Lecture 7
Asymptotics of OLS

1

OLS Estimation - Assumptions

- CLM Assumptions

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\beta + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $\text{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3**) $\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\,\mathbf{I}_T$

(**A4**) $\mathbf{X}$ has full column rank – rank($\mathbf{X}$)=k-, where T ≥ k.

- From (**A1**), (**A2**), and (**A4**) $\Rightarrow \mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'}\,\boldsymbol{y}$
- Using (**A3**) $\Rightarrow \text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2(\mathbf{X'X})^{-1}$
- Adding (**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim iid\,\text{N}(\mathbf{0}, \sigma^2\mathbf{I}_T)$ $\Rightarrow \mathbf{b}|\mathbf{X} \sim iid\,\text{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X'X})^{-1})$

(**A5**) gives us *finite sample* results for **b** (& for the *t-test*, *F-test*, Wald test)

- Now, we relax (**A5**). We study **b** (& the test statistics) when $T \rightarrow \infty$.

OLS Estimation - Assumptions

• In this lecture, we relax (**A5**). We focus on the behavior of **b** (and the test statistics) when $T \to \infty$ –i.e., *large samples*.

• First, we throw away the normality for $\boldsymbol{\varepsilon}|\mathbf{X}$. This is not bad. In many econometric situations, normality is not a realistic assumption (daily, weekly, or monthly stock returns do not follow a normal).

• Second, we relax the *i.i.d.* assumption for $\boldsymbol{\varepsilon}|\mathbf{X}$. This is also not bad. In many econometric situations, identical distributions are not realistic (different means and variances are common).

• Q: How does **b** (and all the tests) behave without this normality assumption? We will not be able to say much for small samples. But, we can say a lot about the behavior of **b** when $T \to \infty$.

Brief Review: Plims and Consistency

• The asymptotic properties of estimators are their properties as the number of observations in a sample becomes very large and tends to infinity.

• Q: Why are we interested in large sample properties, like consistency, when in practice we have finite samples?

A: As a first approximation, the answer is that if we can show that an estimator has good large sample properties, then we may be optimistic about its finite sample properties. For example, if an estimator is inconsistent, we know that for finite samples it will definitely be biased.

• We will review the concepts of probability limits, consistency, and the CLT.

Probability Limit: Convergence in probability

• Definition: Convergence in probability

Let θ be a constant, $\varepsilon > 0$, and n be the index of the sequence of RV x_n. If $\lim_{n\to\infty}$ Prob[$|x_n - \theta| > \varepsilon$] = 0 for any $\varepsilon > 0$, we say that x_n *converges in probability* to θ.

That is, the probability that the difference between x_n and θ is larger than any $\varepsilon > 0$ goes to zero as n becomes bigger.

Notation: $x_n \xrightarrow{p} \theta$

plim $x_n = \theta$

• If x_n is an estimator (for example, the sample mean) and if plim $x_n = \theta$, we say that x_n is a *consistent* estimator of θ.

Estimators can be *inconsistent*. For example, when they are consistent for something other than our parameter of interest.

Probability Limit: Weak Law of Large Numbers

• **Theorem**: Convergence for sample moments.

Under certain assumptions (for example, *i.i.d.* with *finite mean*), sample moments converge in probability to their population counterparts.

We saw this theorem before. It's the (Weak) Law of Large Numbers (LLN). Different assumptions create different versions of the LLN.

Note: The LLN is very general:

$$(1/n) \sum_i^n f(z_i) \xrightarrow{p} \text{E}[f(z_i)]$$

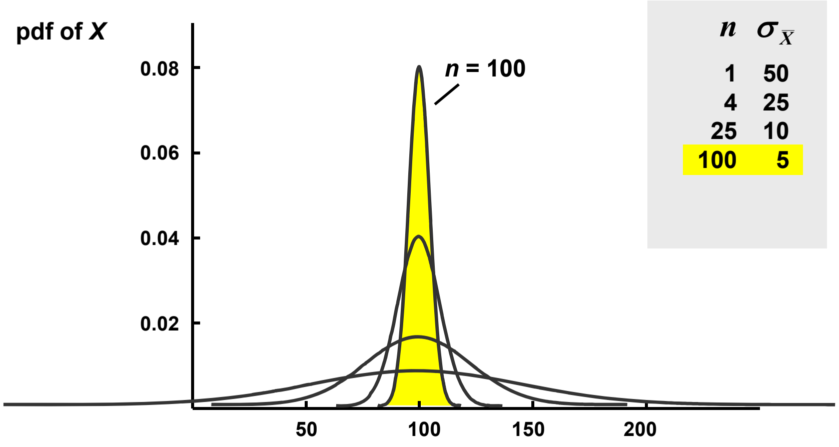
• The usual version in Greene assumes *i.i.d.* with finite mean. This is the Khinchin's (1929) (weak) LLN. (Khinchin is also spelled *Khintchine)*

Probability Limit: Weak Law of Large Numbers

• When $\{X_n\}$ is not *i.i.d.*, extra conditions are needed for the convergence of $(1/n)\sum_i^n f(x_i)$. Such conditions are typically imposed on higher-order moments of X_n.

• For the non-*i.i.d.* case, we have Chebychev's version, which assumes *independence* and *finite mean* and *finite variance*.

Plims and Consistency: Review



n	$\sigma_{\bar{X}}$
1	50
4	25
25	10
100	5

• **C**onsider the mean of a sample, $\bar{X}$, of observations generated from a RV X with mean μ_X and variance σ_X^2. Recall Var$[\bar{X}] = \sigma_X^2/n$. Then, as n grows, the sampling distribution becomes more concentrated.

Slutsky's Theorem: Review

Let x_n be a RV such that plim $x_n = \theta$. (We assume θ is a constant.) Let $g(.)$ be a continuous function with continuous derivatives. $g(.)$ is not a function of n. Then

$$\text{plim}[g(x_n)] = g[\text{plim}(x_n)] = g[\theta] \qquad \text{(provided } g[\text{plim}(x_n)] \text{ exists)}$$

When $g(.)$ is continuous, this result is sometimes referred as the *continuity theorem.*

Note 1: This theorem extends to sequences of random vectors and vector-valued X-continuous functions.

Note 2: This theorem is extremely useful and has many applications

Plims and Expectations: Review

- Q: What is the difference between E[x_n] and plim x_n?
  – E[x_n] reflects an average
  – plim x_n reflects a (probabilistic) limit of a sequence.

Slutsky's Theorem works for plims, but not for expectations. That is,

$$\text{plim}[s^2] = \sigma^2 \quad \Rightarrow \text{plim}[s = \sqrt{s^2}] = \sigma$$
$$E[s^2] = \sigma^2 \quad \Rightarrow E[s] = ?$$

Note: This very simple result is one of the motivations of using asymptotic theory. Plims are easy to manipulate, expectations are not. For example, the expectation of a product of RVs is complicated to derive, but the plim is not difficult.

Properties of plims: Review

- These properties are derived from Slutsky's Theorem.

Let x_n have plim $x_n = \theta$ and y_n have plim $y_n = \psi$. Let c be a constant. Then,

1) plim $c = c$.

2) plim $(x_n + y_n) = \theta + \psi$.

3) plim $(x_n * y_n) = \theta * \psi$. (plim $(c\, x_n) = c\, \theta$.)

4) plim $(x_n / y_n) = \theta / \psi$. (provided $\psi \neq 0$)

5) plim$[g(x_n, y_n)] = g(\theta, \psi)$. (assuming it exists and $g(.)$ is cont. diff.)

- We can generalize Slutsky's Theorem to matrices.

Let plim $\mathbf{A}_n = \mathbf{A}$ and plim $\mathbf{B}_n = \mathbf{B}$ (element by element). Then

1) plim$(\mathbf{A}_n^{-1})$ = [plim $\mathbf{A}_n]^{-1} = \mathbf{A}^{-1}$

2) plim$(\mathbf{A}_n\mathbf{B}_n)$ = plim$(\mathbf{A}_n)$ plim$(\mathbf{B}_n) = \mathbf{AB}$

Convergence in Mean(*r*): Review

- Definition: Convergence in mean *r*

Let θ be a constant, and n be the index of the sequence of RV x_n. If

$$\lim_{n\to\infty} \mathrm{E}[(x_n - \theta)^r] = 0 \text{ for any } r \geq 1,$$

we say that x_n *converges in mean r* to θ.

The most used version is mean-squared convergence, which sets $r = 2$.

Notation: $x_n \xrightarrow{p} \theta$

$x_n \xrightarrow{m.s.} \theta$ (when $r = 2$)

For the case $r = 2$, the sample mean converges to a constant, since its variance converges to zero.

Theorem: $x_n \xrightarrow{m.s.} \theta \quad \Rightarrow x_n \xrightarrow{p} \theta$

Consistency: Brief Remarks

• *Consistency*

A consistent estimator of a population characteristic satisfies two conditions:

(1) It possesses a probability limit –its distribution collapses to a spike as the sample size becomes large, and

(2) The spike is located at the true value of the population characteristic.

• The sample mean in our example satisfies both conditions and so it is a consistent estimator of μ_X. Most estimators, in practice, satisfy the first condition, because their variances tend to zero as the sample size becomes large.

• Then, the only issue is whether the distribution collapses to a spike at the true value of the population characteristic.

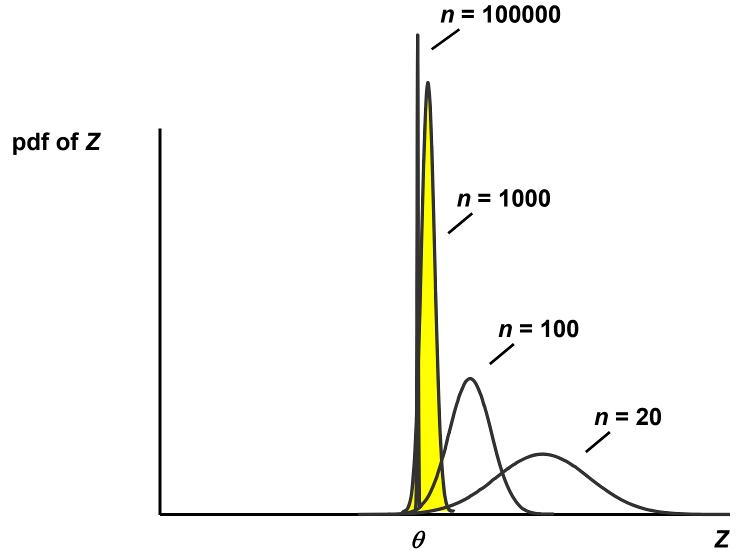
Consistency: Brief Remarks

- A *sufficient* condition for consistency is that the estimator should be unbiased and that its variance should tend to zero as *n* becomes large.

- However the condition is only sufficient, not *necessary*. It is possible that an estimator may be biased in a finite sample, but the bias disappears as the sample size tends to infinity.

⇒ Such an estimator is biased (in finite samples), but consistent because its distribution collapses to a spike at the true value.

Consistency: Brief Remarks



Consistency: Brief Remarks

• Therefore, we should be cautious about preferring consistent estimators to inconsistent ones.

(1) A consistent estimator may be biased for finite samples.

(2) If a consistent estimator has a larger variance than an inconsistent one, the latter might be preferable if judged by the MSE.

