{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})' &\xrightarrow{p} \boldsymbol{R}(\beta)\,\mathrm{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\implies W(\theta) \equiv \sqrt{n}(\hat{\theta} - \theta)'(\widehat{\mathrm{avar}}(\hat{\theta}))^{-1}\sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} \chi^2_q.
\end{aligned}$$

This is the Wald statistic.

- A sufficient condition for the derivation to go through is that both $\boldsymbol{R}(\beta)$ and $\mathrm{avar}(\hat{\beta})$ are of full rank.
- This means the hypotheses must be locally linearly independent at the truth.
- For linear hypotheses, this is a global property and easy to check.

The Linear Model: Inference

Multiple Hypothesis Testing

We next generalize to $\theta = r(\beta)$, where $r(\cdot) : \mathbb{R}^k \to \mathbb{R}^q$ is vector valued.

Defining $\boldsymbol{R}(\cdot)$ as the **Jacobian** of r, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &\stackrel{d}{\to} N\big(0, \text{avar}(\hat{\theta})\big) \\
\text{avar}(\hat{\theta}) &= \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\widehat{\text{avar}}(\hat{\theta}) \equiv \boldsymbol{R}(\hat{\beta})\widehat{\text{avar}}(\hat{\beta})\boldsymbol{R}(\hat{\beta})' &\stackrel{p}{\to} \boldsymbol{R}(\beta)\,\text{avar}(\hat{\beta})\boldsymbol{R}(\beta)' \\
\Longrightarrow \quad W(\theta) \equiv \sqrt{n}(\hat{\theta} - \theta)'(\widehat{\text{avar}}(\hat{\theta}))^{-1}\sqrt{n}(\hat{\theta} - \theta) &\stackrel{d}{\to} \chi^2_q.
\end{aligned}$$

This is the Wald statistic.

- This test statistic induces confidence regions in analogy to CI_α above.
- These are ellipsoids.
- They are appropriate if one is genuinely interested in simultaneous inference on several scalars, e.g. several components of β.
- Their projections onto the axes will be valid but (possibly very) conservative confidence intervals for components of θ.