Colin Cameron: Brief Asymptotic Theory for 240A

For 240A we do not go in to great detail. **Key OLS results are in Section 1 and 4**. The theorems cited in sections 2 and 3 are those from Appendix A of Cameron and Trivedi (2005), *Microeconometrics: Methods and Applications*.

1. A Roadmap for the OLS Estimator

1.1. Estimator and Model Assumptions

Consider the OLS estimator $\widehat{\beta} = \left(\sum_{i=1}^{N} x_i^2\right)^{-1} \sum_{i=1}^{N} x_i y_i$, where there is just one regressor and no intercept. We want to find the properties of $\widehat{\beta}$ as sample size $N \to \infty$.

First, we need to make assumptions about the **data generating process** (dgp). We assume data are independent over i, the model is correctly specified and the error is well behaved:

$$\begin{aligned} y_i &= \beta x_i + u_i \\ u_i|x_i &\sim [0, \sigma^2], \text{ not necessarily normal.} \end{aligned}$$

Then $\mathrm{E}[\widehat{\beta}] = \beta$ and $\mathrm{V}[\widehat{\beta}] = \sigma^2 \left(\sum_{i=1}^{N} x_i^2\right)^{-1}$ for any N.

1.2. Consistency of $\widehat{\beta}$

Intuitively $\widehat{\beta}$ collapses on its mean of β as $N \to \infty$, since $\mathrm{V}[\widehat{\beta}] \to 0$ as $N \to \infty$. The formal term is that $\widehat{\beta}$ **converges in probability** to β, or that $\operatorname{plim} \widehat{\beta} = \beta$, or that $\widehat{\beta} \xrightarrow{p} \beta$. We say that $\widehat{\beta}$ is **consistent** for β.

The usual method of proof is not this simple as our toolkit works separately on each average, and then combines results. Given the dgp, $\widehat{\beta}$ can be written as

$$\widehat{\beta} = \beta + \frac{\frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2}.$$

The numerator can be shown to go to zero, by a **law of large numbers**, while the denominator goes to something nonzero. It follows that the ratio goes to zero.

Formally:

$$\widehat{\beta} \xrightarrow{p} \beta + \frac{\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta.$$

1.3. Asymptotic Normality of $\widehat{\beta}$

Intuitively, if a central limit theorem can be applied, then

$$\widehat{\beta} \overset{a}{\sim} \mathcal{N}[\mathrm{E}[\widehat{\beta}],\, \mathrm{V}[\widehat{\beta}]] \overset{a}{\sim} \mathcal{N}[\beta,\, \sigma^2(\textstyle\sum_{i=1}^{N} x_i^2)^{-1}],$$

where $\overset{a}{\sim}$ means is "**asymptotically distributed as**".

Again our toolkit works separately on each average, and then combines results. The method is to rescale by $\sqrt{N}$, to get something with nondegenerate distribution, and

$$\begin{aligned}\sqrt{N}(\widehat{\beta}-\beta) &= \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2\left(\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i^2\right)\right]}{\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i^2} \\ &\xrightarrow{d} \mathcal{N}\left[0, \sigma^2\left(\text{plim } \frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}\right].\end{aligned}$$

The key component is that $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i$ has a normal distribution as $N \to \infty$, by a **central limit theorem**.

Given this theoretical result we convert from $\sqrt{N}(\widehat{\beta}-\beta)$ to $\widehat{\beta}$ and drop the plim, giving

$$\widehat{\beta} \overset{a}{\sim} \mathcal{N}[\beta,\, \sigma^2(\textstyle\sum_{i=1}^{N} x_i^2)^{-1}].$$

2. Consistency

The keys are (1) convergence in probability; (2) law of large numbers for an average; (3) combining pieces.

2.1. Convergence in Probability

Consider the limit behavior of a **sequence of random variables** b_N as $N \to \infty$. This is a stochastic extension of a sequence of real numbers, such as $a_N = 2 + (3/N)$.

Examples include: (1) b_N is an estimator, say $\widehat{\theta}$; (2) b_N is a component of an estimator, such as $N^{-1}\sum_i x_i u_i$; (3) b_N is a test statistic.

Due to sampling randomness we can never be certain that a random sequence b_N, such as an estimator $\widehat{\theta}_N$, will be within a given small distance of its limit, even if the sample is infinitely large. But we can be almost certain. Different ways of expressing this almost certainty correspond to different types of convergence of a sequence of random variables to a limit. The one most used in econometrics is convergence in probability.

Recall that a sequence of nonstochastic real numbers $\{a_N\}$ converges to a if for any $\varepsilon > 0$, there exists $N^* = N^*(\varepsilon)$ such that for all $N > N^*$,

$$|a_N - a| < \varepsilon.$$

e.g. if $a_N = 2+3/N$, then the limit $a = 2$ since $|a_N - a| = |2+3/N-2| = |3/N| < \varepsilon$ for all $N > N^* = 3/\varepsilon$.

For a sequence of r.v.'s we cannot be certain that $|b_N - b| < \varepsilon$, even for large N, due to the randomness. Instead, we require that the probability of being within ε is arbitrarily close to one.