(3) How can you resolve these issues? Mathematically they are intractable, otherwise we would not have resorted to large sample analysis in the first place.

• A simulation can help to understand the trade-offs.

Almost Sure Convergence: Review

- Definition: Almost sure convergence

Let θ be a constant, and n be the index of the sequence of RV x_n. If

$$P[\ \lim_{n\to\infty} x_n = \theta\ ] = 1,$$

we say that x_n *converges almost surely* to θ.

The probability of observing a realization of $\{x_n\}$ that does not converge to θ is zero. $\{x_n\}$ may not converge everywhere to θ, but the points where it does not converge form a zero measure set (probability sense).

Notation: $x_n \xrightarrow{a.s.} \theta$

This is a stronger convergence than convergence in probability.

Theorem: $x_n \xrightarrow{a.s.} \theta \quad \Rightarrow\ x_n \xrightarrow{p} \theta$

Almost Sure Convergence: Strong LLN

- In almost sure convergence, the probability measure takes into account the joint distribution of $\{X_n\}$. With convergence in probability we only look at the joint distribution of the elements of $\{X_n\}$ that actually appear in x_n.

- Strong Law of Large Numbers

We can state the LLN in terms of almost sure convergence:

Under certain assumptions, sample moments converge almost surely to their population counterparts.

This is the Strong LLN.

- From the previous theorem, the Strong LLN implies the (Weak) LLN.

Almost Sure Convergence: Strong LLN

• Versions used in Greene

(i) Khinchine's Strong LLN.

Assumptions: $\{X_n\}$ is a sequence of *i.i.d.* RVs with E[X_n] = $\mu < \infty$.

(ii) Kolmogorov's Strong LLN.

Assumptions: $\{X_n\}$ is a sequence of *independent.* RVs with E[X_n] = $\mu < \infty$ and Var[X_n] = $\sigma^2 < \infty$.

Convergence for Random Functions: ULLN

• In econometrics, we often deal with sample means of random functions. A random function is a function that is a random variable for each fixed value of its argument.

• In cross section econometrics, random functions usually take the form of a function $g(\boldsymbol{Z}, \theta)$ of a random vector $\boldsymbol{Z}$ and a non-random vector θ.

• For example, consider a Poisson model:

$$\Pr ob(Y_i = y_i \mid X_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

• Let ln λ_i= $\mathbf{X}_i\boldsymbol{\beta}$ and denote $Z_j = (Y_j, \boldsymbol{X}_j)$. Then,

$g(\boldsymbol{Z}_i, \theta) = -\mathbf{X}_i\boldsymbol{\beta} + y_j \ln(\mathbf{X}_i\boldsymbol{\beta}) - \Sigma_i \ln(y_j)$, where $\theta = \boldsymbol{\beta}$.

For these functions we can extend the LLN to a Uniform LLN.

Convergence for Random Functions: ULLN

• **Theorem**: Uniform weak LLN (UWLLN)

Let $\{Z_i, i = 1, 2, .., n\}$ be a random sample from a k-variate distribution. Let $g(z, \theta)$ be a Borel measurable function on $\mathbf{Z} \times \Theta$, where $\mathbf{Z} \in R^k$ is a Borel set such that $P[Z_i \in \mathbf{Z}] = 1$, and Θ is a compact subset of R^m, such that for each $z \in \mathbf{Z}$, $g(z, \theta)$ is a continuous function on Θ. Furthermore, let

$$E[\sup_{\theta \in \Theta} \mid g(Z_i, \theta) \mid] < \infty$$

Then,

$$\text{plim} \sup_{\theta \in \Theta} \mid (1/n) \textstyle\sum_i^n g(Z_i, \theta) - E[g(Z, \theta)] \mid = 0.$$

• That is, for any fixed θ, the sequence $\{g(Z_1, \theta), g(Z_2, \theta), \ldots\}$ is a sequence of *i.i.d.* RVs, and the sample mean of this sequence converges in probability to $E[g(Z, \theta)]$. This is *pointwise* (in θ) convergence.

Note: The condition that the random vectors Z_i are *i.i.d.* can be relaxed.

Back to CLM: New Assumptions

(1) $\{x_i, \varepsilon_i\}$ $i = 1, 2, ..., T$ is a sequence of independent observations.

– **X** is stochastic, but independent of the process generating $\boldsymbol{\varepsilon}$.

– We require that **X** have finite means and variances. Similar requirement for $\boldsymbol{\varepsilon}$, but we also require $E[\boldsymbol{\varepsilon}]=\mathbf{0}$.

(2) Well behaved **X**:

$$\text{plim } (\mathbf{X}'\mathbf{X}/T) = \mathbf{Q} \qquad (\mathbf{Q} \text{ a pd matrix of finite elements})$$

- Q: Why do we need assumption (2) in terms of a ratio divided by T? Each element of $\mathbf{X}'\mathbf{X}$ matrix is a sum of T numbers. As $T \to \infty$, these sums will become large. We divide by T so that the sums will not be too large.

Linear Model: New Assumptions

(2) plim ($\mathbf{X}'\mathbf{X}/T$) = $\mathbf{Q}$ ($\mathbf{Q}$ a pd matrix of finite elements)

Note: This assumption is not a difficult one to make since the LLN suggests that the each component of $\mathbf{X}'\mathbf{X}/T$ goes to the mean values of $\mathbf{X}'\mathbf{X}$. We require that these values are finite.

– Implicitly, we assume that there is not too much dependence in **X**.

Linear Model: New Assumptions

- Now, we have a new set of assumptions in the CLM:

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\beta + \boldsymbol{\varepsilon}$.

(**A2'**) **X** stochastic, but E[$\mathbf{X}'\,\boldsymbol{\varepsilon}$] = 0 and E[$\boldsymbol{\varepsilon}$] = **0.**

(**A3**) Var[$\boldsymbol{\varepsilon}|\mathbf{X}$] = $\sigma^2\,\mathbf{I}_T$

(**A4'**) plim ($\mathbf{X}'\mathbf{X}/\,T$) = $\mathbf{Q}$ (pd matrix with finite elements, rank = k)

- We want to study the large sample properties of OLS:

Q 1: Is **b** consistent? s^2?

Q 2: What is the distribution of **b**?

Q 3: What about the distribution of the tests: *t-tests, F-tests* & Wald tests?

Consistency of OLS: b

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

$\Rightarrow$ plim $\mathbf{b} =$ plim $\boldsymbol{\beta}$ + plim $(\mathbf{X'X}/T)^{-1}$ plim $(\mathbf{X'}\boldsymbol{\varepsilon}/T)$

$= \quad \boldsymbol{\beta} + \mathbf{Q}^{-1}$ plim $(\mathbf{X'}\boldsymbol{\varepsilon}/T)$

• When can we say that plim $(\mathbf{X'}\boldsymbol{\varepsilon}/T) = 0$?

New assumption (1) -or (**A2'**)- $\Rightarrow$ E$[\mathbf{X'}\boldsymbol{\varepsilon}]=0$

Then, using new assumptions (1) and (2), we can use the (weak) LLN:

$\Rightarrow$ plim $(\mathbf{X'}\boldsymbol{\varepsilon}/T) = \mathbf{0}$

$\Rightarrow$ plim $\mathbf{b} = \boldsymbol{\beta}$ $\Rightarrow$ $\mathbf{b}$ is consistent.

Note: This could have been shown through $(\mathbf{X'}\boldsymbol{\varepsilon}/T) \xrightarrow{m.s.} \mathbf{0}$.

Consistency of OLS: s^2

• $s^2 = \mathbf{e'e}/(T-k)$

$\Rightarrow$ plim s^2 = plim$[\mathbf{e'e}/(T-k)]$ = plim$[\mathbf{e'e}/T]$ * plim $[T/(T-k)]$

= plim$[\mathbf{e'e}/T]$

= plim $[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/T]$

= plim $[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/T]$ – plim $[\boldsymbol{\varepsilon}'\boldsymbol{X}\,(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\boldsymbol{\varepsilon}/T]$

= plim $[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/T]$ – plim$(\boldsymbol{\varepsilon}'\boldsymbol{X}/T)$ * plim$(\boldsymbol{X'X}/T)^{-1}$ *

* plim$(\boldsymbol{X'}\boldsymbol{\varepsilon}/T)$

= plim $[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/T]$ – $\mathbf{0}$ * $\mathbf{Q}^{-1}$ * $\mathbf{0}$ = σ^2

$\Rightarrow \sigma^2$ (s^2 is consistent)

Note: Using Slutzky's theorem, we can show that plim $s = \sigma$. Now, recall that we cannot use Slutzky's theorem for expectations when *g*(.) is nonlinear! That is, s is not an unbiased estimator for σ.

Convergence to a Random Variable: Review

• Definition: Limiting Distribution

Let x_n be a random sequence with cdf $F_n(x_n)$. Let x be a random variable with cdf $F(x)$.

When F_n converges to F as $n \to \infty$, for all points x at which $F(x)$ is continuous, we say that x_n converges in distribution to x. The distribution of that random variable is the *limiting distribution* of x_n.

Notation: $x_n \xrightarrow{d} x$

Example: The t_n statistic converges to a N(0, 1): $t_n \xrightarrow{d}$ N(0, 1)

Remark: If plim $x_n = \theta$ (a constant), then $F_n(x_n)$ becomes a point.

Convergence to a Random Variable: Review

Theorem: If $x_n \xrightarrow{d} x$ & plim $y_n = c$. Then, $x_n\, y_n \xrightarrow{d} c\, x$.
That is the limiting distribution of $x_n\ y_n$ is the distribution of $c\, x$.

Also, $x_n + y_n \xrightarrow{d} x + c$

$x_n / y_n \xrightarrow{d} x/c$ (provided $c \neq 0$.)

Slutsky's Theorem for RVs - Review

Let x_n converge in distribution to x and let $g(.)$ be a *continuous* function with continuous derivatives. $g(.)$ is not a function of n.

Then, $g(x_n) \xrightarrow{d} g(x)$.

Example: $t_n \xrightarrow{d}$ N(0,1) $\Rightarrow g(t_n) = (t_n)^2 \xrightarrow{d} [\text{N}(0,1)]^2$.

- Extension

Let $x_n \xrightarrow{d} x$ & $g(x_n, \theta) \xrightarrow{d} g(x)$ (θ: parameter).
Let plim $y_n = \theta$ (y_n is a consistent estimator of θ)

Then, $g(x_n, y_n) \xrightarrow{d} g(x)$.

That is, replacing θ by a consistent estimator leads to the same limiting distribution.

Extension of Slutsky's Theorem: Examples

Example 1: t_n statistic

$$z = \sqrt{n}\,(\bar{x} - \mu)/\sigma \xrightarrow{d} \text{N}(0, 1)$$

$$t_n = \sqrt{n}\,(\bar{x} - \mu)/s_n \xrightarrow{d} \text{N}(0, 1) \qquad (\text{where plim } s_n = \sigma)$$

Example 2: F-statistic for testing J restrictions in a regression ($\boldsymbol{e*}$ & $\boldsymbol{e}$ are restricted and unrestricted residuals, respectively)

$$F = [(\boldsymbol{e*}'\boldsymbol{e*} - \boldsymbol{e}'\boldsymbol{e})/\text{J}]/[\boldsymbol{e}'\boldsymbol{e}/(T-k)]$$
$$= [(\boldsymbol{e*}'\boldsymbol{e*} - \boldsymbol{e}'\boldsymbol{e})/(\sigma^2\text{J})]/[\boldsymbol{e}'\boldsymbol{e}/(\sigma^2(T-k))]$$

The denominator: $\boldsymbol{e}'\boldsymbol{e}/[\sigma^2(T-k)] \xrightarrow{p} 1$.

