Chapter 22: Asymptotic Properties of OLS Estimator and Test Statistics

22.1 THE I.I.D. CASE

In this section, we consider a random design regression model where the observations (x_i, ε_i) are independent and identically distributed (i.i.d.). The model is given by:

$$y_i = \beta^T x_i + \varepsilon_i, \quad x_i = (x_{i1}, \ldots, x_{iK})^T$$

where y_i is the dependent variable, x_i is a $K \times 1$ vector of regressors, β is a $K \times 1$ vector of coefficients, and ε_i is the error term.

Assumption B

We assume that (x_i, ε_i) are i.i.d. and that:

1. $E(x_i) = E(\varepsilon_i) = 0$ and $E[|x_{ji}\varepsilon_i|] < \infty$, for $j = 1, \ldots, K$, and for some positive definite $K \times K$ matrix M:

$$E[x_i x_i^T] = M$$

2. For some finite positive definite $K \times K$ matrix Ω:

$$E[x_i x_i^T \varepsilon_i^2] = \Omega$$

We also consider the **homoskedastic** special case where $E(\varepsilon_i^2 | x_i) = \sigma^2$ for some σ^2, in which case:

$$\Omega = \sigma^2 M$$

Note: Assumptions A1 and A2 from the i.i.d framework imply conditions B1 and B2, respectively.

In B1, we assume the **unconditional moment condition**:

$$\text{cov}(x_i, \varepsilon_i) = E(x_i \varepsilon_i) = 0$$

This condition is sufficient for the asymptotic properties but is insufficient for the proper interpretation of the model as a regression since we would get the Best Linear Predictor (Chapter 7) and it does not ensure that the OLS estimator is unbiased.

We might assume the stronger **conditional moment condition**:

$$E[\epsilon_i | x_i] = 0$$

$$E[\epsilon_i^2 | x_i] = v(x_i)$$

$$\Omega = E[v(x_i) x_i x_i^T]$$

Example 22.1

$x_i = \sin(\theta_i)$ and $\varepsilon_i = \cos(\theta_i)$, where $\theta_i \sim [0, 2\pi]$. We have:

$$E(x_i \varepsilon_i) = 0$$

But knowing x_i tells you ε_i, and it is not always zero. However, $E(\varepsilon_i|x_i) = 0$.

The conditional moment condition also ensures that the OLS estimator is unbiased because:

$$\hat{\beta} = \beta + (X^T X)^{-1}(X^T \varepsilon) = \beta + \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \sum_{i=1}^{n} x_i \varepsilon_i$$

$$E[\hat{\beta}|X] = \beta + \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \sum_{i=1}^{n} x_i E[\varepsilon_i|x_i] = \beta$$

We assume B1 and B2 and are concerned with large sample properties.

Theorem 22.1

Suppose that B1 holds. Then we have

$$\hat{\beta} \xrightarrow{p} \beta$$

Proof:

We have

$$\hat{\beta} - \beta = \left(\frac{X^T X}{n}\right)^{-1} \left(\frac{X^T \varepsilon}{n}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

By the Weak Law of Large Numbers, we know:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \xrightarrow{p} E(x_i x_i^T) = M$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} E(x_i \varepsilon_i) = 0$$

Then the result follows by applying the Law of Large Numbers to both terms and then applying Slutsky's theorem.

Theorem 22.2

Suppose that B1 and B2 hold. Then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, M^{-1}\Omega M^{-1})$$

Proof: We have

$$\frac{X^T X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \xrightarrow{p} M$$

using LLN element by element. Then for any $c \in \mathbb{R}^K$:

$$\frac{c^T X^T \varepsilon}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} c^T x_i \varepsilon_i \xrightarrow{d} N(0, c^T \Omega c)$$

by the CLT for i.i.d. random variables. Then apply the Crámer-Wald, the Mann-Wald, and the Slutsky Theorems.

In the special case of homoskedasticity, we have:

$$M^{-1} \Omega M^{-1} = M^{-1} \sigma^2 M M^{-1} = \sigma^2 M^{-1}$$

Theorem 22.3

Suppose that B1 and B2 hold and let $\sigma^2 = E[\varepsilon_i^2] < \infty$. Then

$$s_*^2 \xrightarrow{p} \sigma^2$$

where

$$s_*^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - K} = \frac{1}{n-K} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-K} \varepsilon^T M_X \varepsilon$$

Proof:

We have

$$\begin{aligned} s_*^2 &= \frac{1}{n-K} \varepsilon^T M_X \varepsilon \\ &= \frac{1}{n-K} [\varepsilon^T \varepsilon - \varepsilon^T X (X^T X)^{-1} X^T \varepsilon] \\ &= \left(\frac{n}{n-K} \right) \left(\frac{\varepsilon^T \varepsilon}{n} - \frac{\varepsilon^T X}{n} \left(\frac{X^T X}{n} \right)^{-1} \frac{X^T \varepsilon}{n} \right) \\ &\xrightarrow{p} \sigma^2 \end{aligned}$$

This follows by the Law of Large Numbers applied to $X^T \varepsilon / n$, $\varepsilon^T \varepsilon / n$, and $X^T X / n$.

Intuition: This theorem shows that the unbiased estimator of the error variance, s_*^2, is a consistent estimator of the true error variance, σ^2.

Theorem 22.4

Suppose B1 and B2 and (22.1) hold. Then, under H_0, as $n \to \infty$, we have:

$$t \xrightarrow{d} N(0, 1)$$

and

$$qF \xrightarrow{d} \chi^2(q)$$

where:

$$t = \frac{\sqrt{n} c^T \hat{\beta}}{\sqrt{s_*^2 c^T (\frac{X^T X}{n})^{-1} c}}$$

$$F = n(R\hat{\beta} - r)^T \left[s_*^2 R \left(\frac{X^T X}{n} \right)^{-1} R^T \right]^{-1} (R\hat{\beta} - r)/q$$

Proof:

It follows from above that under H_0:

$$\frac{c^T\hat{\beta} - \gamma}{\sqrt{\text{Var}(c^T\hat{\beta} - \gamma)}} = \frac{\sqrt{n}c^T(\hat{\beta} - \beta)}{\sigma\sqrt{c^T(\frac{X^TX}{n})^{-1}c}} \xrightarrow{d} N(0,1)$$

Then, using Slutsky and Mann-Wald, we know that:

$$\frac{\sqrt{n}c^T(\hat{\beta} - \beta)}{s_*\sqrt{c^T(\frac{X^TX}{n})^{-1}c}} = \frac{\sqrt{n}c^T(\hat{\beta} - \gamma)}{\sigma\sqrt{c^T(\frac{X^TX}{n})^{-1}c}} \times \frac{\sigma}{s_*} \xrightarrow{d} N(0,1)$$

Similarly, for the F-test, it follows from (20.15) that, by applying the delta method and Slutsky's theorem:

$$W, LR, LM \xrightarrow{d} \chi^2(q)$$

If the homoskedasticity condition fails, then $t \xrightarrow{d} N(0, v)$ for some $v \neq 1$ and the testing strategy would use the wrong critical values. Likewise, the Wald statistic would not have the claimed chi-squared distribution.

Estimating Ω

In general, we work with B1 and B2 and we don't assume homoskedasticity. Therefore, the asymptotic variance of $\hat{\beta}$ involves both M and Ω, where:

$$\Omega = E(x_i x_i^T \epsilon_i^2)$$

We estimate Ω with:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \hat{\varepsilon}_i^2$$

Then the matrix:

$$\hat{H} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \right)^{-1} \hat{\Omega} \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \right)^{-1} = (\hat{H}_{j,k})_{j,k=1}^{K}$$

contains all the estimators of $\text{var}(\sqrt{n}\hat{\beta}_k)$ and $\text{cov}(\sqrt{n}\hat{\beta}_j, \sqrt{n}\hat{\beta}_k)$. In this case, we consider robust test statistics (White, 1980):

$$t_R = \frac{c^T\hat{\beta}}{\sqrt{c^T(X^TX)^{-1}\hat{\Omega}(X^TX)^{-1}c}}$$

$$W_R = (R\hat{\beta} - r)^T[R(X^TX)^{-1}\hat{\Omega}(X^TX)^{-1}R^T]^{-1}(R\hat{\beta} - r)$$

where

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \hat{\varepsilon}_i^2 = \frac{1}{n} X^T S X$$

where S is the $n \times n$ diagonal matrix with typical element $\hat{\varepsilon}_i^2$.

Theorem 22.5

Suppose that B1 and B2 hold. Suppose also that $E|x_{ij}x_{ik}x_{il}x_{ir}| < \infty$, $E|x_{ij}x_{ik}x_{il}\varepsilon_i| < \infty$ and $E(x_{ij}x_{ik}\varepsilon_i x_{il}) = 0$. Then

$$t_R \xrightarrow{d} N(0,1)$$

$$W_R \xrightarrow{d} \chi^2(q)$$

Proof:

We have:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \varepsilon_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \varepsilon_i x_i^T (X^T X)^{-1} X^T \varepsilon$$
$$+ \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T x_i^T (X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1} x_i$$

where $\tilde{\varepsilon_i} = \varepsilon_i - x_i^T (X^T X)^{-1} X^T \varepsilon$. We have:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \epsilon_i x_i^T (X^T X)^{-1} X^T \epsilon = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \epsilon_i x_i^T (X^T X/n)^{-1} (X^T \epsilon/n) \xrightarrow{p} 0$$

since

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} x_{ik} \varepsilon_i x_{il} \xrightarrow{p} 0$$

while

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T x_i^T \left(\frac{X^T X}{n}\right)^{-1} \frac{X^T \varepsilon \varepsilon^T X}{n} \left(\frac{X^T X}{n}\right)^{-1} x_i \xrightarrow{p} 0$$

because $X^T X/n \xrightarrow{p} M$ and $X^T \varepsilon / \sqrt{n} \Rightarrow N(0, \Omega)$, while for any j,k,l,r:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} x_{ik} x_{il} x_{ir} \xrightarrow{p} E[x_{ij} x_{ik} x_{il} x_{ir}]$$

It follows that $\hat{\Omega} \xrightarrow{p} \Omega$.

22.2 THE NON-I.I.D. CASE

We next consider the "fixed" design setting of **Assumption A** and make general assumptions on the matrix X. These conditions work in more general sampling schemes including trending variables.