Thus $\{b_N\}$ **converges in probability** to b if

$$\lim_{N\to\infty} \Pr[|b_N - b| < \varepsilon] = 1,$$

for any $\varepsilon > 0$. A more formal definition is the following.

Definition A1: (*Convergence in Probability*) A sequence of random variables $\{b_N\}$ **converges in probability** to b if for any $\varepsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\varepsilon, \delta)$ such that for all $N > N^*$,

$$\Pr[|b_N - b| < \varepsilon] > 1 - \delta.$$

We write plim $b_N = b$, where plim is short-hand for **probability limit**, or $b_N \xrightarrow{p} b$. The limit b may be a constant or a random variable. The usual definition of convergence for a sequence of real variables is a special case of A1.

For **vector random variables** we can apply the theory for each element of $\mathbf{b}_N$. [Alternatively replace $|b_N - b|$ by the scalar $(\mathbf{b}_N - \mathbf{b})'(\mathbf{b}_N - \mathbf{b}) = (b_{1N} - b_1)^2 + \cdots + (b_{KN} - b_K)^2$ or by its square root $||\mathbf{b}_N - \mathbf{b}||$.]

Now consider $\{\mathbf{b}_N\}$ to be a sequence of parameter estimates $\widehat{\boldsymbol{\theta}}$.

Definition A2: (*Consistency*) An estimator $\widehat{\boldsymbol{\theta}}$ is **consistent** for $\boldsymbol{\theta}_0$ if

$$\operatorname{plim} \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0.$$

Unbiasedness $\not\Rightarrow$ consistency. Unbiasedness states $\mathrm{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}_0$. Unbiasedness permits variability around $\boldsymbol{\theta}_0$ that need not disappear as the sample size goes to infinity.

Consistency $\not\Rightarrow$ unbiasedness. e.g. add $1/N$ to an unbiased and consistent estimator - now biased but still consistent.

A useful property of plim is that it can apply to **transformations of random variables**.

Theorem A3: (*Probability Limit Continuity*). Let $\mathbf{b}_N$ be a finite-dimensional vector of random variables, and $g(\cdot)$ be a real-valued function continuous at a **constant vector point b**. Then

$$\mathbf{b}_N \xrightarrow{p} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{p} g(\mathbf{b}).$$

This theorem is often referred to as *Slutsky's Theorem*. We instead call Theorem A12 Slutsky's theorem.

Theorem A3 is one of the major reasons for the prevalence of asymptotic results versus finite sample results in econometrics. It states a very convenient property that does not hold for expectations.

For example, $\operatorname{plim}(a_N, b_N) = (a, b)$ implies $\operatorname{plim}(a_N b_N) = ab$, whereas $\mathrm{E}[a_N b_N]$ generally differs from $\mathrm{E}[a]\mathrm{E}[b]$.

Similarly $\operatorname{plim}[a_N/b_N] = a/b$ provided $b \neq 0$.

There are several ways to establish convergence in probability. The brute force method uses Definition A1, this is rarely done. It is often easier to establish alternative modes of convergence, notably convergence in mean square or use of Chebychev's inequality, which in turn imply convergence in probability. But it is usually easiest to use a law of large numbers.

2.2. Laws of Large Numbers

Laws of large numbers are theorems for **convergence in probability** (or *almost surely*) in the special case where the sequence $\{b_N\}$ is a **sample average**, i.e. $b_N = \bar{X}_N$ where

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Note that X_i here is general notation for a random variable, and in the regression context does not necessarily denote the regressor variables. For example $X_i = x_i u_i$.

Definition A7: (*Law of Large Numbers*) A **weak law of large numbers** (LLN) specifies conditions on the individual terms X_i in $\bar{X}_N$ under which

$$(\bar{X}_N - \mathrm{E}[\bar{X}_N]) \xrightarrow{p} 0.$$

For a **strong law of large numbers** the convergence is instead almost surely.

If a LLN can be applied then

$$\begin{array}{lll} \operatorname{plim} \bar{X}_N & = \lim \mathrm{E}[\bar{X}_N] & \text{in general} \\ & = \lim N^{-1} \sum_{i=1}^{N} \mathrm{E}[X_i] & \text{if } X_i \text{ independent over } i \\ & = \mu & \text{if } X_i \text{ iid.} \end{array}$$

The simplest laws of large numbers assume that X_i is iid.

Theorem A8: (*Kolmogorov LLN*) Let $\{X_i\}$ be iid (independent and identically distributed). If and only if $\mathrm{E}[X_i] = \mu$ exists and $\mathrm{E}[|X_i|] < \infty$, then $(\bar{X}_N - \mu) \xrightarrow{as} 0$.

The Kolmogorov LLN gives almost sure convergence. Usually convergence in probability is enough and we can use the weaker *Khinchine's Theorem*.

Theorem A8b: (*Khinchine's Theorem*) Let $\{X_i\}$ be iid (independent and identically distributed). If and only if $\mathrm{E}[X_i] = \mu$ exists, then $(\bar{X}_N - \mu) \xrightarrow{p} 0$.

There are other laws of large numbers. In particular, if X_i are independent but not identically distributed we can use the Markov LLN.