Then, the limiting distribution of the F statistic will be given by the limiting distribution of the numerator.

The CLT: Review

• The CLT states conditions for the sequence of RV $\{x_n\}$ under which the mean or a sum of a sufficiently large number of x_i *'s* will be approximately normally distributed.

CLT: Under some conditions, $z = \sqrt{n}\,(\bar{x} - \mu)/\sigma \xrightarrow{d}$ N(0,1)

• It is a general result. When sums of random variables are involved, eventually (sometimes after transformations) the CLT can be applied.

The CLT: Review

• Two popular versions in Greene, used in economics and finance:

Lindeberg-Levy: $\{x_n\}$ are *i.i.d.,* with finite μ and finite σ^2.

Lindeberg-Feller: $\{x_n\}$ are independent, with finite $\mu_i, \sigma_i^2 < \infty$, $S_n = \sum_i^n x_i$, $s_n^2 = \sum_i^n \sigma_i^2$ and for ε > 0,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|x_i - \mu_i| > \varepsilon s_n} (x_i - \mu_i)^2 f(x_i)dx = 0$$

Note:

Lindeberg-Levy assumes random sampling – observations are *i.i.d.,* with the same mean and same variance.

Lindeberg-Feller allows for heterogeneity in the drawing of the observations --through different variances. The cost of this more general case: More assumptions about how the $\{x_n\}$ vary.

Asymptotic Distribution of OLS

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

Using Slutzky's theorem for RV, we know the limiting distribution of **b** is not affected by replacing **(X'X)** by its plim. That is, we examine the limiting distribution of

$$\boldsymbol{\beta} + \mathbf{Q}^{-1}\,\mathbf{X'}\boldsymbol{\varepsilon}/T$$

• Notice $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$. But, it has no distribution! It is O$(1/T)$.

We need to do a stabilizing transformation –i.e., the moments do not depend on *T*. Steps:

(1) Stabilize the variance: Var[$\sqrt{T}\,\mathbf{b}$] ~ $\sigma^2\mathbf{Q}^{-1}$ is O(1)

(2) Stabilize the mean: E[$\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta})$] = **0**

Now, we have a RV, $\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta})$, with finite mean and variance .

Asymptotic Distribution of OLS

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

The stabilizing transformation of **b** gives us:

$$\begin{aligned}\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta}) &= \sqrt{T}\,(\mathbf{X'X})^{-1}\mathbf{X'}\,\boldsymbol{\varepsilon}\\ &= \sqrt{T}\,(\mathbf{X'X}/T)^{-1}(\mathbf{X'}\boldsymbol{\varepsilon}\,/T)\end{aligned}$$

The limiting behavior of $\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta})$ is the same as that of

$$\sqrt{T}\,\mathbf{Q}^{-1}\,(\mathbf{X'}\boldsymbol{\varepsilon}/T)$$

Q is a fixed matrix. Asymptotic behavior depends on the RV

$$\sqrt{T}\,(\mathbf{X'}\boldsymbol{\varepsilon}/T)$$

• $\sqrt{T}\,(\mathbf{X'}\boldsymbol{\varepsilon}/T) = \sqrt{T}\,\sum_i^n \boldsymbol{x}_i\varepsilon_i/T = \sqrt{T}\,[1/T\,\sum_i^n \boldsymbol{w}_i] = \sqrt{T}\,\bar{w}$

⇒ $\bar{w}$ is a sample mean of independent observations.

Asymptotic Distribution of OLS

⇒ $\bar{w} = \sum_i^n \boldsymbol{x}_i \varepsilon_i / T$ is a sample mean of independent observations.

- Under the new assumptions, $\bar{w} \xrightarrow{p} \mathbf{0}$ (already shown).

- Since Var[$\mathbf{x}_i\varepsilon_i$] = $\sigma^2\, \mathbf{x}_i{}'\mathbf{x}_i$ ⇒ Var[$\bar{w}$] $\xrightarrow{p} \sigma^2\, \mathbf{Q}/T$

  ⇒ Var[$\sqrt{T}\,\bar{w}$] $\xrightarrow{p} \sigma^2\mathbf{Q}$

- We can apply the Lindeberg-Feller CLT: $\sqrt{T}\,\bar{w} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{Q})$

  ⇒ $\sqrt{T}\,\bar{w} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{Q})$

  ⇒ $\mathbf{Q}^{-1}\sqrt{T}\,\bar{w} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{Q}^{-1}\, \mathbf{Q}\, \mathbf{Q}^{-1}) = \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{Q}^{-1})$

  ⇒ $\sqrt{T}\,(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{Q}^{-1})$

  ⇒ $\mathbf{b} \xrightarrow{a} \mathrm{N}(\boldsymbol{\beta}, (\sigma^2/T)\, \mathbf{Q}^{-1})$

Asymptotic Distribution of OLS

- Note: The last step is a significant jump. We go from an asymptotic distribution to an approximation that we use in small samples. That is, the last step embodies a significant assumption. Now, we say:

$$\mathbf{b} \xrightarrow{a} \mathrm{N}(\boldsymbol{\beta}, (\sigma^2/\,T)\mathbf{Q}^{-1})$$

Phillips (1983, *Handbook of Econometrics*) remarks:

> "For the process by which asymptotic machinery works inevitably washes out sensitivities that are present and important in finite samples".

Asymptotic Results

- How 'fast' does **b** converges to **β**?
  Asy.Var[**b**] = $\sigma^2/T\ \mathbf{Q}^{-1}$ is O(1/*T*)
  – Convergence is at the rate of $1/\sqrt{T}$ –usual square root rate
  – $\sqrt{T}$ **b** has variance of O(1)

- Distribution of **b** does not depend on normality of **ε**

- Estimator of the Asy Var[**b**] = $(\sigma^2/T)\mathbf{Q}^{-1}$ is $(s^2/T)\ (\mathbf{X}'\mathbf{X}/T)^{-1}$. (The degrees of freedom correction is irrelevant. It may matter for small sample behavior.)

- Slutsky's theorem and the delta method apply to functions of **b**.

Test Statistics

- We have established the asymptotic distribution of **b**. We now turn to the construction of test statistics, which are functions of **b**.

- Again, we know that if (**A5**) $\boldsymbol{\varepsilon}|\mathbf{X}$ ~N(**0**, $\sigma^2\mathbf{I}_T$), the Wald statistic
  $F[J, T-k] = (1/J)(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}\ s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \sim\ F_{J,T-k}$

- Q: What is the distribution of F when (**A5**) is no longer assumed? Again, we will study the distribution of test statistics when $T \to \infty$.

Wald Statistics: Cheat sheet

• Recall these results:

- A square of a N(0, 1) RV ~ χ_1^2

- If z ~ N[μ, σ^2] $\Rightarrow [(z-\mu)/\sigma]^2 \sim \chi_1^2$

- Let z_n be not normally distributed, with E[z_n] = μ & Var[z_n] = σ^2. Then, by CLT,

$$\Rightarrow \tau_n = (z_n - \mu)/\sigma \xrightarrow{d} N[0, 1].$$

- If the preceding holds $\Rightarrow (\tau_n)^2 = [(z_n - \mu)/\sigma\,]^2 \xrightarrow{d} \chi_1^2.$

- Let σ be unknown. We use s_n such that plim $s_n = \sigma$.

$$\Rightarrow t_n = [(z_n - \mu)/s_n] \xrightarrow{d} N(0,1) \quad \text{(Slutzky's theorem)}$$

Wald Statistics: Cheat sheet

- A sum of k independent squared N(0, 1) RV ~ χ_k^2

- If **z** is a Txk vector, where **z** ~ N(**μ, Σ**)

$$\Rightarrow W = (\mathbf{x} - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_{Rank(\boldsymbol{\Sigma}))}$$

- Let $\boldsymbol{z}_n \xrightarrow{d}$ N(**μ**, **Σ**). Also, suppose that **Σ** is replaced by a consistent matrix $\boldsymbol{S}_n$ –i.e., plim $\mathbf{S}_T$ = **Σ**. Then,

$$W = (\boldsymbol{z}_n - \boldsymbol{\mu})' \, \boldsymbol{S}_n^{-1} (\boldsymbol{z}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_{Rank(\boldsymbol{S}_n)}$$

• <u>Note</u>: No normal distribution for **z** needed. What we used is consistency of a certain estimator ($\boldsymbol{S}_n$) and the CLT for $\boldsymbol{z}_n$.

Wald Statistics: The *F-test*

- Let $\boldsymbol{z_n} \xrightarrow{d}$ N(μ, Σ) and plim $\boldsymbol{S_n}$ = Σ. Then,

$$W = (\boldsymbol{z_n} - \mu)' \boldsymbol{S_n}^{-1} (\boldsymbol{z_n} - \mu) \xrightarrow{d} \chi^2_{Rank(\boldsymbol{S_n})}$$

- Now, we derive the asymptotic distribution for the *F*-statistic for H$_0$: $\mathbf{R\beta} - \mathbf{q} = \mathbf{0}$

$$F = (1/J)\,(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}\, s^2 (\mathbf{X'X})^{-1}\mathbf{R'}]^{-1}(\mathbf{Rb} - \mathbf{q})$$

Let $\mathbf{m} = (\mathbf{R}\, \mathbf{b}_n - \mathbf{q})$,

⇒ Under H$_0$, plim $\mathbf{m} = \mathbf{0}$ and Var[$\mathbf{m}$] = $\mathbf{R}(\sigma^2/T)\mathbf{Q}^{-1}\mathbf{R}'$.

By CLT,

$$\sqrt{T}\,\mathbf{m} \xrightarrow{d} N[0, \mathbf{R}(\sigma^2)\mathbf{Q}^{-1}\mathbf{R}']$$

Then, by Slutzky's theorem (using plim $s^2 = \sigma^2$)

$$J\,F \xrightarrow{d} \chi^2_{Rank(\text{Var}[\boldsymbol{m}])}$$

Hypothesis Test: Central and Non-central χ^2

- Recall: The *noncentral* χ^2 distribution is "pushed to the right" relative to the (*central*) χ^2. For a given value q,

Prob[χ^2_1 * [½μ^2] > q] is larger than Prob[χ^2_1 > q].

- In our hypothesis testing context: H$_0$:$\mathbf{R\beta} - \mathbf{q} = \mathbf{0}$. The "**z**" in the quadratic form is $\mathbf{Rb} - \mathbf{q}$. The hypothesis is that E[$\mathbf{Rb} - \mathbf{q}$] = **0.**

- If H$_0$ is true –i.e., the expectation really is **0**–, *W* will follow a χ^2 distribution. If the mean is not zero –i.e., H$_1$ is true–, *W* is likely to be larger than we would "predict" based on the (central) χ^2.