Theorem 22.6

Suppose Assumptions A1-A2 hold and with probability one:

$$\lambda_{max}(X^T\Sigma X)\lambda_{min}^{-2}(X^TX) \to 0, \quad \text{as } n \to \infty$$

Then, $\hat{\beta} \xrightarrow{p} \beta$. **Proof:** We show that $c^T\hat{\beta} \xrightarrow{p} c^T\beta$ for all vectors c. First, with probability 1:

$$E[c^T\hat{\beta}|X] = c^T\beta$$

$$\text{var}(c^T\hat{\beta}|X) = c^T(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}c$$

We have:

$$\begin{aligned} c^T(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}c &= c^Tc\frac{c^T(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}c}{c^Tc} \\ &\leq c^Tc \cdot \max_{u\in\mathbb{R}^k} \frac{u^T(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}u}{u^Tu} \\ &\leq c^Tc \cdot \lambda_{max}((X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}) \\ &= c^Tc \cdot \lambda_{max}((X^TX)^{-1})\lambda_{max}(X^T\Sigma X) \\ &= c^Tc \cdot \frac{\lambda_{max}(X^T\Sigma X)}{\lambda_{min}^2(X^TX)} \end{aligned}$$

and provided $\lambda_{max}(X^T\Sigma X)\lambda_{min}^{-2}(X^TX) \to 0$, we have $\text{var}(c^T\hat{\beta}) \to 0$.

When $\Sigma = \sigma^2 I_n$, it suffices that $\lambda_{min}(X^TX) \to \infty$.

If we do have a random design, then the conditions and conclusion should be interpreted as holding with probability one in the conditional distribution given X. Deterministic design settings arise commonly in practice, either in the context of trends or dummy variables.

Example 22.2

Suppose that

$$X = \begin{pmatrix} 1 & D_1 \\ \vdots & \vdots \\ 1 & D_n \end{pmatrix}$$

$$D_i = \begin{cases} 1 & \text{if } i \in I \subset \{1, \ldots, n\} \\ 0 & \text{else} \end{cases}$$

Here, I is the set defining the dummy variable. Then,

$$X^TX = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n D_i \\ \sum_{i=1}^n D_i & \sum_{i=1}^n D_i \end{pmatrix} = \begin{pmatrix} n & n_I \\ n_I & n_I \end{pmatrix}$$

where n_I is the number of observations in set I, because $D_i^2 = D_i$. Therefore,

$$(X^TX)^{-1} = \frac{1}{(n-n_I)n_I}\begin{pmatrix} n_I & -n_I \\ -n_I & n \end{pmatrix}$$

In this case (using Mathematica):

$$\lambda_{min}(X^TX) = \frac{1}{2}n_I + \frac{1}{2}n - \frac{1}{2}\sqrt{5n_I^2 - 2n_In + n^2} = \frac{1}{2}n_I + \frac{1}{2}n\left(1 - \sqrt{5\frac{n_I^2}{n^2} - 2\frac{n_I}{n} + 1}\right)$$

If $n_I \to \infty$ as $n \to \infty$, then $\lambda_{min}(X^TX) \to \infty$. This is the usual case, say, for day-of-the-week dummies. On the other hand, if n_I is fixed, then we can see that $\lambda_{min}(X^TX) \to n_I/2$ as $n \to \infty$ and $\hat{\beta}$ is not consistent. However, even in this case, the first component is consistent because

$$((X^TX)^{-1})_{11} = \frac{1}{n - n_I} \to 0$$

but the second component is inconsistent. One can't estimate consistently the effect of a dummy variable when it only affects a relatively small number of observations.

Example 22.3

Trends. Suppose that

$$X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & n \end{pmatrix}$$

which says that there is potentially a linear trend in the data. In this case,

$$X^TX = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n i \\ \sum_{i=1}^n i & \sum_{i=1}^n i^2 \end{pmatrix} = \begin{pmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{pmatrix}$$

which you know from A-level mathematics. Furthermore (by Mathematica):

$$\lambda_{min}(X^TX) = \frac{7}{12}n - \frac{1}{2}n^2 - \frac{1}{9}n^3 + \frac{1}{36}n^4 + \frac{25}{36}n^2 + \frac{7}{6}n^3 + \frac{1}{4}n^4 + \frac{1}{6}n^2 + \frac{1}{36}n^3$$

This looks nasty, but it can be shown by very tedious work that this goes to infinity. It is easier to see that

$$((X^TX)^{-1})_{11} = \frac{\frac{1}{6}n(n+1)(2n+1)}{\frac{1}{12}n^4 - \frac{1}{12}n^2} \to 0$$

$$((X^TX)^{-1})_{22} = \frac{n}{\frac{1}{12}n^4 - \frac{1}{12}n^2} \to 0$$

Example 22.4

Consider the high-low method for determining average cost:

$$\hat{\beta}_{H-L} = \frac{y_H - y_L}{x_H - x_L}$$

where x_H, x_L are the highest and lowest achieved by the covariate, respectively, and y_H, y_L are the "concomitants", that is, the corresponding values of the outcome variable. This estimator is conditionally unbiased and satisfies:

$$\text{Var}(\hat{\beta}_{H-L}|X) = \frac{2\sigma^2}{(x_H - x_L)^2}$$

For consistency, it suffices that $x_H - x_L \to \infty$ as the sample size increases. This happens for example if the covariates have support on the whole real line, say, are normally distributed.

Example 22.5

Suppose that the regressors are orthogonal, meaning $X^T X = I_K$, but that $K = K(n)$ is large. Then consider a linear combination $c^T \hat{\beta}$. We have:

$$\mathrm{Var}(c^T \hat{\beta}) = \frac{\sigma^2}{n} c^T c = \frac{\sigma^2 K(n)}{n} \frac{1}{K} \sum_{k=1}^{K} c_k^2$$

This says that the individual estimates ($c_k = 1$, $c_j = 0$ for $j \neq k$) are consistent provided $K(n)/n \to 0$.

We next consider the limiting distribution of $\hat{\beta}$. We shall suppose that the covariates are nonrandom, possibly containing trends or dummy variables, but the error terms are independent. We make the following assumptions.

Assumptions R

1. u_i are independent random variables with mean zero and variance σ_i^2 such that $0 < \underline{\sigma}^2 \leq \sigma_i^2 \leq \bar{\sigma}^2 < \infty$ and $E(|u_i|^{2+\delta}) < C < \infty$ for some $\delta > 0$.

2. X is non-stochastic and full rank and satisfy for all $j = 1, \ldots, K$:

$$d_j^2 = \sum_{i=1}^{n} x_{ij}^2 \to \infty$$

$$\frac{\max_{1 \leq i \leq n} x_{ij}^2}{\sum_{i=1}^{n} x_{ij}^2} \to 0$$

3. The $K \times K$ matrices below satisfy for positive definite matrices M, Ψ

$$M_n = \Delta^{-1} X^T X \Delta^{-1} \to M$$

$$\Psi_n = \Delta^{-1} X^T \Sigma X \Delta^{-1} \to \Psi$$

   where $\Delta = \mathrm{diag}\{d_1, \ldots, d_K\}$ and $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$.

Theorem 22.7

Suppose that Assumptions R1-R3 hold. Then

$$\Delta(\hat{\beta} - \beta) \xrightarrow{d} N(0, M^{-1} \Psi M^{-1})$$

Proof:

We write

$$\Delta(\hat{\beta} - \beta) = (\Delta^{-1} X^T X \Delta^{-1})^{-1} \Delta^{-1} X^T u = M_n^{-1} r_n$$

By assumption, M_n converges to M. We consider for any $c \in \mathbb{R}^K$

$$c^T\Delta^{-1}X^Tu = \sum_{j=1}^K c_j d_j^{-1}\sum_{i=1}^n x_{ij}u_i = \sum_{i=1}^n w_{ni}u_i$$

where $w_{ni} = \sum_{j=1}^K c_j d_j^{-1}x_{ij}$. We next apply the Lindeberg CLT (generalized to **triangular arrays** - w_{ni} depends on n). We show that the standardized random variable

$$T_n = \frac{\sum_{i=1}^n w_{ni}u_i}{\left(\sum_{i=1}^n w_{ni}^2\sigma_i^2\right)^{1/2}}$$

converges to a standard normal random variable.

For any i we have

$$\begin{aligned} E\left[u_i^2 1(w_{ni}^2u_i^2 > \varepsilon s_n^2)\right] &\leq E\left[|u_i|^{2+\delta}\left(\frac{w_{ni}^2}{\varepsilon s_n^2}\right)^{\delta/2} 1\left(u_i^2 > \frac{\varepsilon s_n^2}{w_{ni}^2}\right)\right] \\ &\leq \left(\frac{w_{ni}^2}{\varepsilon s_n^2}\right)^{\delta/2} E[|u_i|^{2+\delta}] \\ &\leq C\left(\frac{w_{ni}^2}{\varepsilon s_n^2}\right)^{\delta/2} \end{aligned}$$

where $s_n^2 = \sum_{i=1}^n w_{ni}^2\sigma_i^2$. Therefore, provided

$$\frac{\max_{1\leq i\leq n} w_{ni}^2}{s_n^2} \to 0,$$

we have

$$\frac{1}{\sum_{i=1}^n w_{ni}^2\sigma_i^2}\sum_{i=1}^n w_{ni}^2 E\left[u_i^2 1(w_{ni}^2u_i^2 > \varepsilon s_n^2)\right] \leq \frac{1}{\sum_{i=1}^n w_{ni}^2\sigma_i^2} \times C\left(\frac{1}{\varepsilon^{\delta/2}}\right) \times \left(\frac{\max_{1\leq i\leq n} w_{ni}^2}{s_n^2}\right)^{\delta/2} \to 0$$

We have by R1, R2, and R3 and the Cauchy-Schwarz inequality

$$\begin{aligned} \frac{\max_{1\leq i\leq n}\left(\sum_{j=1}^K c_j d_j^{-1}x_{ij}\right)^2}{\sum_{i=1}^n\left(\sum_{j=1}^K c_j d_j^{-1}x_{ij}\right)^2\sigma_i^2} &\leq \frac{\left(\sum_{j=1}^K |c_j|\max_{1\leq i\leq n}|d_j^{-1}x_{ij}|\right)^2}{\underline{\sigma}^2\sum_{i=1}^n\left(\sum_{j=1}^K c_j d_j^{-1}x_{ij}\right)^2} \\ &= \frac{\left(\sum_{j=1}^K |c_j|\max_{1\leq i\leq n}|d_j^{-1}x_{ij}|\right)^2}{\underline{\sigma}^2 c^T\Delta^{-1}X^TX\Delta^{-1}c} \\ &\leq \frac{K\times c^Tc\times\max_{1\leq j\leq K}\max_{1\leq i\leq n} d_j^{-2}x_{ij}^2}{\underline{\sigma}^2\times c^Tc\times(\lambda_{min}(M)+o(1))} \to 0 \end{aligned}$$

It follows that $T_n \xrightarrow{d} N(0,1)$ for any $c \neq 0$ and hence $\Delta^{-1}X^Tu \xrightarrow{d} N(0,\Psi)$. We apply the multivariate version of the continuous mapping and Slutsky theorem to conclude.