2.3. Consistency of OLS Estimator

Obtain the probability limit of $\widehat{\beta} = \beta + [\frac{1}{N}\sum_{i=1}^{N} x_i u_i]/[\frac{1}{N}\sum_{i=1}^{N} x_i^2]$, under simple random sampling with iid errors. Assume x_i iid with mean μ_x and u_i iid with mean 0.

As $x_i u_i$ are iid, apply Khinchine's Theorem yielding $N^{-1}\sum_i x_i u_i \xrightarrow{p} \mathrm{E}[xu] = \mathrm{E}[x] \times \mathrm{E}[u] = 0$.

As x_i^2 are iid, apply Khinchine's Theorem yielding $N^{-1}\sum_i x_i^2 \xrightarrow{p} \mathrm{E}[x^2]$ which we assume exists.

By Theorem A3 (Probability Limit Continuity) $\operatorname{plim}[a_N/b_N] = a/b$ if $b \neq 0$. Then

$$\operatorname{plim} \widehat{\beta} = \beta + \frac{\operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\operatorname{plim} \frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\mathrm{E}[x^2]} = \beta.$$

3. Asymptotic Normality

The keys are (1) convergence in distribution; (2) central limit theorem for an average; (3) combining pieces.

Given consistency, the estimator $\widehat{\theta}$ has a *degenerate distribution* that collapses on θ_0 as $N \to \infty$. So cannot do statistical inference. [Indeed there is no reason to do it if $N \to \infty$.] Need to magnify or *rescale* $\widehat{\theta}$ to obtain a random variable with *nondegenerate distribution* as $N \to \infty$.

3.1. Convergence in Distribution

Often the appropriate scale factor is $\sqrt{N}$, so consider $b_N = \sqrt{N}(\widehat{\theta} - \theta_0)$. b_N has an extremely complicated cumulative distribution function (cdf) F_N. But like any other function F_N it may have a limit function, where convergence is in the usual (nonstochastic) mathematical sense.

Definition A10: (*Convergence in Distribution*) A sequence of random variables $\{b_N\}$ is said to **converge in distribution** to a random variable b if

$$\lim_{N\to\infty} F_N = F,$$

at every continuity point of F, where F_N is the distribution of b_N, F is the distribution of b, and convergence is in the usual mathematical sense.

We write $b_N \xrightarrow{d} b$, and call F the **limit distribution** of $\{b_N\}$.

$b_N \xrightarrow{p} b$ implies $b_N \xrightarrow{d} b$.

In general, the reverse is not true. But if b is a constant then $b_N \xrightarrow{d} b$ implies $b_N \xrightarrow{p} b$.

To extend limit distribution to **vector random variables** simply define F_N and F to be the respective cdf's of vectors $\mathbf{b}_N$ and $\mathbf{b}$.

A useful property of convergence in distribution is that it can apply to **transformations of random variables**.

Theorem A11: (*Limit Distribution Continuity).* Let $\mathbf{b}_N$ be a finite-dimensional vector of random variables, and $g(\cdot)$ be a continuous real-valued function. Then

$$\mathbf{b}_N \xrightarrow{d} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{d} g(\mathbf{b}). \tag{3.1}$$

This result is also called the *Continuous Mapping Theorem*.

Theorem A12: (*Slutsky's Theorem*) If $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$, where a is a random variable and b is a constant, then

$$\begin{array}{ll} (i) & a_N + b_N \xrightarrow{d} a + b \\ (ii) & a_N b_N \xrightarrow{d} ab \\ (iii) & a_N / b_N \xrightarrow{d} a/b, \quad \text{provided } \Pr[b = 0] = 0. \end{array} \tag{3.2}$$

Theorem A12 (also called *Cramer's Theorem*) permits one to separately find the limit distribution of a_N and the probability limit of b_N, rather than having to consider the joint behavior of a_N and b_N. Result (ii) is especially useful and is sometimes called the **Product Rule**.

3.2. Central Limit Theorems

Central limit theorems give **convergence in distribution** when the sequence $\{b_N\}$ is a **sample average**. A CLT is much easier way to get the plim than brute force use of Definition A10.

By a LLN $\bar{X}_N$ has a degenerate distribution as it converges to a constant, $\lim \mathrm{E}[\bar{X}_N]$. So scale $(\bar{X}_N - \mathrm{E}[\bar{X}_N])$ by its standard deviation to construct a random variable with unit variance that will have a nondegenerate distribution.

Definition A13: (*Central Limit Theorem*) Let

$$Z_N = \frac{\bar{X}_N - \mathrm{E}[\bar{X}_N]}{\sqrt{\mathrm{V}[\bar{X}_N]}},$$

where $\bar{X}_N$ is a sample average. A **central limit theorem** (CLT) specifies the conditions on the individual terms X_i in $\bar{X}_N$ under which

$$Z_N \xrightarrow{d} \mathcal{N}[0, 1],$$

i.e. Z_N converges in distribution to a standard normal random variable.