- Thus, we construct a test statistic, the Wald statistic, based on the (central) χ^2 distribution. Most Neyman-Pearson tests can be cast in this form. Analysis of the power of a test will need a noncentral $\chi^{2.}$.

The Delta Method

• The *delta method* is used to obtain the asymptotic distribution of a non-linear function of random variables (usually, estimators). It uses a first-order Taylor series expansion and Slutsky's theorem.

• **Univariate case**

Let x_n be a RV, with plim $x_n = \theta$ and Var(x_n) = $\sigma^2 < \infty$.

We can apply the CLT to obtain $\sqrt{n}\,(x_n - \mu)/\sigma \xrightarrow{d}$ N(0, 1).

Goal: $g(x_n) \xrightarrow{a}$? (g(x_n) is a continuous differentiable function, independent of n.)

Steps:

(1) Taylor series approximation around θ:

$$g(x_n) \approx g(\theta) + g'(\theta)\,(x_n - \theta) + \text{higher order terms}$$

We will assume the higher order terms are o(n).

The Delta Method

Remark: o(n): as n grows the higher order terms vanish.

(2) Use Slutsky theorem: plim $g(x_n) = g(\theta)$
plim $g'(x_n) = g'(\theta)$

Then, as n grows, $g(x_n) \approx g(\theta) + g'(\theta)\,(x_n - \theta)$

$$\Rightarrow \quad \sqrt{n}\,[g(x_n) - g(\theta)]) \approx g'(\theta)\,[\sqrt{n}(x_n - \theta)]$$

$$\Rightarrow \quad \sqrt{n}\,([g(x_n) - g(\theta)]\,/\sigma) \approx g'(\theta)\,[\sqrt{n}(x_n - \theta)/\sigma]$$

If *g*(.) does not behave badly, the asymptotic distribution of $(g(x_n) - g(\theta))$ is given by that of $[\sqrt{n}(x_n - \theta)/\sigma]$, which is a standard normal. For the approximation to work well, we want σ to be "small."

Then,

$$\sqrt{n}\,[g(x_n) - g(\theta)] \xrightarrow{a} \text{N}(0, [g'(\theta)]^2\,\sigma^2).$$

The Delta Method

Then,

$$\sqrt{n}\,[\mathrm{g}(x_n) - \mathrm{g}(\theta)] \xrightarrow{a} \mathrm{N}(0, [\mathrm{g}'(\theta)]^2\, \sigma^2).$$

After some work ("inversion"), we obtain:

$$\mathrm{g}(x_n) \xrightarrow{a} \mathrm{N}(\mathrm{g}(\theta), [\mathrm{g}'(\theta)]^2\, \sigma^2).$$

- If we want to test H_0: $\mathrm{g}(\theta) = g_0$, we can do a Wald test:

$$\mathrm{W} = [\mathrm{g}(x_n) - g_0]^2/[[\mathrm{g}(x_n)]^2\, s^2/n] \xrightarrow{a} \chi_1^2$$

The Delta Method

- **Multivariate case**

The extension is straightforward.

Now, we have a vector, $\boldsymbol{x}_n$, that can be asymptotically approximated by a multivariate normal:

$$\boldsymbol{x}_n \xrightarrow{a} \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

Then,

$$\mathrm{g}(\boldsymbol{x}_n) \xrightarrow{a} \mathrm{N}(g(\boldsymbol{\theta}), [\mathrm{g}'(\boldsymbol{\theta})]'\, \boldsymbol{\Sigma}\, [\mathrm{g}'(\boldsymbol{\theta})]).$$

The Delta Method – Example 1

Let $x_n \xrightarrow{a} \mathrm{N}(\theta, \sigma^2/n)$

Then, $g(x_n) = \delta/x_n \xrightarrow{a}$? (δ is a constant)

(1) Calculate the plims of $g(x_n)$ and $g'(x_n)$:

$g(x_n) = \delta/x_n \quad \Rightarrow \text{plim } g(x_n) = (\delta/\theta)$

$g'(x_n) = -(\delta/x_n^2) \quad \Rightarrow \text{plim } g'(x_n) = -(\delta/\theta^2)$

(2) Use delta method formula: $g(x_n) \xrightarrow{a} \mathrm{N}(g(\theta), [g'(\theta)]^2\, \sigma^2/n)$.

$$\Rightarrow \quad g(x_n) \xrightarrow{a} \mathrm{N}(\delta/\theta, (\delta^2/\theta^4)\sigma^2/n)$$

- If we want to test H_0: $g(x_n) = g_0$, we can do a Wald test:

$$W = [\delta/x_n - g_0]^2/[(\delta^2/x_n^4)s^2/n] \xrightarrow{a} \chi_1^2$$

The Delta Method – Example 2

Let

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} \xrightarrow{a} N\left(\begin{bmatrix} \theta_x \\ \theta_y \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}\right)$$

Define R = x_n / y_n.

Q: What is the Var(R) = ?

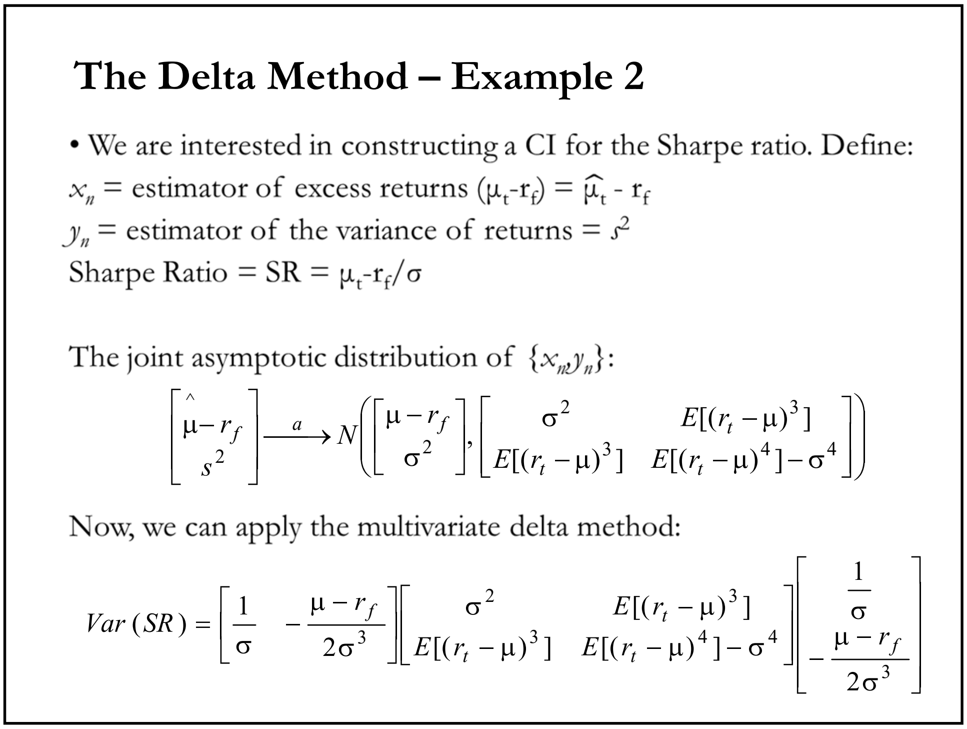
(1) Calculate the plims of $g(x_n)$ and $g'(x_n)$:

$g(R_n) = x_n / y_n \quad \Rightarrow \text{plim } g(R_n) = (\theta_x/\theta_y)$

$g'(R_n) = [(1/y_n) \quad (-x_n / y_n^2)] \quad \Rightarrow \text{plim } g'(R_n) = [(1/\theta_y) \quad (-\theta_x/\theta_y^2)]'$

(2) Multivariate delta method: $g(x_n) \xrightarrow{a} \mathrm{N}(g(\theta), [g'(\theta)]'\, \boldsymbol{\Sigma}\, [g'(\theta)]/n)$.

$$Var(R_n) = \begin{bmatrix} \frac{1}{\theta_y} & -\frac{\theta_x}{\theta_y^2} \end{bmatrix} \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} \frac{1}{\theta_y} \\ -\frac{\theta_x}{\theta_y^2} \end{bmatrix} = \frac{\sigma_{xx}}{\theta_y^2} - \frac{\theta_x \sigma_{xy}}{\theta_y^3} - \frac{\theta_x \sigma_{xy}}{\theta_y^3} + \frac{\theta_x^2 \sigma_{yy}}{\theta_y^4}$$



The Delta Method – Example 2

- We are interested in constructing a CI for the Sharpe ratio. Define:

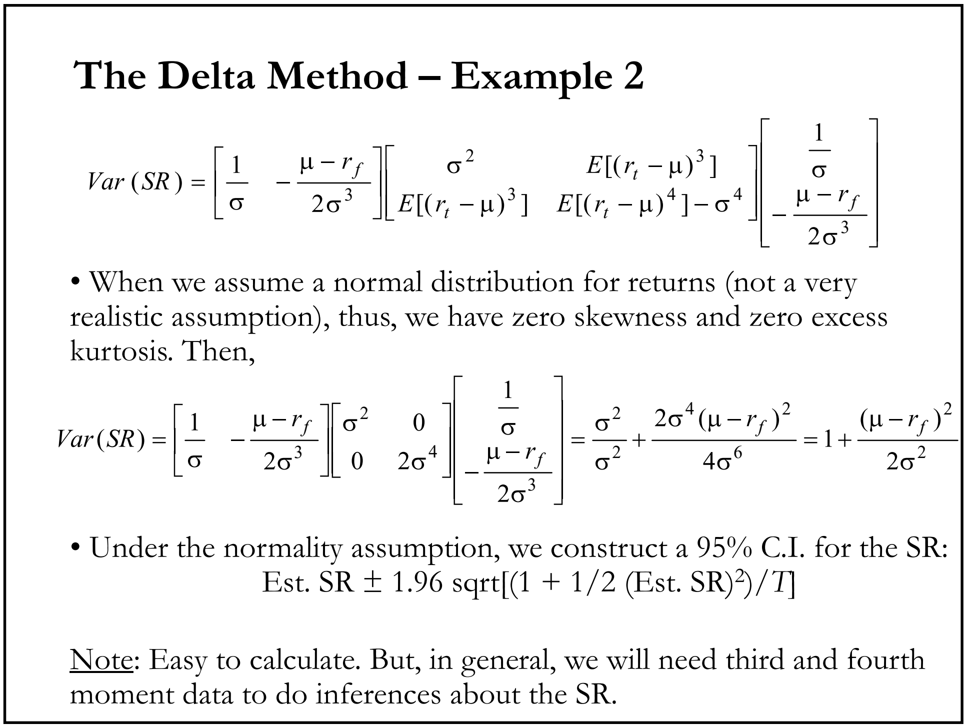
x_n = estimator of excess returns (μ_t-r_f) = $\hat{\mu}_t$ - r_f
y_n = estimator of the variance of returns = s^2
Sharpe Ratio = SR = μ_t-r_f/σ

The joint asymptotic distribution of $\{x_n, y_n\}$:

$$\begin{bmatrix} \hat{\mu} - r_f \\ s^2 \end{bmatrix} \xrightarrow{a} N\left(\begin{bmatrix} \mu - r_f \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & E[(r_t-\mu)^3] \\ E[(r_t-\mu)^3] & E[(r_t-\mu)^4]-\sigma^4 \end{bmatrix} \right)$$

Now, we can apply the multivariate delta method:

$$Var(SR) = \begin{bmatrix} \frac{1}{\sigma} & -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & E[(r_t-\mu)^3] \\ E[(r_t-\mu)^3] & E[(r_t-\mu)^4]-\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix}$$



The Delta Method – Example 2

$$Var(SR) = \begin{bmatrix} \frac{1}{\sigma} & -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & E[(r_t-\mu)^3] \\ E[(r_t-\mu)^3] & E[(r_t-\mu)^4]-\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix}$$

- When we assume a normal distribution for returns (not a very realistic assumption), thus, we have zero skewness and zero excess kurtosis. Then,

$$Var(SR) = \begin{bmatrix} \frac{1}{\sigma} & -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\mu - r_f}{2\sigma^3} \end{bmatrix} = \frac{\sigma^2}{\sigma^2} + \frac{2\sigma^4(\mu - r_f)^2}{4\sigma^6} = 1 + \frac{(\mu - r_f)^2}{2\sigma^2}$$

- Under the normality assumption, we construct a 95% C.I. for the SR:

Est. SR ± 1.96 sqrt[(1 + 1/2 (Est. SR)^2)/T]

<u>Note</u>: Easy to calculate. But, in general, we will need third and fourth moment data to do inferences about the SR.