Example 22.6

Suppose that $x_{i1} = 1$, $x_{i2} = i$, and $x_{i3} = i^2$, and u_i are i.i.d. with variance σ^2. Then

$$\Delta = \begin{pmatrix} \sqrt{n} & 0 & 0 \\ 0 & \sqrt{\sum_{i=1}^n i^2} & 0 \\ 0 & 0 & \sqrt{\sum_{i=1}^n i^4} \end{pmatrix} \approx \begin{pmatrix} \sqrt{n} & 0 & 0 \\ 0 & n^{3/2}/\sqrt{3} & 0 \\ 0 & 0 & n^{5/2}/\sqrt{5} \end{pmatrix}$$

$$M = I$$

The negligibility conditions are easily satisfied, for example, if $x_{i2} = i$,

$$\frac{\max_{1 \le i \le n} i^2}{\sum_{j=1}^n j^2} = \frac{n^2}{O(n^3)} \to 0$$

In this case, even though the largest element is increasing with sample size, many other elements are increasing just as fast.

Example 22.7

Consider the dummy variable example $x_i = D_i$, then

$$\frac{\max_{1 \le i \le n} x_i^2}{\sum_{j=1}^n x_j^2} = \frac{1}{\sum_{j=1}^n D_j}$$

which goes to zero if and only if $\sum_{j=1}^n D_j \to \infty$.

Example 22.8

An example where the CLT would fail is

$$x_i = \begin{cases} 1 & \text{if } i < n \\ n & \text{if } i = n \end{cases}$$

In this case, the negligibility condition fails and the distribution of the least squares estimator would be largely determined by the last observation.

We next consider the robust t and Wald statistics in this environment. We let:

$$\hat{\Psi} = \Delta^{-1} X^T S X \Delta^{-1}$$

where S is the $n \times n$ diagonal matrix with a typical element $\hat{\varepsilon}_i^2$. We can write:

$$t_R = \frac{c^T \Delta^{-1} \Delta \hat{\beta}}{\sqrt{c^T \Delta^{-1} (\Delta^{-1} X^T X \Delta^{-1})^{-1} \hat{\Psi} (\Delta^{-1} X^T X \Delta^{-1})^{-1} \Delta^{-1} c}}$$

Define:

$$\hat{x}_i = (x_{i1}, \ldots, x_{iK})^T; \quad \tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^n x_{ij}^2}}$$

Assumption R3 says that $\sum_{i=1}^n \tilde{x}_i \tilde{x}_i^T$ converges to a positive definite matrix. We make further assumption

Assumption R4

For $j,k,l,r = 1,\ldots,K$, we have:

$$\sum_{i=1}^{n} \tilde{x}_{ij}^2 \tilde{x}_{ik}^2, \quad \sum_{i=1}^{n} \tilde{x}_{ij}^2 \tilde{x}_{ik}^2 \tilde{x}_{il}^2, \quad \sum_{i=1}^{n} |\tilde{x}_{ij}\tilde{x}_{ik}\tilde{x}_{il}\tilde{x}_{ir}| \to 0$$

Assumption R5

For some $C < \infty$, we have:

$$E[(\epsilon_i^2 - \sigma_i^2)^2] \leq C$$

Theorem 22.8

Suppose that R1-R5 hold. Then, under H_0:

$$t_R \xrightarrow{d} N(0,1)$$

Proof: We have:

$$\hat{\Psi} = \Delta^{-1} \sum_{i=1}^{n} x_i x_i^T \varepsilon_i^2 \Delta^{-1} - 2\Delta^{-1} \sum_{i=1}^{n} x_i x_i^T \varepsilon_i x_i^T (X^T X)^{-1} X^T \varepsilon + \Delta^{-1} \sum_{i=1}^{n} x_i x_i^T x_i^T (X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1} x_i$$

$$= \hat{\Psi}_1 + \hat{\Psi}_2 + \hat{\Psi}_3$$

where $\tilde{\varepsilon_i} = \varepsilon_i - x_i^T (X^T X)^{-1} X^T \varepsilon$. We have

$$\begin{aligned} \hat{\Psi}_2 &= \Delta^{-1} \sum_{i=1}^{n} x_i x_i^T \tilde{\varepsilon_i} \Delta^{-1} x_i^T (\Delta^{-1} X^T X \Delta^{-1})^{-1} (\Delta^{-1} X^T \varepsilon) \\ &= \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \tilde{\varepsilon_i} (\Delta^{-1} X^T X \Delta^{-1})^{-1} (\Delta^{-1} X^T \varepsilon) \xrightarrow{p} 0 \end{aligned}$$

since $(\Delta^{-1} X^T X \Delta^{-1})^{-1} (\Delta^{-1} X^T \varepsilon) \xrightarrow{d} W \in \mathbb{R}^K$, say, while for any $j,k,l = 1,\ldots,K$ we have

$$E\left[\sum_{i=1}^{n} \tilde{x}_{ij}\tilde{x}_{ik}\tilde{x}_{il}\varepsilon_i\right] = 0$$

$$\begin{aligned} \text{var}\left(\sum_{i=1}^{n} \tilde{x}_{ij}\tilde{x}_{ik}\tilde{x}_{il}\varepsilon_i\right) &= \sum_{i=1}^{n} \tilde{x}_{ij}^2 \tilde{x}_{ik}^2 \tilde{x}_{il}^2 \sigma_i^2 \\ &\leq \bar{\sigma}^2 \sum_{i=1}^{n} \tilde{x}_{ij}^2 \tilde{x}_{ik}^2 \tilde{x}_{il}^2 \to 0. \end{aligned}$$

Likewise

$$\hat{\Psi}_3 = \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T (\Delta^{-1} X^T X \Delta^{-1})^{-1} \Delta^{-1} X^T \varepsilon\varepsilon^T X \Delta^{-1} (\Delta^{-1} X^T X \Delta^{-1})^{-1} \tilde{x}_i \xrightarrow{p} 0$$

because $(\Delta^{-1} X^T X \Delta^{-1})^{-1} \Delta^{-1} X^T \varepsilon\varepsilon^T X \Delta^{-1} (\Delta^{-1} X^T X \Delta^{-1})^{-1} \xrightarrow{d} WW^T$, and the fact that $\sum_{i=1}^{n} |\tilde{x}_{ij}\tilde{x}_{ik}\tilde{x}_{il}\tilde{x}_{ir}| \to 0$. Likewise, for any j,k, we have

$$E\left[\sum_{i=1}^{n}\tilde{x}_{ij}\tilde{x}_{ik}(\epsilon_i^2-\sigma_i^2)\right]=0$$

$$Var\left[\sum_{i=1}^{n}\tilde{x}_{ij}\tilde{x}_{ik}(\epsilon_i^2-\sigma_i^2)\right]=\sum_{i=1}^{n}\tilde{x}_{ij}^2\tilde{x}_{ik}^2E\left[(\epsilon_i^2-\sigma_i^2)^2\right]\leq C\sum_{i=1}^{n}\tilde{x}_{ij}^2\tilde{x}_{ik}^2\to 0$$

Therefore:

$$\hat{\Psi}\xrightarrow{p}\Psi$$

We have

$$t_R\xrightarrow{d}\left(\lim_{n\to\infty}\frac{c^T\Delta^{-1}(M^{-1}\Psi M^{-1})^{-1/2}}{\sqrt{c^T\Delta^{-1}M^{-1}\Psi M^{-1}\Delta^{-1}c}}\right)\times Z=a^TZ$$

where $Z\sim N(0,I_K)$ and a is such that $a^Ta=1$ and the result follows.

Note that a may possess many zeros, so only a few components of Z contribute to the limiting distribution.

Example 22.9

In the case where $x_{ij}=i$ and $x_{ik}=i^2$ we have

$$\sum_{i=1}^{n}\tilde{x}_{ij}^2\tilde{x}_{ik}^2=\frac{\sum_{i=1}^{n}i^6}{\sum_{i=1}^{n}i^2\sum_{i=1}^{n}i^4}\to 0$$

Let's consider the robust Wald statistic, which can be written

$$W_R=(R\Delta^{-1}\Delta(\hat{\beta}-\beta))^T\left[R\Delta^{-1}(\Delta^{-1}X^TX\Delta^{-1})^{-1}\hat{\Psi}(\Delta^{-1}X^TX\Delta^{-1})^{-1}\Delta^{-1}R^T\right]^{-1}R\Delta^{-1}\Delta(\hat{\beta}-\beta)$$

We have $\Delta(\hat{\beta}-\beta)\xrightarrow{d}(M^{-1}\Psi M^{-1})^{1/2}\times x$, where $x\sim N(0,I_K)$ so replacing $\hat{\Psi}$ by its probability limit we have

$$W_R\xrightarrow{d}x^T(M^{-1}\Psi M^{-1})^{1/2}\times\lim_{n\to\infty}\left(\Delta^{-1}R^T\left[R\Delta^{-1}M^{-1}\Psi M^{-1}\Delta^{-1}R^T\right]^{-1}R\Delta^{-1}\right)\times(M^{-1}\Psi M^{-1})^{1/2}\times x$$

The main issue is that we can't guarantee this limit exists, because the matrix $R\Delta^{-1}$ can be of deficient rank in large samples even when R is full rank.