Note that

$$\begin{array}{lll} Z_N & = (\bar{X}_N - \mathrm{E}[\bar{X}_N]) / \sqrt{\mathrm{V}[\bar{X}_N]} & \text{in general} \\ & = \sum_{i=1}^N (X_i - \mathrm{E}[X_i]) / \sqrt{\sum_{i=1}^N \mathrm{V}[X_i]} & \text{if } X_i \text{ independent over } i \\ & = \sqrt{N}(\bar{X}_N - \mu)/\sigma & \text{if } X_i \text{ iid.} \end{array}$$

If $\bar{X}_N$ satisfies a central limit theorem, then so too does $h(N)\bar{X}_N$ for functions $h(\cdot)$ such as $h(N) = \sqrt{N}$, since

$$Z_N = \frac{h(N)\bar{X}_N - \mathrm{E}[h(N)\bar{X}_N]}{\sqrt{\mathrm{V}[h(N)\bar{X}_N]}}.$$

Often we apply the CLT to the normalization $\sqrt{N}\bar{X}_N = N^{-1/2}\sum_{i=1}^{N} X_i$, since $\mathrm{V}[\sqrt{N}\bar{X}_N]$ is finite.

The simplest central limit theorem is the following.

Theorem A14: (*Lindeberg-Levy CLT*) Let $\{X_i\}$ be iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{V}[X_i] = \sigma^2$. Then $Z_N = \sqrt{N}(\bar{X}_N - \mu)/\sigma \xrightarrow{d} \mathcal{N}[0,1]$.

Lindberg-Levy is the CLT in introductory statistics. For the iid case the LLN required μ exists, while CLT also requires σ^2 exists.

There are other central limit theorems. In particular, if X_i are independent but not identically distributed we can use the Liapounov CLT.

3.3. Limit Distribution of OLS Estimator

Obtain the limit distribution of $\sqrt{N}(\widehat{\beta} - \beta) = [\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i]/[\frac{1}{N}\sum_{i=1}^{N} x_i^2]$, under simple random sampling with iid errors. Assume x_i iid with mean μ_x and second moment $\mathrm{E}[x^2]$, and assume u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are iid, with mean $\mathrm{E}[xu] = \mathrm{E}[x]\times\mathrm{E}[u] = 0$ and variance $\mathrm{V}[xu] = \mathrm{E}[(xu)^2] - (\mathrm{E}[xu])^2 = \mathrm{E}[x^2u^2] - 0 = \mathrm{E}[x^2]\mathrm{E}[u^2] = \sigma^2\mathrm{E}[x^2]$. Apply Lindeberg-Levy CLT yielding

$$\sqrt{N}\left(\frac{N^{-1}\sum_{i=1}^{N} x_i u_i - 0}{\sqrt{\sigma^2\mathrm{E}[x^2]}}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2\mathrm{E}[x^2]}} \xrightarrow{d} \mathcal{N}[0,1].$$

Using Slutsky's theorem that $a_N \times b_N \xrightarrow{d} a \times b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$), this implies that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2\mathrm{E}[x^2]}} \times \sqrt{\sigma^2\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}[0,1]\times\sqrt{\sigma^2\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}[0, \sigma^2\mathrm{E}[x^2]].$$

Then using Slutsky's theorem that $a_N/b_N \xrightarrow{d} a/b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$)

$$\begin{aligned}\sqrt{N}(\widehat{\beta}-\beta) &= \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \\ &\xrightarrow{d} \frac{\mathcal{N}\left[0,\sigma^2\mathrm{E}[x^2]\right]}{\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0,\sigma^2\mathrm{E}[x^2]\right]}{\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}\left[0,\sigma^2\left(\mathrm{E}[x^2]\right)^{-1}\right],\end{aligned}$$

where we use result from consistency proof that $\operatorname{plim} N^{-1}\sum_{i=1}^{N} x_i^2 = \mathrm{E}[x^2]$.

3.4. Asymptotic Distribution of OLS Estimator

From consistency we have that $\widehat{\beta}$ has a degenerate distribution with all mass at β, while $\sqrt{N}(\widehat{\beta}-\beta)$ has a limit normal distribution. For formal asymptotic theory, such as deriving hypothesis tests, we work with this limit distribution. But for exposition it is convenient to think of the distribution of $\widehat{\beta}$ rather than $\sqrt{N}(\widehat{\beta}-\beta)$. We do this by introducing the artifice of "asymptotic distribution".

Specifically we consider N large but not infinite, and drop the probability limit in the preceding result, so that

$$\sqrt{N}(\widehat{\beta}-\beta) \sim \mathcal{N}\left[0,\sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

It follows that the **asymptotic distribution** of $\widehat{\beta}$ is

$$\widehat{\beta} \overset{a}{\sim} \mathcal{N}\left[\beta,\sigma^2\left(\sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

Note that this is exactly the same result as we would have got if $y_i = \beta x_i + u_i$ with $u_i \sim \mathcal{N}[0,\sigma^2]$.

4. OLS Estimator with Matrix Algebra

Now consider $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, so

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Note that the $k \times k$ matrix $\mathbf{X}'\mathbf{X} = \sum_i \mathbf{x}_i\mathbf{x}_i'$ is a sum, where $\mathbf{x}_i$ is a $k \times 1$ vector of regressors for the i^{th} observation. Dividing by N gives an average.

4.1. Multivariate CLT

Now we need to work with vectors. Two useful results follow.