Trilogy of Asymptotic Tests: LR, Wald, and LM

- We assume we know the distribution for **ε**|**X**. to build a likelihood function, L(**β**). We present three asymptotic tests: LR, Wald, and LM.

- Notation:
  $\hat{\beta}^U$: Unrestricted estimator (MLE, $\hat{\beta}^U = \hat{\beta}_{ML}$)
  $\hat{\beta}^R$: Restricted estimator, imposing H_0: g(**β**)= **Rβ - q** = **0**.

- From a Taylor expansion, we get the basic tests:

$$(\hat{\theta}\text{-}\theta_0)'Var(\theta)^{-1}(\hat{\theta}\text{-}\theta_0)\xrightarrow{d}\chi^2_J$$

- Differences in the construction.

- Likelihood Ratio (LR) test: Use both $\hat{\beta}^U$ and $\hat{\beta}^R$
- Wald test: Use $\hat{\beta}^U$
- Lagrange Multiplier (LM) test: Use $\hat{\beta}^R$ (a test of H_0: **λ**=**0**.)

Trilogy of Asymptotic Tests: Wald, LM and LR

- We want to test J restrictions H_0: g(**β**) = **0**, where g(.) is a nice differentiable vector function, with 1st derivative matrix **G**.

- *The Likelihood Ratio (LR) test*

We estimate the model twice, once unrestricted and once restricted. Then, we compare the two.

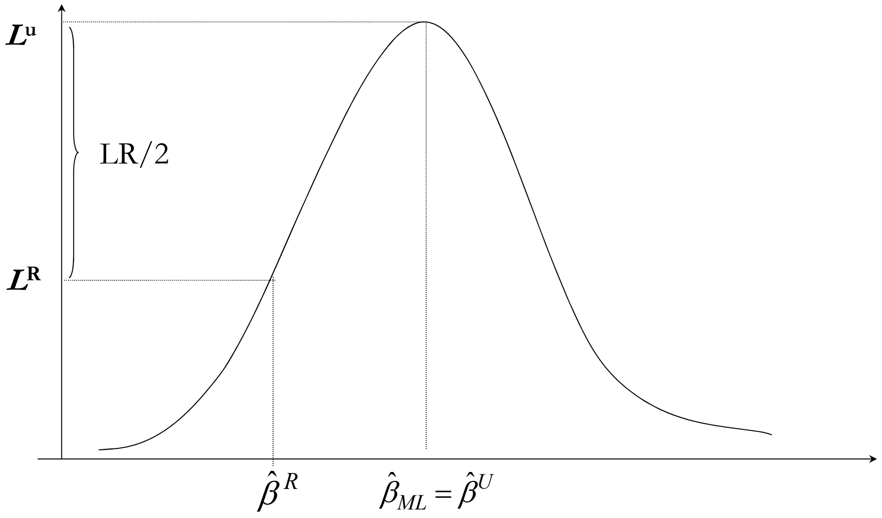
- *The Wald (W) test*

We estimate only the unrestricted model. We use an estimate of the second derivative to `guess' the restricted model.

- *The Lagrange Multiplier (LM) test*

We estimate only the restricted model. We use again an estimate of the second derivatives to guess the restricted model. Under H_0, the LM associated with the constraint, **λ**, is zero.

Trilogy of Asymptotic Tests: LR Test



If the values of $\hat{\beta}^U$ and $\hat{\beta}^R$ are far apart, then L^U and L^R will be far apart, the test statistic will be large, and we will reject H_0.

Trilogy of Asymptotic Tests: LR Test

- Recall that we can approximate the likelihood function, L, using a 2nd-order Taylor series expansion around $\hat{\beta}_{ML}$:

$$\log(L(\beta)) = \log(L(\hat{\beta}_{ML})) + (\hat{\beta}_{ML} - \beta) S(\hat{\beta}_{ML}) + \frac{1}{2} (\hat{\beta}_{ML} - \beta)' I_n(\hat{\beta}_{ML}) (\hat{\beta}_{ML} - \beta)$$

where $\frac{\delta \log(L(\beta))}{\delta\beta} = \sum_{i=1}^{n} \frac{\delta \log(f(x_i \mid \beta))}{\delta\beta} = S(\beta) \quad \Rightarrow S(\hat{\beta}_{ML}) = 0$

$$\frac{\delta^2 \log(L(\beta))}{\delta\beta\delta\beta'} = I_n(\beta) = n \, \mathrm{I}(\beta)$$

$$E\left[\left(\frac{\partial \log f(x;\beta)}{\partial\beta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(x;\beta)}{\partial\beta\partial\beta'}\right] = I(\beta)$$

- Then,

$$\log(L(\beta)) = \log(L(\hat{\beta}_{ML})) + 0.5(\hat{\beta}_{ML} - \beta)' I_n(\beta)(\hat{\beta}_{ML} - \beta)$$

Trilogy of Asymptotic Tests: LR Test

- Under H$_0$, we do the approximation around $\hat{\beta}^R$:

$$\log(L(\beta)) = \log(L(\hat{\beta}^R)) + \frac{1}{2}(\hat{\beta}^R - \beta)' I_n(\hat{\beta}^R)(\hat{\beta}^R - \beta)$$

- Subtracting both expression, under H$_0$, we get

$$\log(L(\hat{\beta}_{ML})) - \log(L(\hat{\beta}^R)) = \frac{1}{2}(\hat{\beta}_{ML} - \hat{\beta}^R)' I_n(\hat{\beta}_{ML})(\hat{\beta}_{ML} - \hat{\beta}^R)$$

- Under the regularity conditions in SR-11, we have shown:

$$n(\hat{\beta}_{ML} - \hat{\beta}^R)' I(\hat{\beta}_{ML})(\hat{\beta}_{ML} - \hat{\beta}^R) \xrightarrow{d} \chi^2_J$$

- Then,

$$LR = 2[\log(L(\hat{\beta}_{ML})) - \log(L(\hat{\beta}^R))] \xrightarrow{d} \chi^2_J$$

Trilogy of Asymptotic Tests: LR Test

- An asymptotic test which rejects H$_0$ with probability one when the H$_1$ is true is called a *consistent test*. That is, a consistent test has asymptotic power of 1.

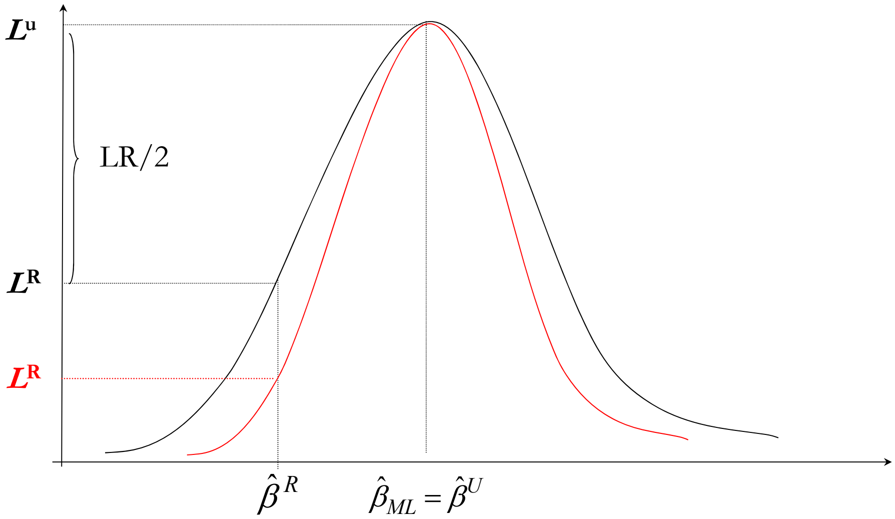
- The LR test is a consistent test. A heuristic argument is that if the alternative hypothesis is true instead of H$_0$, then

$$(\hat{\beta}^U - \hat{\beta}^R) \xrightarrow{p} k_0 \neq 0$$

Then, under H$_1$, $(\hat{\beta}_{ML} - \hat{\beta}^R)' I(\hat{\beta}_{ML})(\hat{\beta}_{ML} - \hat{\beta}^R)$ converges to *k*, a constant, not 0. Multiplying a constant by *n,* get the LR to diverge to ∞ as $n \to$ ∞,which implies that we always reject H$_0$ when H$_1$ is true.

Alternatively, we can think of $g(\hat{\beta}^U) \xrightarrow{p} g(\beta_0) \neq 0$ driving the divergence of the LR test under H$_0$ when H$_1$ is true.

Trilogy of Asymptotic Tests: Wald Test



• Consider two likelihood functions, ***L*** and ***L***. They have the same values of $\hat\beta^U$ and $\hat\beta^R$ –i.e., same $|\hat\beta^U$ - $\hat\beta^R|$. But different LR tests.

Trilogy of Asymptotic Tests: Wald Test

• The fact that the LR test statistic is affected by the curvature of L suggests a rescaling for $|\hat\beta^U$ - $\hat\beta^R|$. Use the second derivative of the likelihood function.

• Formally, recall that under the usual regularity conditions, the ML estimator $\hat\beta^U = \hat\beta_{ML}$ is asymptotically normal:

$$\hat\beta_{ML} \xrightarrow{a} N(\beta, I(\beta)^{-1}) \qquad (*)$$

• Then, under H_0, the Wald test can be calculated as:

$$W=(\hat\beta^U-\beta_0)'[Var(\hat\beta^U)]^{-1}(\hat\beta^U-\beta_0)=(\hat\beta^U-\beta_0)'I(\hat\beta^U)(\hat\beta^U-\beta_0)\xrightarrow{d}\chi_J{}^2$$

• We can also derive a Wald test by combining (*) and a 1st-order Taylor expansion of the restriction g(β), around the true value β_0.