Example 22.10

Suppose that $M=\Psi=I_3$ and

$$\Delta=\text{diag}(n^{1/2},n^{3/2},n^{5/2});\quad R=\begin{pmatrix}1&1&1\\0&0&1\end{pmatrix}$$

Then

$$R\Delta^{-1}=\begin{pmatrix}n^{-1/2}&n^{-3/2}&n^{-5/2}\\0&0&n^{-5/2}\end{pmatrix}$$

$$R\Delta^{-1}\Delta^{-1}R^T=\begin{pmatrix}n^{-1}+n^{-3}+n^{-5}&n^{-5}\\n^{-5}&n^{-5}\end{pmatrix}$$

and the matrix $nR\Delta^{-1}\Delta^{-1}R^T$ is asymptotically singular. In this case, if c is the first row of R we have $a^T = (1, 0, 0)$ so the t-statistic is only driven by the first component.

Intuitively, when there are restrictions across variables that are driven by different trend rates, it is the slowest rate that prevails. See Phillips (2007) for further discussion of regression with trending regressors and some of the pitfalls that may occur and the possible remedies.

Exercises

Exercise 1

Solution 1

Consider the regression model $y_i = \beta^T x_i + \varepsilon_i$, where (x_i, ε_i) are i.i.d. and $E[x_i x_i^T] = M$. Explain in intuitive terms what the matrix M represents.

Exercise 2

Solution 2

Under what conditions is the matrix $\Omega = E[x_i x_i^T \varepsilon_i^2]$ equal to $\sigma^2 M$, where σ^2 is a scalar? Explain the meaning of this condition.

Exercise 3

Solution 3

State the unconditional moment condition and explain why it does not guarantee the unbiasedness of the OLS estimator.

Exercise 4

Solution 4

State the conditional moment condition that ensures the unbiasedness of the OLS estimator and explain why this condition achieves this.

Exercise 5

Solution 5

Explain the result of Theorem 22.1, $\hat{\beta} \xrightarrow{p} \beta$, in intuitive terms. What does it tell us about the OLS estimator in large samples?

Exercise 6

Solution 6

In the proof of Theorem 22.1, what is the role of Slutsky's theorem?

Exercise 7

[Solution 7]

Explain the result of Theorem 22.2, $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, M^{-1}\Omega M^{-1})$, in intuitive terms. What is the distribution of the OLS estimator in large samples?

Exercise 8

[Solution 8]

In the special case of homoskedasticity, how does the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ simplify?

Exercise 9

[Solution 9]

Explain why Theorem 22.3 is important for inference in the linear regression model.

Exercise 10

[Solution 10]

If the homoskedasticity condition fails, what is the consequence for the usual t-statistic?

Exercise 11

[Solution 11]

What is the purpose of using robust test statistics (t_R and W_R) as defined in equations (22.2) and (22.3)?

Exercise 12

[Solution 12]

In Theorem 22.5, what do the conditions $E|x_{ij}x_{ik}x_{il}x_{ir}| < \infty$, $E|x_{ij}x_{ik}x_{il}\varepsilon_i| < \infty$, and $E(x_{ij}x_{ik}\varepsilon_i x_{il}) = 0$ ensure?

Exercise 13

[Solution 13]

Explain the non-i.i.d. case considered in section 22.2. How does it differ from the i.i.d. case?

Exercise 14

[Solution 14]

In Theorem 22.6, what is the significance of the condition $\lambda_{max}(X^T\Sigma X)\lambda_{min}^{-2}(X^TX) \to 0$ as $n \to \infty$?

Exercise 15

[Solution 15]

In Example 22.2, why is the OLS estimator of the dummy variable coefficient inconsistent when the number of observations in the set I is fixed?

Exercise 16

Solution 16

Explain Assumptions R1, R2, and R3 in the context of Theorem 22.7.

Exercise 17

Solution 17

What is a triangular array, and why is the Lindeberg CLT used for triangular arrays in the proof of Theorem 22.7?

Exercise 18

Solution 18

In Example 22.6, explain the scaling matrix Δ and why different scaling is needed for different regressors.

Exercise 19

Solution 19

In Theorem 22.8, what are the assumptions R4 and R5 ensuring?

Exercise 20

Solution 20

In Example 22.10, explain why the t-statistic is driven only by the first component when the matrix $R\Delta^{-1}$ is asymptotically singular.

Solutions

Solution 1

Exercise 1

The matrix $M = E[x_i x_i^T]$ represents the expected outer product of the regressor vector x_i with itself. Since x_i is a $K \times 1$ vector, $x_i x_i^T$ is a $K \times K$ matrix. The elements of this matrix are the expected values of the squares and cross-products of the regressors. Specifically, the diagonal elements, $E[x_{ij}^2]$, represent the expected values of the squared regressors, and the off-diagonal elements, $E[x_{ij}x_{ik}]$ for $j \neq k$, represent the expected values of the cross-products of different regressors. The matrix M summarizes the second moments of the regressors. Intuitively, it reflects the spread and the linear relationships between the regressors.

Solution 2

Exercise 2

The matrix $\Omega = E[x_i x_i^T \varepsilon_i^2]$ is equal to $\sigma^2 M$ under the condition of **homoskedasticity**, which means that the variance of the error term ε_i is constant and does not depend on the regressors x_i. Formally, this is stated as

$E[\varepsilon_i^2|x_i] = \sigma^2$ for all i. If this holds, then

$$\Omega = E[x_i x_i^T \varepsilon_i^2] = E[E[x_i x_i^T \varepsilon_i^2|x_i]] = E[x_i x_i^T E[\varepsilon_i^2|x_i]] = E[x_i x_i^T \sigma^2] = \sigma^2 E[x_i x_i^T] = \sigma^2 M$$

Solution 3

Exercise 3

The **unconditional moment condition** is $E[x_i \varepsilon_i] = 0$. This means that the regressors and the error term are uncorrelated *on average*. However, this does *not* guarantee the unbiasedness of the OLS estimator. Unbiasedness requires that $E[\hat{\beta}|X] = \beta$. The unconditional moment condition is not strong enough to imply this conditional expectation. It only implies that $E[X^T \varepsilon] = 0$, but it doesn't guarantee that $E[(X^T X)^{-1} X^T \varepsilon|X] = 0$, as is needed for unbiasedness.

Solution 4

Exercise 4

The **conditional moment condition** that ensures the unbiasedness of the OLS estimator is $E[\varepsilon_i|x_i] = 0$. This condition means that the expected value of the error term, given any value of the regressors, is zero. It is a stronger condition than the unconditional moment condition. This condition ensures unbiasedness because:

$$\begin{aligned} E[\hat{\beta}|X] &= E[\beta + (X^T X)^{-1} X^T \varepsilon|X] \\ &= \beta + E[(X^T X)^{-1} X^T \varepsilon|X] \\ &= \beta + (X^T X)^{-1} X^T E[\varepsilon|X] \\ &= \beta + (X^T X)^{-1} X^T \cdot 0 \\ &= \beta \end{aligned}$$

Here, we use the fact that X is treated as fixed in the conditional expectation, and $E[\varepsilon|X]$ is a vector with elements $E[\varepsilon_i|x_i] = 0$.

Solution 5

Exercise 5

Theorem 22.1, $\hat{\beta} \xrightarrow{p} \beta$, states that the OLS estimator $\hat{\beta}$ **converges in probability** to the true parameter vector β. This means that as the sample size n goes to infinity, the probability that $\hat{\beta}$ is arbitrarily close to β approaches 1. In intuitive terms, this tells us that in large samples, the OLS estimator is likely to be very close to the true value of the parameters. This is the property of **consistency** of an estimator.

Solution 6

Exercise 6

In the proof of Theorem 22.1, Slutsky's theorem is used to combine the results from the Law of Large Numbers applied to the two terms:

$$\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \xrightarrow{p} M^{-1}$$

and

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

Slutsky's theorem states that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then $X_nY_n \xrightarrow{d} cX$, and $X_n + Y_n \xrightarrow{d} X + c$ and, $Y_n^{-1}X_n \xrightarrow{d} c^{-1}X$, if $c \neq 0$. Here we have convergence in probability, which is a special, simpler, case of convergence in distribution, and it allow us to conclude that the product converges in probability to $M^{-1} \cdot 0 = 0$:

$$\hat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right) \xrightarrow{p} M^{-1} \cdot 0 = 0$$

Solution 7

Exercise 7

Theorem 22.2 states that $\sqrt{n}(\hat{\beta} - \beta)$ **converges in distribution** to a normal distribution with mean 0 and variance $M^{-1}\Omega M^{-1}$. This means that, in large samples, the *scaled* difference between the OLS estimator and the true parameter vector is approximately normally distributed. This result is crucial for constructing confidence intervals and hypothesis tests about β. The scaling factor $\sqrt{n}$ indicates the rate of convergence: $\hat{\beta}$ approaches β at a rate proportional to $1/\sqrt{n}$.

Solution 8

Exercise 8

In the special case of homoskedasticity, where $E[\varepsilon_i^2|x_i] = \sigma^2$, the asymptotic variance simplifies to $\sigma^2 M^{-1}$. This is because, under homoskedasticity, $\Omega = \sigma^2 M$, so:

$$M^{-1}\Omega M^{-1} = M^{-1}(\sigma^2 M)M^{-1} = \sigma^2 M^{-1}MM^{-1} = \sigma^2 M^{-1}$$

Solution 9

Exercise 9

Theorem 22.3, $s_*^2 \xrightarrow{p} \sigma^2$, is important for inference because s_*^2 is the estimator of the error variance σ^2 that is used in the calculation of the standard errors of the OLS estimator $\hat{\beta}$. Since s_*^2 is consistent for σ^2, we can use it to consistently estimate the variance-covariance matrix of $\hat{\beta}$, which is needed for constructing t-statistics and F-statistics.

Solution 10

Exercise 10

If the homoskedasticity condition fails, the usual t-statistic does *not* converge to a standard normal distribution. Instead, it converges to a normal distribution with a variance that is not equal to 1. This means that using the standard normal critical values for hypothesis testing will lead to incorrect inference (incorrect size of the test).

Solution 11

Exercise 11

The purpose of using robust test statistics (t_R and W_R) is to obtain valid inference even when the homoskedasticity assumption is violated. These statistics use a consistent estimator of the variance-covariance matrix of $\hat{\beta}$ that does not rely on the assumption of constant error variance. The matrix $\hat{\Omega}$ provides a consistent estimate of Ω even under heteroskedasticity.

Solution 12

Exercise 12

In Theorem 22.5, the conditions $E|x_{ij}x_{ik}x_{il}x_{ir}| < \infty$, $E|x_{ij}x_{ik}x_{il}\varepsilon_i| < \infty$, and $E(x_{ij}x_{ik}\varepsilon_i x_{il}) = 0$ are moment conditions that ensure that certain sample averages involving the regressors and the error terms converge to their population counterparts. These conditions, together with B1 and B2, guarantee the consistency of $\hat{\Omega}$ for Ω under heteroskedasticity. They are technical conditions required for the application of laws of large numbers.