Definition A16a: (*Multivariate Central Limit Theorem*) Let $\boldsymbol{\mu}_N = \mathrm{E}[\bar{\mathbf{X}}_N]$ and $\mathbf{V}_N = \mathrm{V}[\bar{\mathbf{X}}_N]$. A **multivariate central limit theorem** (CLT) specifies the conditions on the individual terms $\mathbf{X}_i$ in $\bar{\mathbf{X}}_N$ under which

$$\mathbf{V}_N^{-1/2}(\bar{\mathbf{X}}_N - \boldsymbol{\mu}_N) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}].$$

Note that if we can apply a CLT to $\bar{\mathbf{X}}_N$ we can also apply it to $N^{-1/2}\bar{\mathbf{X}}_N$.

Theorem A17: (*Limit Normal Product Rule*) If a vector $\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\boldsymbol{\mu}, \mathbf{A}]$ and a matrix $\mathbf{H}_N \xrightarrow{p} \mathbf{H}$, where $\mathbf{H}$ is positive definite, then

$$\mathbf{H}_N\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\mathbf{A}\mathbf{H}'].$$

4.2. Consistency of OLS

To prove consistency we rewrite this as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{X}'\mathbf{u}.$$

The reason for renormalization in the right-hand side is that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ is an average that converges in probability to a finite nonzero matrix if $\mathbf{x}_i$ satisfies assumptions that permit a LLN to be applied to $\mathbf{x}_i\mathbf{x}_i'$.

Then

$$\operatorname{plim} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u}\right),$$

using Slutsky's Theorem (Theorem A.3). The OLS estimator is therefore **consistent** for $\boldsymbol{\beta}$ (i.e., $\operatorname{plim} \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \boldsymbol{\beta}$) if

$$\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}.$$

If a law of LLN can be applied to the average $N^{-1}\mathbf{X}'\mathbf{u} = N^{-1}\sum_i \mathbf{x}_i u_i$ then a necessary condition for this to hold is that $\mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$. The fundamental condition for consistency of OLS is that $\mathrm{E}[u_i|\mathbf{x}_i] = \mathbf{0}$ so that $\mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$.

4.3. Limit Distribution of OLS

Given consistency, the limit distribution of $\widehat{\boldsymbol{\beta}}$ is degenerate with all the mass at $\boldsymbol{\beta}$. To obtain a limit distribution we scale $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ up by a multiple $\sqrt{N}$, so

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1/2}\mathbf{X}'\mathbf{u}.$$

We know plim $N^{-1}\mathbf{X}'\mathbf{X}$ exists and is finite and nonzero from the proof of consistency. For iid errors, $\text{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $\text{V}[\mathbf{X}'\mathbf{u}|\mathbf{X}] = \text{E}[\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}'|\mathbf{X}] = \sigma^2\mathbf{X}'\mathbf{X}$ we assume that a CLT can be applied to $\mathbf{b}_N = N^{-1/2}\mathbf{X}'\mathbf{u}$ to yield

$$\begin{aligned}
&[N^{-1}\mathbf{X}'\mathbf{X}]^{-1/2} \times N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}], \text{ so by Theorem A17} \\
&N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2 \left(\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X}\right)].
\end{aligned}$$

Then by Theorem A17 (the limit normal product rule)

$$\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{d} (\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X})^{-1} \times \mathcal{N}[\mathbf{0}, \sigma^2 (\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X})] \\
&\xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2 (\text{plim}\, N^{-1}\mathbf{X}'\mathbf{X})^{-1}].
\end{aligned}$$

4.4. Asymptotic Distribution of OLS

Then dropping the limits

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}[\mathbf{0}, \sigma^2 \left(N^{-1}\mathbf{X}'\mathbf{X}\right)],$$

so

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

The asymptotic variance matrix is

$$\text{V}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

and is consistently estimated by the estimated variance matrix

$$\widehat{\text{V}}[\widehat{\boldsymbol{\beta}}] = s^2(\mathbf{X}'\mathbf{X})^{-1},$$

where s^2 is consistent for σ^2. For example, $s^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/(N-k)$ or $s^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/N$.

4.5. OLS with Heteroskedastic Errors

What if the errors are heteroskedastic? If $\mathrm{E}[\mathbf{uu}'|\mathbf{X}] = \mathbf{\Sigma} = \mathrm{Diag}[\sigma_i^2]$ then $\mathrm{V}[\mathbf{X}'\mathbf{u}|\mathbf{X}] = \mathrm{E}[\mathbf{X}'\mathbf{uu}'\mathbf{X}'|\mathbf{X}] = \mathbf{X}'\mathbf{\Sigma}\mathbf{X} = \sum_{i=1}^{N} \sigma_i^2 \mathbf{x}_i\mathbf{x}_i'$. A Multivariate CLT gives

$$[N^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}]^{-1/2} \times N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}], \text{ so}$$
$$N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \operatorname{plim} N^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}],$$

leading to

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{d} (\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X})^{-1} \times \mathcal{N}[\mathbf{0}, \operatorname{plim} N^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}] \\ &\xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2 (\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X})^{-1} \times \operatorname{plim} N^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X} \times (\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X})^{-1}].\end{aligned}$$

Then dropping the limits etcetera

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}].$$

The asymptotic variance matrix is

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

White (1980) showed that we can use the estimated asymptotic variance matrix

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}(\sum\nolimits_{i=1}^{N} \widehat{u}_i{}^2\mathbf{x}_i\mathbf{x}_i')(\mathbf{X}'\mathbf{X})^{-1},$$

where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ is the OLS residual.