Trilogy of Asymptotic Tests: Wald Test

- We can also derive a Wald test by combining (*) and a 1st-order Taylor expansion of the restriction $g(\beta)$, around the true value β_0:

$$g(\hat{\beta}^U) = g(\beta_0) + G'(\hat{\beta}^U - \beta_0) + o(1)$$

- A little bit of algebra and (*) deliver:

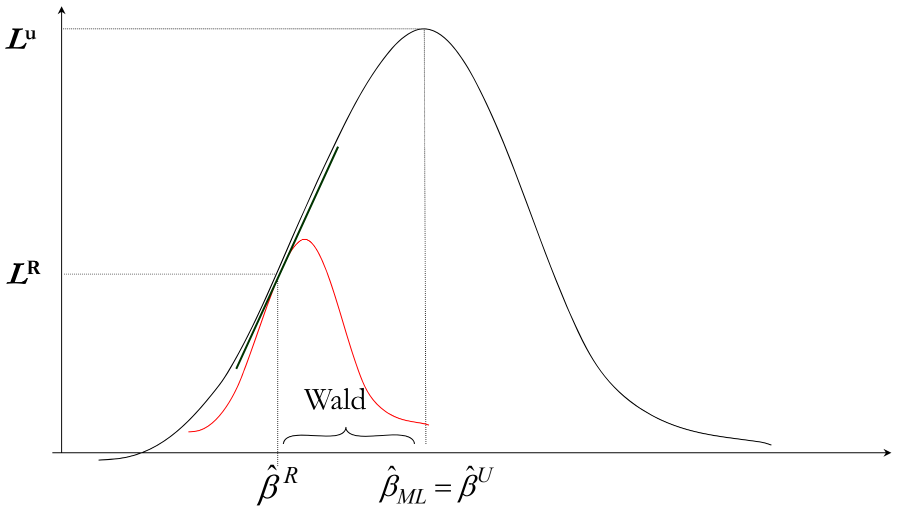
$$\sqrt{n}(g(\hat{\beta}^U) - g(\beta_0)) \xrightarrow{d} N(0, G' I(\beta_0)^{-1} G)$$

- Under H_0: $g(\beta_0)=0$, we form the usual quadratic form for a Wald test:

$$n\, g(\hat{\beta}^U)'[G' I(\hat{\beta}^U)^{-1} G]^{-1} g(\hat{\beta}^U) \xrightarrow{d} \chi^2_J$$

where we use $\hat{\beta}^U$ to evaluate $\mathbf{I}(\beta)$ and $\mathbf{G}$.

Trilogy of Asymptotic Tests: LM Test



- Consider two likelihood functions, ***L*** and ***L***. They have the same slope at $\hat{\beta}^R$. But the distance $|\hat{\beta}^U - \hat{\beta}^R|$. will be greater for ***L***.

Trilogy of Asymptotic Tests: LM Test

• Recall that in the context of the CLM in Chapter 4, we derive a Wald test based on the LM.

Suppose we just test H$_0$: $\boldsymbol{\lambda} = \mathbf{0}$, using the Wald criterion:

$$W = \boldsymbol{\lambda}'(\text{Var}[\boldsymbol{\lambda}\,|\,\mathbf{X}])^{-1}\boldsymbol{\lambda}$$

• The LM test uses the curvature at the restricted estimator, $\hat{\beta}^R$, to test if it is closed to 0 (the f.o.c.).

• Thus, in the basic formula $(\hat{\theta} - \theta_0)' Var(\hat{\theta})^{-1}(\hat{\theta} - \theta_0)$ a rescaling is suggested, by a measure of L's curvature, as given by the score, $S(\beta)$.

Trilogy of Asymptotic Tests: LM Test

• Recall the score function, $S(\beta) \xrightarrow{a} N(0, n\text{I}(\beta))$.

• The LM (*score*) test is based on $S(\beta)$. Under H$_0$, $S(\beta^R)=0$. Then,

$$LM = \frac{1}{n} S(\beta^R)'[I(\beta^R)]^{-1} S(\beta^R) \xrightarrow{\text{d}} \chi_J^{\,2}$$

• The information matrix , $\mathbf{I}(\beta)$ may be evaluated at the hypothesized value β^R or at the $\hat{\beta}_{ML}$. One advantage of using $\hat{\beta}^R$ is that we bypass the calculation of the ML. In practice, score tests are seldom used.

• Recall that we can write $\text{I}(\beta)$ as the expectation of a product of scores, $S(\beta)$. We can rewrite the LM test as

$$LM = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta \log(f(x_i\,|\,\theta))}{\delta\theta'} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\delta \log(f(x_i\,|\,\theta))}{\delta\theta} \text{ x } \frac{\delta \log(f(x_i\,|\,\theta))}{\delta\theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\delta \log(f(x_i\,|\,\theta))}{\delta\theta'}$$

This expression looks like an R^2 from the regression of 1 on $S(\beta)$.

Trilogy of Asymptotic Tests: LM Test

- The (uncentered) R^2: $R^2 = \dfrac{i' X(X' X)^{-1} X' i}{i' i}$

- Then, LM = n R^2, where R^2 is calculated from a regression of a vector of ones on the scores. (This version of the LM test may be referred as Engle's LM test.)

Example: $Y_t = f(X_t, \beta_1, \beta_2) + \varepsilon_t$ subject to $g(\beta_1) = 0$

where β_1 is a subset of parameters restricted by $g(\beta_1) = 0$ (***G*** is the 1st derivative of this restriction).

- Then,

$$S(\beta_1) = \frac{G' e^*}{\sigma^2}$$

$$I(\beta_1) = [\sigma^{-2} E(G' G)]^{-1}$$

e*: estimated errors from the restricted model

Trilogy of Asymptotic Tests: LM Test

- The LM test becomes: LM = σ^{-2} $\mathbf{e}^{*\prime}\mathbf{G}\{\sigma^{-2} E[\mathbf{G}'\mathbf{G}]\}^{-1} \sigma^{-2} \mathbf{G}'\mathbf{e}^*$

If $E[\mathbf{G}'\mathbf{G}] = \mathbf{G}'\mathbf{G}$ $\Rightarrow$ LM = σ^{-2} $\mathbf{e}^{*\prime}\mathbf{e}^*$,

with the interpretation as TR^2 from a regression of **e*** on **G**. This is used in many tests for serial correlation, heteroskedasticity, etc.

Example: Testing for serial correlation
Suppose

$$Y = X\beta + \varepsilon$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

H_0: $\rho=0$ (no serial correlation).

To calculate the Engle's LM test we need **e***(residuals under ρ=0) and **G**.

Trilogy of Asymptotic Tests: LM Test

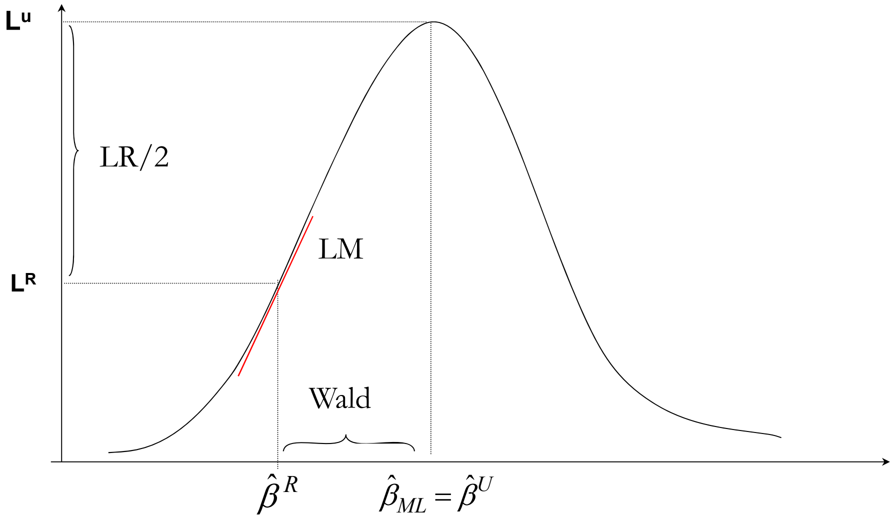
Example(continuation): Serial correlation
Steps to calculate Engle's LM test for serial correlation.:
(1) Estimate restricted model –i.e., without serial correlation: g: ρ=0.
(2) Save the residuals **e***.
(3) Get **G**= **e*$_{-1}$** (lagged residuals). Then, estimate the model:

$$e_t^* = x_t'\gamma + \theta\, e_{t-1}^* + \nu_t$$

and keep the R^2 from this regression .

(4) LM = (T – 1) R^2 ~ χ_1^2

Trilogy of Asymptotic Tests: Wald, LM and LR



If the likelihood function were quadratic then LR = LM = W. In general, however W > LR > LM.

Trilogy of Asymptotic Tests: Wald, LM and LR

• The three tests have the same asymptotic distribution: Equivalent as $T \to \infty$. Thus, they are expected to give similar results in large samples.

• Their small-sample properties are not known. Simulation studies suggest that the likelihood ratio test may be better in small samples.

• The three tests are consistent, pivotal tests.

Asymptotic Tests: Small sample behavior?

• The *p-values* from asymptotic tests are approximate for small samples. We worry that tests based on them may over-reject in small samples (or under-reject). The conclusions may end up being too liberal (or too conservative).

• Whether they over-reject or under-reject, and how severely, depends on many things:

(1) Sample size, T.

(2) Distribution of the error terms, $\boldsymbol{\varepsilon}$.

(3) The number of regressors, k, and their properties

(4) The relationship between the error terms and the regressors.

• A simulation can help.

Pivot Tests: Review

Definition: Pivot Test

A *pivot* is a statistic whose sampling distribution does not depend on unknown parameters. A test whose distribution does not depend on unknown parameters is called a *pivot test*.

Example: Suppose you draw X from a N(μ, σ^2).

Asymptotic theory implies that $\bar{x} \xrightarrow{d} N(\mu, \sigma^2/N)$.

This statistic is *not* asymptotically pivotal statistic because it depends on an unknown parameter, σ^2 (even if you specify μ_0 under H_0).

On the other hand, the *t-statistic*, $t = (\bar{x} - \mu_0)/s \xrightarrow{d} N(0,1)$.

⇒ The *t-statistic* is asymptotically pivotal since 0 and 1 are known!

Pivot Tests: Review

• Most statistics are *not* asymptotically pivotal. Popular test statistics in econometrics, however, are asymptotically pivotal. For example, Wald, LR and LM tests are distributed as χ^2 with known df.

• Keep in mind that in finite samples, most tests are not still asymptotically pivotal in finite samples.

• Under the usual assumptions, the $t = (b - \beta)/SE(b)$ is an example of a pivotal statistic: $t \sim t_{T-k}$, which does not depend on β.

Note: These functions depend on the parameters, but the distribution is independent of the parameters.

Bootstrap: Review

• Recall the *Fundamental Theorem of Statistics*:

The empirical cdf → true CDF.

• Then, if we knew that a certain test statistic was pivotal but did not know how it was distributed, we could select any DGP in H_0 and generate simulated samples from it.

Trick: Simulation to get the empirical cdf of a statistic. Apply *FTS*.
⇒ Sample from our ED to collect *B* samples of size *T*.

• The simulated samples can be used to construct a C.I., which easily allows us to test a H_0.

Bootstrap: Testing – Example IBM

• We want to test if IBM has an average return equal to the market, proxied by the S&P 500. H_0: $\mu_{IBM} = \mu_{Market}$ = **0.76%** monthly (or 9.5% annualized, based on 1928-2015 data).