Solution 13

Exercise 13

The non-i.i.d. case considered in section 22.2 refers to situations where the observations (x_i, ε_i) are *not* independently and identically distributed. This can occur, for example, when the regressors include deterministic trends or dummy variables or when the errors have different variances (heteroscedasticity), or are autocorrelated. This contrasts with the i.i.d. case, where each observation is assumed to be drawn from the same distribution and independent of other observations.

Solution 14

Exercise 14

In Theorem 22.6, the condition $\lambda_{max}(X^T\Sigma X)\lambda_{min}^{-2}(X^TX) \to 0$ as $n \to \infty$ is a condition that ensures the consistency of the OLS estimator in the non-i.i.d. case. It essentially requires that the variance of the error terms (represented by Σ) does not grow too fast relative to the information in the regressors (represented by X^TX). The smallest eigenvalue of X^TX, $\lambda_{min}(X^TX)$, measures how much information X contains. The largest eigenvalue $\lambda_{max}(X^T\Sigma X)$ can be thought of as the largest variance in the errors projected onto the space of X. If $\lambda_{max}(X^T\Sigma X)$ grows at a faster rate than the square of $\lambda_{min}(X^TX)$, then the increased variability in y due to the non-constant variance in Σ dominates the increased information in X^TX obtained by increasing the sample size, n.

Solution 15

Exercise 15

In Example 22.2, the OLS estimator of the dummy variable coefficient is inconsistent when the number of observations in the set I (n_I) is fixed because the information about the effect of that dummy variable does not grow with the sample size. As n increases, the number of observations *outside* the set I grows, but the number of observations *inside* the set I remains constant. This means that we are not getting more information about the effect of the dummy variable as the sample size increases, and thus we cannot consistently estimate its coefficient. This corresponds to the fact that, in this case, $\lambda_{min}(X^TX)$ does not tend to infinity.

Solution 16

Exercise 16

Assumptions R1, R2, and R3 in the context of Theorem 22.7 are conditions on the error terms and the regressors that are needed to establish the asymptotic normality of the OLS estimator in the non-i.i.d. case.

- **R1:** The error terms u_i are independent (but not necessarily identically distributed) with mean zero and bounded variances. The condition $E(|u_i|^{2+\delta}) < C$ is a technical condition (Lyapunov condition) that ensures that the central limit theorem can be applied.
- **R2:** The regressors are non-stochastic (fixed) and full rank. The condition $d_j^2 = \sum_{i=1}^n x_{ij}^2 \to \infty$ ensures that the information in each regressor grows with the sample size. The condition $\frac{\max_{1\le i\le n} x_{ij}^2}{\sum_{i=1}^n x_{ij}^2} \to 0$ is a "no single observation dominates" condition. It prevents any single observation from having an overwhelming influence on the estimator.
- **R3:** This condition states that the scaled regressor matrices converge to positive definite matrices M and Ψ. This ensures that the information in the regressors, appropriately scaled, stabilizes as the sample size increases, and that there is sufficient variation in the data. The matrix Δ scales the regressors based on their individual sums of squares.

Solution 17

Exercise 17

A **triangular array** is a sequence of random variables where the number of variables in each row depends on the row number. In the proof of Theorem 22.7, the w_{ni} terms form a triangular array because they depend on both i (the observation index) and n (the sample size). The Lindeberg CLT is used for triangular arrays because the classical CLT applies to sums of i.i.d. random variables, which is not the case here. The Lindeberg CLT provides conditions under which the sum of independent, but not necessarily identically distributed, random variables (properly normalized) converges to a standard normal distribution. The Lindeberg condition (Equation 22.5) ensures that no single term in the sum dominates the others, preventing the sum from being driven by a few observations.

Solution 18

Exercise 18

In Example 22.6, the scaling matrix Δ is a diagonal matrix with elements that scale each regressor differently. The diagonal elements are the square roots of the sums of squares of the corresponding regressors: $\sqrt{n}$, $\sqrt{\sum_{i=1}^n i^2}$, and $\sqrt{\sum_{i=1}^n i^4}$. Different scaling is needed because the regressors (1, i, and i^2) grow at different rates as the sample size n increases. The constant regressor grows at rate n, the linear trend grows at rate n^3, and the quadratic trend grows at rate n^5. To apply the central limit theorem, we need to normalize each regressor by its "size" so that the scaled regressors have comparable magnitudes.

Solution 19

Exercise 19

In Theorem 22.8, assumptions R4 and R5 are additional conditions needed to prove the consistency of the robust variance estimator $\hat{\Psi}$ in the non-i.i.d. case.

- **R4:** This assumption places restrictions on the higher-order moments of the scaled regressors ($\tilde{x}_{ij}$). It helps ensure that certain sums involving products of the scaled regressors converge to zero, which is necessary for the consistency of $\hat{\Psi}$.
- **R5:** This assumption bounds the variance of the squared errors, which is necessary to show $\hat{\Psi}_1 \xrightarrow{p} \Psi$

Solution 20

Exercise 20

In Example 22.10, the t-statistic is driven only by the first component when the matrix $nR\Delta^{-1}\Delta^{-1}R^T$ is asymptotically singular because the restriction matrix R effectively eliminates the influence of the faster-growing regressors. The matrix Δ^{-1} scales the regressors by $n^{-1/2}$, $n^{-3/2}$, and $n^{-5/2}$. The matrix $R = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ imposes the restrictions. When we multiply R by Δ^{-1}, we get a matrix where the elements corresponding to the faster-growing regressors ($n^{-3/2}$ and $n^{-5/2}$) are much smaller than the element corresponding to the slowest-growing regressor ($n^{-1/2}$). The restriction effectively says that the coefficients sum to the same, but since the coefficients are associated with trends growing at different speeds, that sum is determined by the slowest. Asymptotically, the t-statistic only reflects the variation associated with the slowest-growing component (the constant term in this case).

R Script Examples

R Script 1: Illustration of Consistency (Theorem 22.1)

```
# Load necessary libraries
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Set seed for reproducibility
set.seed(123)

# Define a function to run the simulation
run_simulation <- function(n) {
  # Generate i.i.d. data
  x <- rnorm(n, mean = 2, sd = 1) # x_i ~ N(2, 1)
  epsilon <- rnorm(n, mean = 0, sd = 1) # epsilon_i ~ N(0, 1)
  beta_true <- c(1, 2) # True beta
  y <- beta_true[1] + beta_true[2] * x + epsilon # y_i = 1 + 2x_i + epsilon_i

  # Calculate OLS estimator
  X <- cbind(1, x) # Design matrix
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y

  return(beta_hat)
}

# Run simulations for different sample sizes
sample_sizes <- c(10, 50, 100, 500, 1000, 5000)
results <- lapply(sample_sizes, run_simulation)

# Convert results to a data frame
results_df <- data.frame(
  n = sample_sizes,
  beta_0_hat = sapply(results, function(x) x[1]),
  beta_1_hat = sapply(results, function(x) x[2])
)
```

```
# Print the results
print(results_df)
```

```
     n  beta_0_hat beta_1_hat
1   10 -0.09561438   2.628661
2   50  1.26398748   1.838626
3  100  1.11766332   1.991450
4  500  0.97820790   2.024957
5 1000  0.90675784   2.045172
6 5000  0.98676169   2.003172
```

```
# Plot the results
results_df_long <- results_df %>%
  pivot_longer(cols = c(beta_0_hat, beta_1_hat),
               names_to = "Coefficient",
               values_to = "Estimate")

ggplot(results_df_long, aes(x = n, y = Estimate, color = Coefficient)) +
  geom_line() +
  geom_point() +
  geom_hline(yintercept = c(1, 2), linetype = "dashed", color = c("red", "blue")) +
  scale_x_log10() +
  labs(title = "OLS Estimates vs. Sample Size",
       x = "Sample Size (n) - Log Scale",
       y = "Estimate") +
  theme_bw()
```

Explanation:

1. **Load Libraries:** The `tidyverse` library is loaded for data manipulation and visualization.
2. **Set Seed:** `set.seed(123)` ensures that the results are reproducible.
3. **`run_simulation` Function:** This function simulates data from the linear model $y_i = 1 + 2x_i + \varepsilon_i$, where $x_i \sim N(2, 1)$ and $\varepsilon_i \sim N(0, 1)$. It takes the sample size `n` as input. The true parameters are $\beta_0 = 1$ and $\beta_1 = 2$. The function calculates the OLS estimator $\hat{\beta}$ using the formula $\hat{\beta} = (X^T X)^{-1} X^T y$.
4. **Simulations:** The `run_simulation` function is called for various sample sizes (`sample_sizes`).
5. **Data Frame:** The results are stored in a data frame `results_df`.
6. **Print results:** The resulting estimates are printed.
7. **Plotting:** The `ggplot2` library (part of `tidyverse`) is used to create a plot showing how the OLS estimates of β_0 and β_1 change as the sample size increases. The x-axis is on a log scale to better visualize the convergence. The dashed horizontal lines represent the true values of β_0 and β_1.

Connection to the Text: This script illustrates **Theorem 22.1**, which states that the OLS estimator is consistent ($\hat{\beta} \xrightarrow{p} \beta$). As the sample size (`n`) increases, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ get closer and closer to the true values (1 and 2, respectively), demonstrating convergence in probability.