Why does this work? There are N variances σ_i^2 and only N observations, so we cannot consistently estimate $\mathbf{\Sigma}$ by $\widehat{\mathbf{\Sigma}} = \mathrm{Diag}[\widehat{u}_i{}^2]$. But this is not necessary! We just need to consistently estimate the $k \times k$ matrix $\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X} = \operatorname{plim} N^{-1}\sum_{i=1}^{N} \sigma_i^2\mathbf{x}_i\mathbf{x}_i'$.
Since $\mathrm{E}[u_i^2\mathbf{x}_i\mathbf{x}_i'|\mathbf{x}_i] = \sigma_i^2\mathbf{x}_i\mathbf{x}_i'$, by a LLN, if the error u_i was observed, we could use

$$\operatorname{plim} N^{-1}\sum\nolimits_{i=1}^{N} u_i^2\mathbf{x}_i\mathbf{x}_i' = \operatorname{plim} N^{-1}\sum\nolimits_{i=1}^{N} \sigma_i^2\mathbf{x}_i\mathbf{x}_i'.$$

The error is not observed, so we instead use the residual. Formally, since $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, $\widehat{u}_i^2 - u_i^2 \xrightarrow{p} 0$ so replacing u_i^2 by $\widehat{u}_i^2$ makes no difference asymptotically. This gives White's estimator.

This is a fundamental result. Without specifying a model for heteroskedasticity we can do OLS and get heteroskedastic robust standard errors. This generalizes to other failures of $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, notably serially correlated errors (use Newey-West) and clustered errors. And it generalizes to estimators other than OLS (e.g. IV, MLE).

5. Optional Extra: Different Sampling Schemes

The key for consistency is obtaining the probability of the two averages (of $x_i u_i$ and of x_i^2), by use of laws of large numbers (LLN). And for asymptotic normality the key is the limit distribution of the average of $x_i u_i$, obtained by a central limit theorem (CLT).

Different assumptions about the stochastic properties of x_i and u_i lead to different properties of x_i^2 and $x_i u_i$ and hence different LLN and CLT.

5.1. Sampling Schemes

For the data different **sampling schemes** assumptions include:

1. Simple Random Sampling (SRS).

   SRS is when we randomly draw (y_i, x_i) from the population. Then x_i are iid. So x_i^2 are iid, and $x_i u_i$ are iid if the errors u_i are iid.

2. Fixed regressors.
   This occurs in an experiment where we fix the x_i and observe the resulting random y_i. Given x_i fixed and u_i iid it follows that $x_i u_i$ are inid (even if u_i are iid), while x_i^2 are nonstochastic.

3. Exogenous Stratified Sampling
   This occurs when we oversample some values of x and undersample others. Then x_i are inid, so $x_i u_i$ are inid (even if u_i are iid) and x_i^2 are inid.

The simplest results assume simple random sampling, as we have done.

5.2. LLN and CLT for inid Data

Suppose rather than X_i iid $[\mu, \sigma^2]$ we have have X_i inid $[\mu, \sigma^2]$. Then we often use the following LLN and CLT.

Theorem A9: (*Markov LLN*) Let $\{X_i\}$ be inid (independent but not identically distributed) with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2$. If $\sum_{i=1}^{\infty} (\mathrm{E}[|X_i - \mu_i|^{1+\delta}]/i^{1+\delta}) < \infty$, for some $\delta > 0$, then $(\bar{X}_N - N^{-1} \sum_{i=1}^{N} \mathrm{E}[X_i]) \xrightarrow{as} 0$.

Compared to Kolmogorov LLN, the Markov LLN allows nonidentical distribution, at expense of require existence of an absolute moment beyond the first.

The rest of the side-condition is likely to hold with cross-section data. e.g. if set $\delta = 1$, then need variance plus $\sum_{i=1}^{\infty}(\sigma_i^2/i^2) < \infty$ which happens if σ_i^2 is bounded.

Theorem A15: (*Liapounov CLT*) Let $\{X_i\}$ be independent with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2$. If $\lim \left(\sum_{i=1}^{N} \mathrm{E}[|X_i - \mu_i|^{2+\delta}]\right) / \left(\sum_{i=1}^{N} \sigma_i^2\right)^{(2+\delta)/2} = 0$, for some choice of $\delta > 0$, then $Z_N = \sum_{i=1}^{N}(X_i - \mu_i)/\sqrt{\sum_{i=1}^{N}\sigma_i^2} \xrightarrow{d} \mathcal{N}[0,1]$.

Compared to Lindberg-Levy, the Liapounov CLT additionally requires existence of an absolute moment of higher order than two.

5.3. Consistency of OLS Estimator

Obtain probability limit of $\widehat{\beta} = \beta + [\frac{1}{N}\sum_{i=1}^{N} x_i u_i]/[\frac{1}{N}\sum_{i=1}^{N} x_i^2]$.