We have monthly data from 1990: Jan to 2016: August (*T*=320). The average IBM return was **0.9%**. We do a bootstrap to check the sampling distribution of IBM mean returns.

Steps:
(1) Draw ***B* = 10,000** bootstrap subsamples of size *T* = 320.
(2) For each subsample, calculate the observed sample mean, $\bar{x}_j^*$.
(3) Draw a histogram

From the histogram, it is straightforward to build a (1-α)% C.I. and check if the market return (**0.76%**) falls within it.

Bootstrap: Testing – Example IBM

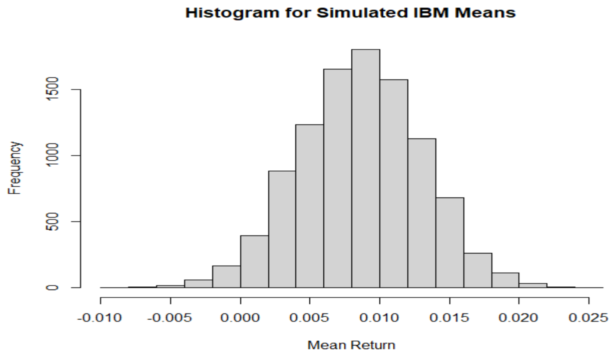
- R-code

```
dat_xy <- read.csv("C:/IFM/datastream-K-DIS.csv", head=TRUE, sep=",")
x <- dat_xy$IBM
 sim_size = 10000

# bootstrap
 bstrap <- c()
for (i in 1:sim_size){
 newsample <- sample(x, 320, replace=T)
 bstrap <- c(bstrap, mean(newsample))}
hist(bstrap,main="Histogram for Simulated IBM Means",  xlab="Mean
Return", breaks=20)

# 95% Confidence Interval
> quantile(bstrap,.025)
> 9.031304e-05
> quantile(bstrap,.975)
```

Bootstrap: Testing – Example IBM



Notes: The simulated IBM mean is **0.0086%.**

- Normal approximation seems OK ⇒ usual t-test should work fine!.

- 95% C.I: [0.00001, 0.01711] ⇒ a sample mean of **0.0076** is possible!

Bootstrap critical values

- Suppose we are interested in testing in the CLM, H_0: $\beta_2 = \beta_{2,0}$.

We use a t-statistic: $t = (b_2 - \beta_{2,0})/SE(b_2)$, which is *asymptotically pivotal* ($\xrightarrow{d}$N(0,1)).

- Bootstrapping can provide some finite sample correction (*refinement*), providing more accurate estimates of the t-test.

Steps:

(1) Draw *B* (say, 999) bootstrap subsamples of size *T* from your data. Denote the subsamples as $w_j^* = \{\mathbf{y}_j, \mathbf{x}_j\}$, j = 1, 2, …, *B*.

(2) For each subsample calculate the observed $t^*_j = (b_{2,j} - \beta_{2,0})/SE(b_{2,j})$.

(3) Sort these *B* estimates from smallest to largest.

Bootstrap critical values

- In the last step, we sort the *B* t_j^* estimates from smallest to largest.

- For an (upper) one-tailed test the *bootstrap critical value* (at level α) is the "upper α*th* quantile" of the *B* estimates of t_j^*. For example, if *B* = 999, choose the critical value of *t* at the 5% level as t_{950}^*.

- For two-sided tests, there are two possibilities.

(1) A *non-symmetrical* or *equal-tailed* test has the same number of bootstrapped estimates in the two tails, which may implies that the critical values for the upper and lower tails are not necessarily equal in absolute value.

For example, for inference at the 5% significance level $|t_{25}^*|$ may not be equal to $|t_{975}^*|$.

Bootstrap critical values

(2) In contrast, a *symmetric test* orders the t_j*'s in terms of their absolute values and the critical value is the single value in absolute value terms. For example, with $B = 999$ and a 5% significance level, order all 999 t_j*'s in terms of their absolute value from smallest to the largest. Pick the one that is 50th from the top and compare that to $|t|$.

<u>Note</u>: An alternative (but "*unrefined*") method to test H_0 is to just use the bootstrapped standard errors to compute

$$t = (b_2 - \beta_{2,0})/SE_{boot}(b_2).$$

Bootstrap critical values

• Another "unrefined" method is not to calculate any SE at all. We use the B b_2*'s estimates to build a confidence interval. Suppose we are interested in a 2-sided test.

Steps:

(1) As usual, calculate B estimates of b_2, call them $b_{2,1}^*, \ldots , b_{2,B}^*$.

(2) At the α% significance level, cut out the bottom (α/2)% and the top (α/2)% estimates of β_2

(3) Reject H_0 if $\beta_{2,0}$ falls outside this range.

This method is called the *percentile method* for conducting hypothesis tests.

Bootstrap *p-values*

- Recall that the *p-value* is the probability, under H_0, of obtaining a more extreme (or equal to) result than what was actually observed.

- We compute *p*-values through a simulation for a test statistic, τ, with observed value $\hat{\tau}$. (τ follows a distribution under H_0; $\hat{\tau}$ is a realization.)

(1) Choose any DGP in H_0, and draw *B* (say, 999) samples of size *T* from it. Denote the simulated samples as y_j^*, j = 1, 2, …, *B*.

(2) For each simulated sample calculate the observed τ^*_j.

(3) Count the number of times the simulated values (τ^*_j's) exceeds the observed value $\hat{\tau}$. Divide by *B*. This is the bootstrapped *p*-value:

$$\hat{p}^*(\hat{\tau}) \equiv \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* > \hat{\tau})$$

Bootstrap *p-values*

- Since the EDF converges to the true CDF, it follows that, if *B* were infinitely large, this procedure would yield an exact test.

- Simulating a pivotal statistic is called a *Monte Carlo test*; Dufour and Khalaf (2001) provides a more detailed introduction and references.

Bootstrap *p-values*: Example IBM

• Now, we do a simulation assuming H_0 is true.

1) We shift the mean of the data in the sample (the "*fake population*")

```
> new_x <- x – mean(x) + .0076
```

2) We do another bootstrap (using the same code as before) with these data to compare with the observed **0.89%** IBM return

3) The *p-value* is the probability of getting something more extreme than what we observed, 0.89%, which is 0.89-0.76%=0.13% units from H_0. For a two-sided test, the *p-value* is given by:

```
> p_val <- (sum(bstrap < 0.0063) + sum(bstrap > 0.0089)/sim_size
> print(p_val)
[1] 0.7494
```

⇒ cannot reject H_0!

Bootstrap Testing: Remarks

• Two types of errors associated with bootstrap testing with p-values:

(1) Most tests are not asymptotically pivotal in finite samples.

The distribution of most test statistics depend on unknown parameters (or other unknown features) of the DGP. Then, bootstrapped *p-values* will be inaccurate, because of the differences between the bootstrap DGP and the true DGP.

Q: How serious is this problem?

Beran (1988), Hall (1992), and Davidson and MacKinnon (1999) argue that bootstrap tests tend to perform better than tests based on approximate asymptotic distributions.

⇒ The errors committed by both tests diminish as T increases, but those committed by bootstrap tests diminish more rapidly.

Bootstrap Testing: Remarks

- Two types of errors associated with bootstrap testing:

(2) B is finite.

An ideal bootstrap test rejects H_0 at level α whenever $p^*(\hat{\tau}) < \alpha$. But, our "feasible" bootstrap test reject H_0 whenever $\hat{p}^*(\hat{\tau}) < \alpha$. If B is extremely large, the difference between feasible and ideal tests will be small. In practice, because of costs, we use small B.

Two consequences of small *B?*

(a) The test may depend on the sequence of random numbers used to generate the bootstrap samples (the seed).

(b) Whenever $B < \infty$, there is loss of power, as discussed in Hall and Titterington (1989). This loss of power is often small, but as pointed out by Davidson and MacKinnon (2001) can get big.

Example: Gasoline Demand (Greene)

- Based on the gasoline data: The regression equation is

$$G = \beta_1 + \beta_2 y + \beta_3 pg + \beta_4 pnc + \beta_5 puc + \\ + \beta_6 ppt + \beta_7 pd + \beta_8 pn + \beta_9 ps + \beta_{10} t + \varepsilon$$

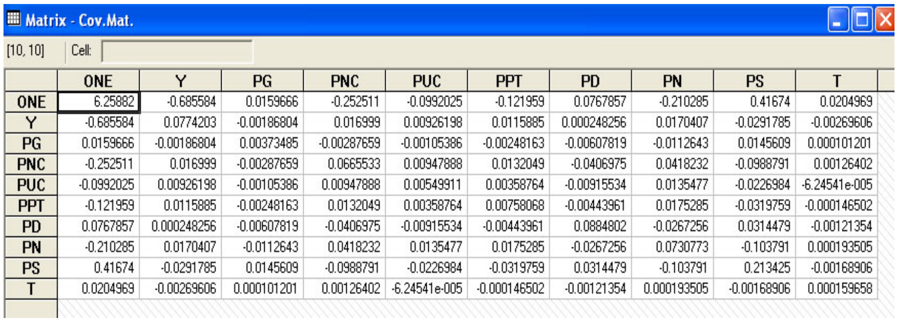
All variables are logs of the raw variables, so that coefficients are elasticities.

The new variable, t, is a time trend, 0, 1, …, 26, so that β_{10} is the autonomous yearly proportional growth in G.

Gasoline Demand - OLS Results (Greene)

```
+----------------------------------------------------+
| Ordinary    least squares regression               |
| LHS=G        Mean                   =   5.308616    |
|              Standard deviation     =   .2313508    |
| Model size   Parameters             =         10    |
|              Degrees of freedom     =         17    |
| Residuals    Sum of squares         =   .003776938  |
|              Standard error of e    =   .01490546   |
| Fit          R-squared              =   .9972859    |
|              Adjusted R-squared     =   .9958490    |
| Model test   F[  9,    17] (prob) = 694.07 (.0000) |
|              Chi-sq [  9]  (prob) = 159.55 (.0000) |
+----------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 Constant     -5.97984140     2.50176400     -2.390   .0287
 Y             1.39438363      .27824509      5.011   .0001    9.03448264
 PG            -.58143705      .06111346     -9.514   .0000     .47679491
 PNC           -.29476979      .25797920     -1.143   .2690     .28100132
 PUC           -.20153591      .07415599     -2.718   .0146     .40523616
 PPT            .08050720      .08706712       .925   .3681     .47071442
 PD            1.50606609      .29745626      5.063   .0001    -.44279509
 PN             .99947385      .27032812      3.697   .0018    -.58532943
 PS            -.81789420      .46197918     -1.770   .0946    -.62272267
 T             -.01251291      .01263559      -.990   .3359    13.0000000
```

Gasoline Demand Covariance Matrix (Greene)