R Script 2: Illustration of Asymptotic Normality (Theorem 22.2)

```
# Load necessary libraries
library(tidyverse)

# Set seed for reproducibility
set.seed(456)

# Simulation parameters
n <- 1000        # Sample size
num_simulations <- 10000 # Number of simulations
beta_true <- c(1, 2)      # True beta
```

```
# Function to generate data and calculate scaled difference
run_simulation_normality <- function() {
  # Generate i.i.d. data
  x <- rnorm(n, mean = 2, sd = 1)
  epsilon <- rnorm(n, mean = 0, sd = 1)
  y <- beta_true[1] + beta_true[2] * x + epsilon

  # Calculate OLS estimator
  X <- cbind(1, x)
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y

  # Calculate M
  M <- matrix(c(1, mean(x), mean(x), mean(x^2)), nrow = 2)

  # Calculate Omega (homoskedastic case)
  sigma2_hat <- sum((y - X %*% beta_hat)^2) / (n - 2)
  Omega <- sigma2_hat * M

  # Calculate scaled difference
  scaled_diff <- sqrt(n) * (beta_hat - beta_true)

  return(scaled_diff)
}

# Run simulations
results <- replicate(num_simulations, run_simulation_normality())

# Convert results to a data frame
results_df <- data.frame(
  beta_0_scaled = results[1, , ],
  beta_1_scaled = results[2, , ]
)

# Plot the distributions
results_df_long <- results_df %>%
  pivot_longer(cols = everything(),
               names_to = "Coefficient",
               values_to = "Scaled_Difference")

# Find the standard deviations
sd <- results_df_long %>%
  group_by(Coefficient) %>%
  summarize(sd = sd(Scaled_Difference)) %>%
  pull(sd)

# Plot 1
results_df_long %>%
  filter(str_detect(Coefficient, "0")) %>%
  ggplot(aes(x = Scaled_Difference)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50, alpha = 0.7) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = sd[1]),
                color = "red", linewidth = 1) +
  facet_wrap(~ Coefficient, scales = "free") +
  labs(title = "Distribution of Scaled Difference (sqrt(n)*(beta_hat - beta))",
       x = "Scaled Difference",
       y = "Density") +
  theme_bw()
```

```
# Plot 2
results_df_long %>%
  filter(str_detect(Coefficient, "1")) %>%
  ggplot(aes(x = Scaled_Difference)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50, alpha = 0.7) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = sd[2]),
                color = "red", linewidth = 1) +
  facet_wrap(~ Coefficient, scales = "free") +
  labs(title = "Distribution of Scaled Difference (sqrt(n)*(beta_hat - beta))",
       x = "Scaled Difference",
       y = "Density") +
  theme_bw()
```

Explanation:

1. **Load Libraries:** The `tidyverse` library is loaded.
2. **Set Seed:** `set.seed(456)` ensures reproducibility.
3. **Simulation Parameters:** The sample size (`n`), the number of simulations (`num_simulations`), and the true parameter vector (`beta_true`) are defined.
4. **`run_simulation_normality` Function:** This function generates data, calculates the OLS estimator $\hat{\beta}$, and then calculates the *scaled difference* $\sqrt{n}(\hat{\beta} - \beta)$. It also estimates M and Ω (assuming homoskedasticity).
5. **Simulations:** The `replicate` function runs the `run_simulation_normality` function many times.
6. **Data Frame:** The results (the scaled differences) are stored in a data frame.
7. **Plotting:** Histograms of the scaled differences for $\hat{\beta}_0$ and $\hat{\beta}_1$ are created. A standard normal density curve (red line) is overlaid on each histogram to show how closely the distributions of the scaled differences resemble a normal distribution. `facet_wrap` creates separate plots for each coefficient.

Connection to the Text: This script illustrates **Theorem 22.2**, which states that $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a normal distribution. The histograms show that the distributions of the scaled differences are approximately normal, even for a moderate sample size (n=1000).

R Script 3: Heteroskedasticity and Robust Standard Errors

```
# Load necessary libraries
library(tidyverse)
library(sandwich) # For robust standard errors
library(lmtest)   # For coeftest
```

```
Loading required package: zoo
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
# Set seed for reproducibility
set.seed(789)

# Simulation parameters
n <- 200
beta_true <- c(1, 2)

# Generate data with heteroskedasticity
x <- runif(n, min = 1, max = 5)
epsilon <- rnorm(n, mean = 0, sd = x) # Heteroskedastic errors: sd depends on x
```

```
y <- beta_true[1] + beta_true[2] * x + epsilon

# Fit OLS model
model <- lm(y ~ x)

# Calculate usual standard errors
summary(model)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-10.003  -1.505   0.287   1.925  12.512

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0214     0.6534   1.563     0.12
x             1.8903     0.2079   9.091   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.266 on 198 degrees of freedom
Multiple R-squared:  0.2945,    Adjusted R-squared:  0.2909
F-statistic: 82.64 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
# Calculate heteroskedasticity-robust standard errors (White's standard errors)
vcov_robust <- vcovHC(model, type = "HC1") # HC1 is a common choice
coeftest(model, vcov. = vcov_robust)
```

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.02140    0.54616  1.8702   0.06294 .
x            1.89026    0.22883  8.2604 2.048e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Create a plot showing the heteroskedasticity
data_df <- data.frame(x = x, y = y, residuals = residuals(model))

ggplot(data_df, aes(x = x, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs. x (Illustrating Heteroskedasticity)",
       x = "x",
       y = "Residuals") +
  theme_bw()
```

Explanation:

1. **Load Libraries:** `tidyverse`, `sandwich`, and `lmtest` are loaded. `sandwich` provides functions for robust covariance matrix estimation, and `lmtest` provides functions for hypothesis testing using these robust estimates.
2. **Set Seed:** `set.seed(789)` ensures reproducibility.
3. **Simulation Parameters:** `n` and `beta_true` are defined.
4. **Generate Data:** Data is generated with **heteroskedasticity**: the standard deviation of the error term `epsilon` depends on the value of `x`. This violates the homoskedasticity assumption.
5. **Fit OLS Model:** `lm(y ~ x)` fits the linear regression model.

6. **Usual Standard Errors:** `summary(model)` provides the usual OLS output, including standard errors that assume homoskedasticity.
7. **Robust Standard Errors:** `vcovHC(model, type = "HC1")` calculates White's heteroskedasticity-consistent covariance matrix. `coeftest(model, vcov. = vcov_robust)` displays the coefficient estimates along with the robust standard errors.
8. **Plot:** A scatterplot of the residuals against `x` is created to visualize the heteroskedasticity. The spread of the residuals increases with `x`, indicating that the variance of the error term is not constant.

Connection to the Text: This script demonstrates the issue of **heteroskedasticity** and how to use **robust standard errors** to address it. The text discusses how the usual standard errors are inconsistent under heteroskedasticity, leading to incorrect inference. The script shows how to calculate White's robust standard errors, which provide consistent estimates of the standard errors even when heteroskedasticity is present, as suggested by Theorem 22.5 and equations (22.2) and (22.3).

R Script 4: Non-IID Case - Dummy Variable with Fixed Number of Observations

```
# Load necessary libraries
library(tidyverse)

# Set seed for reproducibility
set.seed(101112)

# Simulation parameters
n_total <- 1000  # Total sample size
n_group1 <- 10   # Fixed size of group 1 (dummy variable = 1)
beta_true <- c(1, 3)

# Generate data
x <- c(rep(1, n_group1), rep(0, n_total - n_group1)) # Dummy variable
epsilon <- rnorm(n_total, mean = 0, sd = 1)
y <- beta_true[1] + beta_true[2] * x + epsilon

# Create a sequence of increasing sample sizes, keeping n_group1 fixed
sample_sizes <- seq(50, n_total, by = 50)

# Function to run simulation for a given total sample size
run_dummy_simulation <- function(n_total) {
  x <- c(rep(1, n_group1), rep(0, n_total - n_group1))
  epsilon <- rnorm(n_total, mean = 0, sd = 1)
  y <- beta_true[1] + beta_true[2] * x + epsilon
  X <- cbind(1, x)
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}

# Run simulations for different total sample sizes
results <- lapply(sample_sizes, run_dummy_simulation)

# Convert results to a dataframe
results_df <- data.frame(
  n_total = sample_sizes,
  beta_0_hat = sapply(results, function(x) x[1]),
  beta_1_hat = sapply(results, function(x) x[2])
)

# Print results
print(results_df)
```

```
   n_total beta_0_hat beta_1_hat
1       50  0.8499523   3.079081
2      100  0.9987144   3.102146
```

```
3      150  0.9555374   3.551848
4      200  0.9345526   3.403887
5      250  0.9677228   3.460532
6      300  1.0198353   3.161775
7      350  1.0231876   3.139851
8      400  0.9620140   3.474453
9      450  0.9695749   3.209462
10     500  0.9712435   3.641636
11     550  1.0383135   3.618084
12     600  1.0249506   3.379474
13     650  0.9827814   2.916996
14     700  1.0979926   2.868049
15     750  0.9586777   3.084809
16     800  1.0458507   2.441791
17     850  0.9990818   3.494196
18     900  0.9706746   2.549287
19     950  1.0672488   3.148554
20    1000  1.0090117   3.315724
```

```
# Plot the results
results_df_long <- results_df %>%
  pivot_longer(cols = c(beta_0_hat, beta_1_hat),
               names_to = "Coefficient",
               values_to = "Estimate")

ggplot(results_df_long, aes(x = n_total, y = Estimate, color = Coefficient)) +
  geom_line() +
  geom_point() +
  geom_hline(yintercept = c(beta_true[1], beta_true[2]), linetype = "dashed",
             color = c("red","blue")) +
  labs(title = "OLS Estimates with Fixed Dummy Group Size",
       x = "Total Sample Size (n_total)",
       y = "Estimate") +
  theme_bw()
```

Explanation:

1. **Load Libraries:** `tidyverse` is loaded.
2. **Set Seed:** `set.seed(101112)` for reproducibility.
3. **Simulation Parameters:** `n_total` (total sample size) and `n_group1` (fixed size of the group where the dummy variable is 1) are defined, along with `beta_true`.
4. **Generate Data:** A dummy variable `x` is created. `n_group1` observations have `x = 1`, and the rest have `x = 0`. The error term is i.i.d. normal.
5. **Increasing Sample Sizes:** A sequence of increasing *total* sample sizes (`sample_sizes`) is created.
6. **`run_dummy_simulation` Function:** This function simulates data with a fixed number of observations in group 1 (`n_group1`) and a varying total number of observations, then estimates and returns the betas.
7. **Simulations:** The `run_dummy_simulation` function is called for each total sample size in `sample_sizes`.
8. **Data Frame and Print:** The results are stored in a dataframe and printed.
9. **Plotting:** The plot shows how the OLS estimates of β_0 and β_1 change as the *total* sample size increases, while the number of observations with `x = 1` remains fixed.

Connection to the Text: This script illustrates the scenario described in **Example 22.2**, where a dummy variable has a fixed number of observations in one group. As discussed in the text, the OLS estimator for the coefficient of the dummy variable (β_1 in this case) is *inconsistent* because the information about that coefficient does not grow as the total sample size increases. You should observe in the plot and the results table that $\hat{\beta}_0$ converges to β_0 but $\hat{\beta}_1$ *does not* converge to β_1.

R Script 5: Illustration of Lindeberg CLT (Theorem 22.7)