5.3.1. Simple Random Sampling (SRS) with iid errors

Assume x_i iid with mean μ_x and u_i iid with mean 0.

As $x_i u_i$ are iid, apply Khinchine's Theorem yielding $N^{-1}\sum_i x_i u_i \xrightarrow{p} \mathrm{E}[xu] = \mathrm{E}[x]\times\mathrm{E}[u] = 0$.

As x_i^2 are iid, apply Khinchine's Theorem yielding $N^{-1}\sum_i x_i^2 \xrightarrow{p} \mathrm{E}[x^2]$ which we assume exists.

By Theorem A3 (Probability Limit Continuity) $\mathrm{plim}[a_N/b_N] = a/b$ if $b \neq 0$. Then

$$\mathrm{plim}\,\widehat{\beta} = \beta + \frac{\mathrm{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\mathrm{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\mathrm{E}[x^2]} = \beta.$$

5.3.2. Fixed Regressors with iid errors

Assume x_i fixed and that u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are inid with mean $\mathrm{E}[x_i u_i] = x_i \mathrm{E}[u_i] = 0$ and variance $\mathrm{V}[x_i u_i] = x_i^2\sigma^2$. Apply Markov LLN yielding $N^{-1}\sum_i x_i u_i - N^{-1}\sum_i \mathrm{E}[x_i u_i] \xrightarrow{p} 0$, so $N^{-1}\sum_i x_i u_i \xrightarrow{p} 0$. The side-condition with $\delta = 1$ is $\sum_{i=1}^{\infty} x_i^2\sigma^2/i^2$ which is satisfied if x_i is bounded.

We also assume $\lim N^{-1}\sum_i x_i^2$ exists.

By Theorem A3 (Probability Limit Continuity) $\mathrm{plim}[a_N/b_N] = a/b$ if $b \neq 0$. Then

$$\mathrm{plim}\,\widehat{\beta} = \beta + \frac{\mathrm{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\lim\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\lim\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta.$$

5.3.3. Exogenous Stratified Sampling with iid errors

Assume x_i inid with mean $\mathrm{E}[x_i]$ and variance $\mathrm{V}[x_i]$ and u_i iid with mean 0.

Now $x_i u_i$ are inid with mean $\mathrm{E}[x_i u_i] = \mathrm{E}[x_i]\mathrm{E}[u_i] = 0$ and variance $\mathrm{V}[x_i u_i] = \mathrm{E}[x_i^2]\sigma^2$, so need Markov LLN. This yields $N^{-1}\sum_i x_i u_i \xrightarrow{p} 0$, with the side-condition satisfied if $\mathrm{E}[x_i^2]$ is bounded.

And x_i^2 are inid, so need Markov LLN with side-condition that requires e.g. existence and boundedness of $\mathrm{E}[x_i^4]$.

Combining again get $\operatorname{plim}\widehat{\beta} = \beta$.

5.4. Limit Distribution of OLS Estimator

Obtain limit distribution of $\sqrt{N}(\widehat{\beta} - \beta) = [\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i]/[\frac{1}{N}\sum_{i=1}^N x_i^2]$.

5.4.1. Simple Random Sampling (SRS) with iid errors

Assume x_i iid with mean μ_x and second moment $\mathrm{E}[x^2]$, and assume u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are iid, with mean $\mathrm{E}[xu] = \mathrm{E}[x]\times\mathrm{E}[u] = 0$ and variance $\mathrm{V}[xu] = \mathrm{E}[(xu)^2] - (\mathrm{E}[xu])^2 = \mathrm{E}[x^2u^2] - 0 = \mathrm{E}[x^2]\mathrm{E}[u^2] = \sigma^2\mathrm{E}[x^2]$. Apply Lindeberg-Levy CLT yielding

$$\sqrt{N}\left(\frac{N^{-1}\sum_{i=1}^N x_i u_i - 0}{\sqrt{\sigma^2\mathrm{E}[x^2]}}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i}{\sqrt{\sigma^2\mathrm{E}[x^2]}} \xrightarrow{d} \mathcal{N}[0,1].$$

Using Slutsky's theorem that $a_N \times b_N \xrightarrow{d} a \times b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$), this implies that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i}{\sqrt{\sigma^2\mathrm{E}[x^2]}} \times \sqrt{\sigma^2\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}[0,1]\times\sqrt{\sigma^2\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}[0,\sigma^2\mathrm{E}[x^2]].$$

Then using Slutsky's theorem that $a_N/b_N \xrightarrow{d} a/b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$)

$$\begin{aligned}\sqrt{N}(\widehat{\beta} - \beta) &= \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i}{\frac{1}{N}\sum_{i=1}^N x_i^2} \\ &\xrightarrow{d} \frac{\mathcal{N}\left[0,\sigma^2\mathrm{E}[x^2]\right]}{\operatorname{plim}\frac{1}{N}\sum_{i=1}^N x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0,\sigma^2\mathrm{E}[x^2]\right]}{\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}\left[0,\sigma^2\left(\mathrm{E}[x^2]\right)^{-1}\right],\end{aligned}$$

where we use result from consistency proof that $\operatorname{plim} N^{-1}\sum_{i=1}^N x_i^2 = \mathrm{E}[x^2]$.