	ONE	Y	PG	PNC	PUC	PPT	PD	PN	PS	T
ONE	6.25882	-0.685584	0.0159666	-0.252511	-0.0992025	-0.121959	0.0767857	-0.210285	0.41674	0.0204969
Y	-0.685584	0.0774203	-0.00186804	0.016999	0.00926198	0.0115885	0.000248256	0.0170407	-0.0291785	-0.00269606
PG	0.0159666	-0.00186804	0.00373485	-0.00287659	-0.00105386	-0.00248163	-0.00607819	-0.0112643	0.0145609	0.000101201
PNC	-0.252511	0.016999	-0.00287659	0.0665533	0.00947888	0.0132049	-0.0406975	0.0418232	-0.0988791	0.00126402
PUC	-0.0992025	0.00926198	-0.00105386	0.00947888	0.00549911	0.00358764	-0.00915534	0.0135477	-0.0226984	-6.24541e-005
PPT	-0.121959	0.0115885	-0.00248163	0.0132049	0.00358764	0.00758068	-0.00443961	0.0175285	-0.0319759	-0.000146502
PD	0.0767857	0.000248256	-0.00607819	-0.0406975	-0.00915534	-0.00443961	0.0884802	-0.0267256	0.0314479	-0.00121354
PN	-0.210285	0.0170407	-0.0112643	0.0418232	0.0135477	0.0175285	-0.0267256	0.0730773	-0.103791	0.000193505
PS	0.41674	-0.0291785	0.0145609	-0.0988791	-0.0226984	-0.0319759	0.0314479	-0.103791	0.213425	-0.00168906
T	0.0204969	-0.00269606	0.000101201	0.00126402	-6.24541e-005	-0.000146502	-0.00121354	0.000193505	-0.00168906	0.000159658

Gasoline Demand - Linear Hypothesis (Greene)

H_0: Aggregate price variables are not significant determinants of gasoline consumption

H_0: $\beta_7 = \beta_8 = \beta_9 = 0$

H_1: At least one is nonzero

$$\mathbf{R\beta - q = 0}$$

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Gasoline Demand - Wald Test (Greene)

```
R = [0,0,0,0,0,0,1,0,0,0/
     0,0,0,0,0,0,0,1,0,0/
     0,0,0,0,0,0,0,0,1,0];
q = [0 / 0 / 0 ] ;

m = R*b - q ;
Vm = R*Varb*R‘
Wald = m‘*inv(Vm)*m;

 WALD = 66.91506
```

Gasoline Demand - Nonlinear Restrictions

• Suppose we are interested in testing the hypothesis that certain ratios of elasticities are equal. In particular,

$$\phi_1 = \beta_4/\beta_5 - \beta_7/\beta_8 = 0$$
$$\phi_2 = \beta_4/\beta_5 - \beta_9/\beta_8 = 0$$

• To do the Wald test, first we estimate the asymptotic covariance matrix for the sample estimates of ϕ_1 and ϕ_2. After estimating the regression by least squares, the estimates are

$$f1 = b_4/b_5 - b_7/b_8$$
$$f2 = b_4/b_5 - b_9/b_8.$$

Then, using the delta method, we estimate the asymptotic variances and covariances of f1 and f2.

Gasoline Demand - Setting Up the Wald Stat

• After estimating the regression by least squares, the estimates are

$$f1 = b_4/b_5 - b_7/b_8$$
$$f2 = b_4/b_5 - b_9/b_8.$$

Then, we use the delta method to get the asymptotic covariance matrix for f1 and f2. We write f1 = f1(**b**), a function of the entire 10×1 coefficient vector. Then, we compute the 1×10 derivative vector, **d**1 = ∂f1(**b**)/∂**b′**. This vector is

$$\begin{array}{lcccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \mathbf{d}1 = & 0, & 0, & 0, & 1/b5, & -b4/b5^2, & 0, & -1/b8, & b7/b8^2, & 0, & 0 \end{array}$$

• Similarly for **d2** = ∂f2(**b**)/∂**b′**.

$$\mathbf{d}2 = 0,\ 0,\ 0,\ 1/b5, -b4/b5^2,\ 0,\ \ 0,\ \ b9/b8^2, -1/b8,\ 0$$

Gasoline Demand - Wald Statistics (Greene)

Then, $\mathbf{D}$ = the 2×10 matrix with first row $\mathbf{d}1$ and second row $\mathbf{d}2$. The estimator of the asymptotic covariance matrix of [f1,f2]′ (a 2×1 column vector) is

$$\mathbf{V} = \mathbf{D} \times s^2 (\mathbf{X}'\mathbf{X})^{-1} \times \mathbf{D}'.$$

Finally, the test of H_0: $\boldsymbol{\phi} = 0$ is done with

$$W = (\mathbf{f\text{-}0})'\mathbf{V}^{-1}(\mathbf{f\text{-}0}) \sim \chi_2^2.$$

The critical value from the chi-squared table is 5.99 at the 5% level. Then, if $W > 5.99 \Rightarrow$ reject H_0.

Computation: $W = 22.65 > 5.99 \quad \Rightarrow$ reject H_0.

Wald Test: Manipulation of variables (Greene)

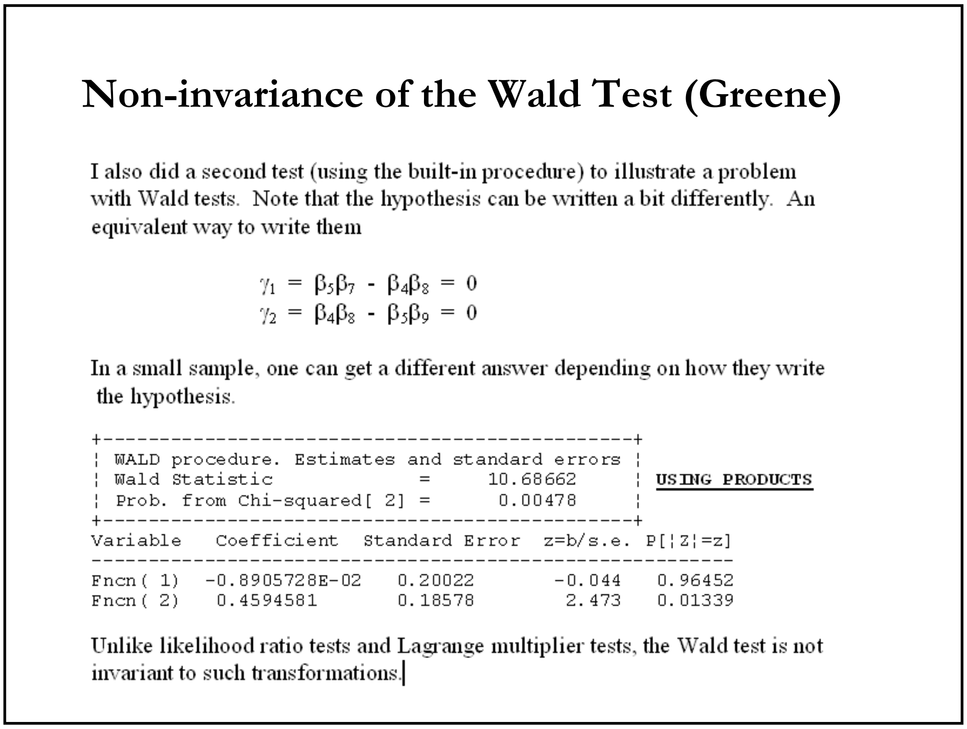
- In the example below, to make this a little simpler, Greene computed the 10 variable regression, then extracted the 5×1 subvector of the coefficient vector $\mathbf{c} = (b_4,b_5,b_7,b_8,b_9)$ and its associated part of the 10×10 covariance matrix. Then, Greene manipulated this smaller set of values.

Application of the Wald Statistic (Greene)

```
? Extract subvector and submatrix for the test
matrix;list ; c=[b(4)/b(5)/b(7)/b(8)/b(9)]$
matrix;list ; vc=[varb(4,4)/
                  varb(5,4),varb(5,5)/
                  varb(7,4),varb(7,5),varb(7,7)/
              varb(8,4),varb(8,5),varb(8,7),varb(8,8)/
              varb(9,4),varb(9,5),varb(9,7),varb(9,8),varb(9,9)]$
? Compute derivatives
calc  ;list
; g11=1/c(2); g12=-c(1)*g11*g11; g13=-1/c(4); g14=c(3)*g13*g13 ; g15=0
; g21=g11   ; g22=g12    ; g23=0  ; g24=c(5)/c(4)^2 ; g25=-1/c(4)$
? Move derivatives to matrix
matrix;list; dfdc=[g11,g12,g13,g14,g15 / g21,g22,g23,g24,g25]$
? Compute functions, then move to matrix and compute Wald statistic
calc;list ; f1=c(1)/c(2) - c(3)/c(4)
          ; f2=c(1)/c(2) - c(5)/c(4) $
matrix ; list; f = [f1/f2]$
matrix ; list; vf=dfdc * vc * dfdc' $
matrix ; list ; wald = f' * <vf> * f$
(This is all automated in the WALD command.)
```

Computations (Greene)

```
Matrix C  is    5 rows by    1 columns.
            1
   1   -0.2948  -0.2015   1.506   0.9995  -0.8179
Matrix VC       is    5 rows by    5 columns.
            1               2               3              4                5
   1    0.6655E-01  0.9479E-02 -0.4070E-01  0.4182E-01 -0.9888E-01
   2    0.9479E-02  0.5499E-02 -0.9155E-02  0.1355E-01 -0.2270E-01
   3   -0.4070E-01 -0.9155E-02  0.8848E-01 -0.2673E-01  0.3145E-01
   4    0.4182E-01  0.1355E-01 -0.2673E-01  0.7308E-01 -0.1038
   5   -0.9888E-01 -0.2270E-01  0.3145E-01 -0.1038      0.2134
 G11 =  -4.96184       G12 =  7.25755      G13= -1.00054    G14     =   1.50770   G15     =  0.000000
 G21 =  -4.96184       G22 =  7.25755      G23 = 0          G24     = -0.818753  G25     =  -1.00054
DFDC=[G11,G12,G13,G14,G15/G21,G22,G23,G24,G25]
Matrix DFDC     is    2 rows by    5 columns.
            1          2          3          4          5
   1    -4.962      7.258      -1.001      1.508      0.0000
   2    -4.962      7.258      0.0000     -0.8188      -1.001
F1= -0.442126E-01
F2=  2.28098
F=[F1/F2]
VF=DFDC*VC*DFDC'
Matrix VF       is    2 rows by    2 columns.
            1          2
   1    0.9804      0.7846
   2    0.7846      0.8648
WALD   =  22.65
```



Non-invariance of the Wald Test (Greene)

I also did a second test (using the built-in procedure) to illustrate a problem with Wald tests. Note that the hypothesis can be written a bit differently. An equivalent way to write them

$$\gamma_1 = \beta_5\beta_7 - \beta_4\beta_8 = 0$$
$$\gamma_2 = \beta_4\beta_8 - \beta_5\beta_9 = 0$$

In a small sample, one can get a different answer depending on how they write the hypothesis.

```
+-----------------------------------------------+
¦ WALD procedure. Estimates and standard errors ¦
¦ Wald Statistic             =     10.68662     ¦  USING PRODUCTS
¦ Prob. from Chi-squared[ 2] =      0.00478     ¦
+-----------------------------------------------+
Variable   Coefficient  Standard Error  z=b/s.e. P[¦Z¦=z]
---------------------------------------------------------
Fncn( 1)  -0.8905728E-02   0.20022       -0.044   0.96452
Fncn( 2)   0.4594581       0.18578        2.473   0.01339
```

Unlike likelihood ratio tests and Lagrange multiplier tests, the Wald test is not invariant to such transformations.