```r
# Load necessary library
library(tidyverse)

# Set seed for reproducibility
set.seed(2024)

# Simulation parameters
n <- 500
num_simulations <- 10000
beta_true <- c(2,-1, 0.5)
K <- length(beta_true)

# Function for the simulation
lindeberg_sim <- function() {

    # Generate non-iid data.  x will be a triangular array.
    x1 <- runif(n, 0, 1)
    x2 <- 1:n # Linear Trend
    x3 <- rnorm(n, 0, sqrt(1:n)) # Increasing variance
    X <- cbind(x1, x2, x3)

    # Heteroskedastic errors
    sigma <- sqrt(0.5 + 0.1 * x2 + 0.01 * x3^2)
    epsilon <- rnorm(n, 0, sigma)

    y <- X %*% beta_true + epsilon

    # Calculate OLS estimator
    beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y

    # Calculate Delta
    Delta <- diag(c(sqrt(sum(x1^2)), sqrt(sum(x2^2)), sqrt(sum(x3^2))))

    # Calculate scaled difference
    scaled_diff <- Delta %*% (beta_hat - beta_true)

    return(scaled_diff)
}

# Run the simulations
results <- replicate(num_simulations, lindeberg_sim(), simplify = TRUE)

# Transform the results in a data frame
results_df <- data.frame(t(results))
names(results_df) <- paste0("beta_", 0:(K-1), "_scaled")

# Calculate theoretical asymptotic variance (for comparison)
x1 <- runif(n, 0, 1)
x2 <- 1:n # Linear Trend
x3 <- rnorm(n, 0, sqrt(1:n))
X <- cbind(x1, x2, x3)
sigma <- sqrt(0.5 + 0.1 * x2 + 0.01 * x3^2)
Delta <- diag(c(sqrt(sum(x1^2)), sqrt(sum(x2^2)), sqrt(sum(x3^2))))
M <- solve(Delta) %*% t(X) %*% X %*% solve(Delta)
Psi <- solve(Delta) %*% t(X) %*% diag(sigma^2) %*% X %*% solve(Delta)
asymptotic_variance <- solve(M) %*% Psi %*% solve(M)

# Plotting
results_df_long <- results_df %>%
  pivot_longer(cols = everything(),
               names_to = "Coefficient",
               values_to = "Scaled_Difference")
```

```
# Find the standard deviations
sd <- results_df_long %>%
  group_by(Coefficient) %>%
  summarize(sd = sd(Scaled_Difference)) %>%
  pull(sd)

# Generate the plots
results_df_long %>%
  filter(str_detect(Coefficient, "0")) %>%
ggplot(aes(x = Scaled_Difference)) +
    geom_histogram(aes(y = after_stat(density)), bins = 50, alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0,
                                           sd = sd[1]
                                           ),
                  color = "red", linewidth = 1) +
    facet_wrap(~Coefficient, scales = "free") +
    labs(title = "Distribution of Scaled Difference (Delta*(beta_hat - beta))",
         x = "Scaled Difference",
         y = "Density") +
     theme_bw()
```

```
results_df_long %>%
  filter(str_detect(Coefficient, "1")) %>%
ggplot(aes(x = Scaled_Difference)) +
    geom_histogram(aes(y = after_stat(density)), bins = 50, alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0,
                                           sd = sd[2]
                                           ),
                  color = "red", linewidth = 1) +
    facet_wrap(~Coefficient, scales = "free") +
    labs(title = "Distribution of Scaled Difference (Delta*(beta_hat - beta))",
         x = "Scaled Difference",
         y = "Density") +
     theme_bw()
```