5.4.2. Fixed Regressors with iid errors

Assume x_i fixed and and u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are inid with mean 0 and variance $\mathrm{V}[x_i u_i] = x_i^2 \sigma^2$. Apply Liapounov LLN yielding

$$\sqrt{N}\left(\frac{N^{-1}\sum_{i=1}^{N} x_i u_i - 0}{\sqrt{\lim N^{-1}\sum_{i=1}^{N} x_i^2\sigma^2}}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2 \lim N^{-1}\sum_{i=1}^{N} x_i^2}} \xrightarrow{d} \mathcal{N}[0,1].$$

Using Slutsky's theorem that $a_N \times b_N \xrightarrow{d} a \times b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$), this implies

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i &= \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2 \lim N^{-1}\sum_{i=1}^{N} x_i^2}} \times \sqrt{\sigma^2 N^{-1}\sum_{i=1}^{N} x_i^2} \\ &= \mathcal{N}[0,1] \times \sqrt{\sigma^2 \lim N^{-1}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \mathcal{N}[0, \sigma^2 \lim \frac{1}{N}\sum_{i=1}^{N} x_i^2].\end{aligned}$$

Then using Slutsky's theorem that $a_N/b_N \xrightarrow{d} a/b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$)

$$\sqrt{N}(\widehat{\beta}-\beta) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2 \lim \frac{1}{N}\sum_{i=1}^{N} x_i^2\right]}{\lim \frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \mathcal{N}\left[0, \sigma^2\left(\lim \frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

5.4.3. Exogenous Stratified Sampling with iid errors

Assume x_i inid with mean $\mathrm{E}[x_i]$ and variance $\mathrm{V}[x_i]$ and u_i iid with mean 0.

Similar to fixed regressors will need to use Liapounov CLT. We will get

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left[0, \sigma^2\left(\mathrm{plim}\, \frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

6. Simulation Exercise

6.1. Data Generating Process

Suppose y_i are iid $\chi^2(1)$, which has mean 1 and variance 2.

Then $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ has $\mathrm{E}[\bar{y}] = 1$, $\mathrm{V}[\bar{y}] = \mathrm{V}[\bar{y}]/N = 2/N$, and using the result that the sum of N independent $\chi^2(1)$ is $\chi^2(N)$, we know that $\bar{y} \sim \chi^2(N)/N$.

6.2. Distribution of $\bar{y}$

From a CLT we know that as $N \to \infty$

$$\bar{y} \sim \mathcal{N}[1, 2/N].$$

Question: how well does this apply in finite samples? Answer: do a simulation:

- For the s^{th} of S times, generate N observations from $\chi^2(1)$ and calculate $\bar{y}_s$.

Then does the density of the S generated $\bar{y}_s$ look like the density of a standard normal with mean 1 and variance $2/N$? We can do this in several ways: comparing key moments, or comparing key percentiles, or compare a histogram or (kernel) density estimate to the normal.

6.3. Distribution of t-test

The t-test statistic of $H_0 : \mu = 1$ is

$$t = \frac{\bar{y} - 1}{s_{\bar{y}}},$$

where $s_{\bar{y}}^2 = s_y^2/N$ and $s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N} y_i$.

From asymptotic theory we know that $t \sim \mathcal{N}[0, 1]$ as $N \to \infty$. As a finite sample correction we will instead suppose it is $t(N-1)$ degrees distributed, which is the exact result if y_i were normal rather than chi-square.

Question: how well does this apply in finite samples? Answer: do a simulation:

- For the s^{th} of S times, generate N observations from $\chi^2(1)$ and calculate $\bar{y}_s$, $s_{\bar{y}_s}$ and hence $t_s = \frac{\bar{y}_s - 1}{s_{\bar{y}_s}}$.

Then does the density of the S generated t_s look like a $t(N-1)$ density? Again we can test the similarity using moments, percentiles or kernel density estimate.

6.4. Size of t-test

For the t-test it is **behavior in the tails** that really matter. In addition to view the 2.5 and 97.5 percentiles, we investigate the **size** of the t-test. Recall

$$\text{Size } = \Pr[\text{reject } H_0 : \mu = 1 | H_0 \text{ true}]$$

The **nominal size** of the test is the size that we think we are testing at. Assume that $t \sim T(N-1)$ distributed under H_0. Then for two-tailed testing at 5 percent

$$\text{Nominal size} = 0.05 = \Pr[|t| > t_{.025;N-1} | t \sim T(N-1)].$$

But we do not know that $t \sim T(N-1)$ under H_0, and in small samples it is not this. The **true size** of the test is the actual size that we are testing at.

$$\text{True size} = \Pr[|t| > t_{.025;N-1} | \text{true distribution of } t \text{ under } H_0].$$

To obtain this unknown true distribution we do a simulation. For the s^{th} of S times, generate N observations from $\chi^2(1)$, calculate $t_s = \frac{\bar{y}_s - 1}{s_{\bar{y}_s}}$, and record whether $|t_s| > t_{.025;N-1}$, i.e. whether we reject H_0.

Then the simulation estimate of true size is the proportion of rejections:

$$\text{True size} = (\#\text{ simulations with } |t| > t_{.025;N-1})/(\#\text{ simulations}).$$