```
results_df_long %>%
  filter(str_detect(Coefficient, "2")) %>%
ggplot(aes(x = Scaled_Difference)) +
    geom_histogram(aes(y = after_stat(density)), bins = 50, alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0,
                                           sd = sd[3]
                                           ),
                  color = "red", linewidth = 1) +
    facet_wrap(~Coefficient, scales = "free") +
    labs(title = "Distribution of Scaled Difference (Delta*(beta_hat - beta))",
         x = "Scaled Difference",
         y = "Density") +
     theme_bw()
```

Explanation:

1. **Load Libraries:** The `tidyverse` package is loaded.
2. **Set Seed:** `set.seed(2024)` ensures reproducibility.
3. **Simulation Parameters:** `n`, `num_simulations`, `beta_true`, and `K` are defined.
4. **`lindeberg_sim` Function:** This function simulates data that violates the i.i.d. assumption.

    - `x1`: Uniformly distributed.
    - `x2`: A linear trend (1, 2, 3, …, n). This creates a triangular array because each value depends on its index within the sequence.
    - `x3`: Normally distributed with increasing variance.
    - `epsilon`: The errors are heteroskedastic, with variance depending on `x2` and `x3`.
    - The function computes $\hat{\beta}$ and then $\Delta(\hat{\beta} - \beta)$
5. **Simulations:** The function is called multiple times.
6. **Theoretical Asymptotic Variance:** The theoretical asymptotic variance ($M^{-1}\Psi M^{-1}$) is calculated using the true data-generating process, for comparison.
7. **Plotting:** Histograms of the scaled differences ($\Delta(\hat{\beta} - \beta)$) are plotted, with normal density curves overlaid. The `facet_wrap` function creates separate plots for each coefficient.

Connection to the Text: This script illustrates **Theorem 22.7**, which deals with the asymptotic distribution of the OLS estimator in the *non-i.i.d.* case. The data are generated to satisfy Assumptions R1-R3, allowing the application of the Lindeberg CLT. Even though the regressors and errors are not i.i.d., the scaled differences $\Delta(\hat{\beta} - \beta)$ are approximately normally distributed in large samples, as the histograms demonstrate. The scaling matrix Δ is crucial for achieving this normality, as different regressors have different rates of growth.

YouTube Videos

Here are some YouTube videos that explain concepts related to Chapter 22, "Asymptotic Properties of OLS Estimator and Test Statistics." I have verified that all links are currently working (as of October 26, 2023).

1. Consistency of OLS

- **Video Title:** Consistency of OLS
- **Channel:** Ben Lambert
- **Link:** https://www.youtube.com/watch?v=tG0nkHFy-Ik
- **Relation to Text:** This video directly addresses the concept of **consistency** of the OLS estimator, which is the main topic of **Theorem 22.1** ($\hat{\beta} \xrightarrow{p} \beta$). The video explains what it means for an estimator to be consistent, and it provides a visual and intuitive explanation of why OLS is consistent under the standard assumptions. It covers similar ground to the proof of Theorem 22.1, although perhaps less formally.

2. Asymptotic Normality of OLS

- **Video Title:** Asymptotic Normality of OLS
- **Channel:** Ben Lambert
- **Link:** https://www.youtube.com/watch?v=DAbE7l-j-9A
- **Relation to Text:** This video covers the **asymptotic normality** of the OLS estimator, which is precisely the statement of **Theorem 22.2** ($\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, M^{-1}\Omega M^{-1})$). The video explains how the central limit theorem is used to derive the asymptotic distribution of the OLS estimator and discusses the implications for inference (hypothesis testing and confidence intervals). It provides the intuition behind the theorem, although without a full formal proof. It focuses on the i.i.d. case.

3. Heteroskedasticity and Robust Standard Errors

- **Video Title:** (1 of 4) Heteroskedasticity
- **Channel:** jbstatistics
- **Link:** https://www.youtube.com/watch?v=V2Z8qa5J2bk

- **Relation to Text:** This is part 1 of a four-part series. It introduces **heteroskedasticity**, explaining what it is and why it's a problem. This relates to the discussion in the text before and after Theorem 22.3, and the motivation for **Theorem 22.5**. The text mentions that if homoskedasticity fails, the usual t-statistic does not have a standard normal distribution. This video explains that visually.

- **Video Title:** (4 of 4) Heteroskedasticity: Estimation using the White standard errors

- **Channel:** jbstatistics

- **Link:** https://www.youtube.com/watch?v=9ujhcWwMlxw

- **Relation to Text:** This video discusses **White's standard errors** (also called heteroskedasticity-consistent or robust standard errors), which provide valid inference even when heteroskedasticity is present. This directly corresponds to the content surrounding **Theorem 22.5**, and equations (22.2) and (22.3) where t_R and W_R (robust test statistics) are defined. The video explains the concept of robust standard errors and shows how they are calculated.

4. Law of Large Numbers and Central Limit Theorem

- **Video Title:** (Weak) Law of Large Numbers

- **Channel:** jbstatistics

- **Link:** https://www.youtube.com/watch?v=MntX3zWNWec

- **Relation to Text:** This video explains the **Weak Law of Large Numbers**, which is crucial for understanding the consistency of the OLS estimator (Theorem 22.1). The proofs in the text rely on the WLLN.

- **Video Title:** Central Limit Theorem

- **Channel:** jbstatistics

- **Link:** https://www.youtube.com/watch?v=YAlJCEDH2uY

- **Relation to Text:** This video presents the **Central Limit Theorem (CLT)**, another fundamental concept underlying the asymptotic normality of the OLS estimator (Theorem 22.2). The proofs and discussions in the text heavily rely on the CLT.

5. Slutsky's Theorem

- **Video Title:** Slutsky's Theorem
- **Channel:** Modus Ponens
- **Link:** https://www.youtube.com/watch?v=tMHEf0c7B8k
- **Relation to Text:** This video explains **Slutsky's Theorem**, which is used extensively in the proofs and derivations throughout the chapter, particularly in Theorems 22.1, 22.2, and 22.4. The video provides the statement of the theorem and illustrates its use with examples.

6. Lindeberg Central Limit Theorem (Advanced)

- **Video Title:** Lindeberg-Feller Central Limit Theorem
- **Channel:** Murtaza Ali
- **Link:** https://www.youtube.com/watch?v=SYLD2uJk_ts
- **Relation to Text:** This video discusses the **Lindeberg-Feller Central Limit Theorem**, which is a more general version of the CLT that applies to independent but *not* identically distributed random variables. This is directly relevant to **Theorem 22.7**, which deals with the non-i.i.d. case. The proof of Theorem 22.7 invokes the Lindeberg CLT. This is a more advanced topic.

These videos provide a good complement to the textbook chapter, offering visual explanations and examples to solidify understanding of the key concepts. They range in difficulty, with the Lindeberg CLT video being the most mathematically advanced.

Multiple Choice Exercises

MC Exercise 1

MC Solution 1

Under the i.i.d. assumption, the matrix $M = E[x_i x_i^T]$ represents:

a. The variance-covariance matrix of the error terms.
b. The expected outer product of the regressor vector with itself.
c. The inverse of the variance-covariance matrix of the OLS estimator.
d. The expected value of the dependent variable.

MC Exercise 2

MC Solution 2

The condition $E[\varepsilon_i^2|x_i] = \sigma^2$ is known as:

a. Linearity
b. Homoskedasticity
c. Normality
d. Exogeneity

MC Exercise 3

MC Solution 3

The unconditional moment condition $E[x_i\varepsilon_i] = 0$ implies:

a. The OLS estimator is unbiased.
b. The regressors and the error term are uncorrelated.
c. The error term has a mean of zero.
d. The OLS estimator is BLUE.

MC Exercise 4

MC Solution 4

Which of the following conditional moment conditions ensures the unbiasedness of the OLS estimator?

a. $E[\varepsilon_i] = 0$
b. $E[x_i] = 0$
c. $E[\varepsilon_i|x_i] = 0$
d. $E[x_i\varepsilon_i] = 0$

MC Exercise 5

MC Solution 5

Theorem 22.1 ($\hat{\beta} \xrightarrow{p} \beta$) states that the OLS estimator is:

a. Unbiased
b. Consistent
c. Normally distributed
d. Efficient

MC Exercise 6

MC Solution 6

In the proof of Theorem 22.1, Slutsky's theorem is used to:

a. Show that the sample mean converges to the population mean.
b. Combine the convergence results of two separate terms.
c. Prove the Central Limit Theorem.
d. Demonstrate the unbiasedness of the OLS estimator.

MC Exercise 7

MC Solution 7

Theorem 22.2 states that in large samples, the distribution of $\sqrt{n}(\hat{\beta} - \beta)$ is approximately:

a. Uniform
b. Chi-squared
c. Normal
d. t-distributed

MC Exercise 8

MC Solution 8

Under homoskedasticity, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ simplifies to:

a. $M^{-1}\Omega M^{-1}$
b. $\sigma^2 M^{-1}$
c. $\sigma^2 \Omega^{-1}$
d. M^{-1}

MC Exercise 9

MC Solution 9

Theorem 22.3 ($s_*^2 \xrightarrow{p} \sigma^2$) is important for:

a. Proving the consistency of the OLS estimator.
b. Constructing confidence intervals and hypothesis tests.
c. Demonstrating the unbiasedness of the OLS estimator.
d. Deriving the asymptotic distribution of the OLS estimator.

MC Exercise 10

MC Solution 10(#sec-ch22mcsolution10}

If the homoskedasticity assumption is violated, the usual t-statistic:

a. Still converges to a standard normal distribution.
b. Converges to a t-distribution with n-K degrees of freedom.
c. Does not converge to a standard normal distribution.
d. Converges to a chi-squared distribution.

MC Exercise 11

MC Solution 11

Robust test statistics (t_R and W_R) are used to:

a. Obtain valid inference under heteroskedasticity.
b. Improve the efficiency of the OLS estimator.
c. Ensure the unbiasedness of the OLS estimator.
d. Simplify the calculation of standard errors.

MC Exercise 12

MC Solution 12

The moment conditions in Theorem 22.5, such as $E|x_{ij}x_{ik}x_{il}x_{ir}| < \infty$, are necessary to: (a) Ensure OLS estimator is unbiased. (b) Ensure the consistency of the heteroskedasticity-robust variance estimator. (c) Ensure that regressors are non-stochastic. (d) Ensure that OLS estimator is BLUE.

MC Exercise 13

MC Solution 13

The non-i.i.d. case considered in section 22.2 allows for:

a. Only independent and identically distributed observations.
b. Observations that are not necessarily independent or identically distributed.
c. Only normally distributed error terms.
d. Only homoskedastic error terms.

MC Exercise 14

MC Solution 14

In Theorem 22.6, the condition $\lambda_{max}(X^T\Sigma X)\lambda_{min}^{-2}(X^TX) \to 0$ ensures:

a. Unbiasedness of the OLS estimator.
b. Consistency of the OLS estimator in the non-i.i.d. case.
c. Normality of the error terms.
d. Homoskedasticity of the error terms.

MC Exercise 15

MC Solution 15

In Example 22.2, the OLS estimator of the dummy variable coefficient is inconsistent when the number of observations in the set I is fixed because:

a. The error terms are heteroskedastic.
b. The regressors are stochastic.
c. Information about that coefficient does not grow with sample size.
d. The model is misspecified.

MC Exercise 16

MC Solution 16

Assumption R2 in Theorem 22.7 includes the condition $\frac{\max_{1\le i\le n} x_{ij}^2}{\sum_{i=1}^n x_{ij}^2} \to 0$. This condition is known as:

a. The homoskedasticity condition
b. The Lyapunov condition
c. The no single observation dominates condition
d. The full rank condition.

MC Exercise 17

MC Solution 17

The Lindeberg CLT is used in the proof of Theorem 22.7 because:

a. The error terms are normally distributed.
b. The regressors are i.i.d.
c. The random variables being summed are independent but not necessarily identically distributed.
d. The OLS estimator is unbiased.

MC Exercise 18

MC Solution 18

In Example 22.6, the scaling matrix Δ is used because:

a. The regressors are heteroskedastic.
b. The regressors grow at different rates.
c. The error terms are not normally distributed.
d. The OLS estimator is biased.

MC Exercise 19

MC Solution 19

Assumptions R4 and R5 in Theorem 22.8 are needed to ensure:

a. The OLS estimator is unbiased
b. The consistency of the robust variance estimator in the non-i.i.d. case.
c. The normality of error terms.
d. The homoscedasticity of error terms

MC Exercise 20

MC Solution 20

In Example 22.10, the t-statistic is driven only by the first component when the matrix $R\Delta^{-1}$ is asymptotically singular. This means that: (a) All regressors contribute equally. (b) The fastest-growing regressors dominate. (c) The slowest-growing regressor dominates. (d) The error terms are homoscedastic.

Multiple Choice Solutions

MC Solution 1

MC Exercise 1

(b) The expected outer product of the regressor vector with itself.

As explained in the text, x_i is a $K \times 1$ vector, so $x_i x_i^T$ is a $K \times K$ matrix containing the squares and cross-products of the regressors. M is the expected value of this matrix.

MC Solution 2

MC Exercise 2

(b) Homoskedasticity

This is the definition of homoskedasticity: the variance of the error term is constant and does not depend on the regressors.

MC Solution 3

MC Exercise 3

(b) The regressors and the error term are uncorrelated.

This is the definition of zero correlation. While $E[\varepsilon_i] = 0$ is often assumed, it's not *implied* by $E[x_i \varepsilon_i] = 0$. Unbiasedness and the BLUE property require stronger assumptions.

MC Solution 4

MC Exercise 4

(c) $E[\varepsilon_i | x_i] = 0$

This is the conditional moment condition that, along with the other standard assumptions, guarantees the unbiasedness of the OLS estimator, as shown in the text.

MC Solution 5

MC Exercise 5

(b) Consistent

Consistency means that the estimator converges in probability to the true parameter value as the sample size increases.

MC Solution 6

MC Exercise 6

(b) Combine the convergence results of two separate terms.

Slutsky's theorem allows us to combine the convergence in probability of $(X^TX/n)^{-1}$ to M^{-1} and the convergence in probability of $X^T\varepsilon/n$ to 0.

MC Solution 7

MC Exercise 7

(c) Normal

This is the statement of the asymptotic normality of the OLS estimator.

MC Solution 8

MC Exercise 8

(b) $\sigma^2 M^{-1}$

Under homoskedasticity, $\Omega = \sigma^2 M$, which simplifies the general expression for the asymptotic variance.

MC Solution 9

MC Exercise 9

(b) Constructing confidence intervals and hypothesis tests.

s_*^2 is used to estimate the variance of the error term, which is needed to calculate standard errors and construct test statistics.

MC Solution 10

MC Exercise 10

(c) Does not converge to a standard normal distribution.

The violation of homoskedasticity invalidates the usual standard errors and, consequently, the usual t-statistic.

MC Solution 11

MC Exercise 11

(a) Obtain valid inference under heteroskedasticity.

Robust standard errors provide consistent estimates of the variance-covariance matrix of the OLS estimator, even when heteroskedasticity is present.

MC Solution 12

MC Exercise 12

(b) Ensure the consistency of the heteroskedasticity-robust variance estimator. These are moment conditions. They ensure that sample averages converge to population quantities.

MC Solution 13

MC Exercise 13

(b) Observations that are not necessarily independent or identically distributed.

The non-i.i.d. case allows for situations like heteroskedasticity, autocorrelation, and deterministic trends in the regressors.

MC Solution 14

MC Exercise 14

(b) Consistency of the OLS estimator in the non-i.i.d. case.

This condition ensures that the variance in the error terms doesn't grow too fast relative to the information in the data, which is necessary for consistency.

MC Solution 15

MC Exercise 15

(c) Information about that coefficient does not grow with sample size.

With a fixed number of observations for the dummy variable, increasing the overall sample size doesn't provide more information about the effect of that specific dummy variable.

MC Solution 16

MC Exercise 16

(c) The no single observation dominates condition This condition prevents any single observation from having too large an influence on the OLS estimator as the sample gets larger.

MC Solution 17

MC Exercise 17

(c) The random variables being summed are independent but not necessarily identically distributed.

The standard CLT requires i.i.d. random variables. The Lindeberg CLT generalizes this to independent but not necessarily identically distributed variables, which is needed for the non-i.i.d. case.

MC Solution 18

MC Exercise 18

(b) The regressors grow at different rates.

The scaling matrix ensures that all regressors, after scaling, contribute to the asymptotic distribution.

MC Solution 19

MC Exercise 19

(b) The consistency of the robust variance estimator in the non-i.i.d. case. These assumptions ensure the sample averages of cross-products converge to their population counterparts, which is needed for consistency.

MC Solution 20

MC Exercise 20

(c) The slowest-growing regressor dominates. When the matrix is asymptotically singular, it means restrictions across variables with different trend rates are effectively dominated by slowest trend rate.

Author: Peter Fuleky

This book was built with Quarto