Chapter 3
Multivariate Limit Theorems

This chapter discusses multivariate limit theorems, and follows Olive (2014, § 8.6, 8.7) closely.

3.1 Multivariate Limit Theorems

Many of the univariate results from Chapter 2 can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 3.1. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 3.1, 3.2 and 3.3 below are the multivariate extensions of the limit theorems in Section 2.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$ and $V(X) = \boldsymbol{\Sigma} = \sigma^2$.

Theorem 3.1: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The MCLT is proven after Theorem 3.8. To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 3.2: the Multivariate Delta Method. If

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

if $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular, where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Example 3.1. If Y has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\phi}{\lambda} y^{\phi - 1} e^{-\frac{y^\phi}{\lambda}}$$

where λ, y, and ϕ are all positive. If $\mu = \lambda^{1/\phi}$ so $\mu^\phi = \lambda$, then the Weibull pdf

$$f(y) = \frac{\phi}{\mu}\left(\frac{y}{\mu}\right)^{\phi - 1} \exp\left[-\left(\frac{y}{\mu}\right)^\phi\right].$$

Let $(\hat{\mu}, \hat{\phi})$ be the MLE of (μ, ϕ). According to Bain (1978, p. 215),

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \mu \\ \phi \end{pmatrix}\right) \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.109\frac{\mu^2}{\phi^2} & 0.257\mu \\ 0.257\mu & 0.608\phi^2 \end{pmatrix}\right)$$

$= N_2(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$ where $\boldsymbol{I}(\boldsymbol{\theta})$ is given in Definition 3.2.

Let column vectors $\boldsymbol{\theta} = (\mu \ \ \phi)^T$ and $\boldsymbol{\eta} = (\lambda \ \ \phi)^T$. Then

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}) = \begin{pmatrix} \lambda \\ \phi \end{pmatrix} = \begin{pmatrix} \mu^{\phi} \\ \phi \end{pmatrix} = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \end{pmatrix}.$$

So

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_1(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} g_2(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_2(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mu} \mu^{\phi} & \frac{\partial}{\partial \phi} \mu^{\phi} \\ \frac{\partial}{\partial \mu} \phi & \frac{\partial}{\partial \phi} \phi \end{bmatrix} = \begin{bmatrix} \phi \mu^{\phi - 1} & \mu^{\phi} \log(\mu) \\ 0 & 1 \end{bmatrix}.$$

Thus by the multivariate delta method,

$$\sqrt{n}\left(\begin{pmatrix} \hat{\lambda} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \lambda \\ \phi \end{pmatrix} \right) \xrightarrow{D} N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where (see Definition 3.4 below)

$$\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\eta})^{-1} = [\boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\theta}))]^{-1} = \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})} =$$

$$\begin{bmatrix} 1.109\lambda^2(1 + 0.4635 \log(\lambda) + 0.5482(\log(\lambda))^2) & 0.257\phi\lambda + 0.608\lambda\phi \log(\lambda) \\ 0.257\phi\lambda + 0.608\lambda\phi \log(\lambda) & 0.608\phi^2 \end{bmatrix}.$$

Definition 3.2. Let X be a random variable with pdf or pmf $f(x|\boldsymbol{\theta})$. Then the **information matrix**

$$\boldsymbol{I}(\boldsymbol{\theta}) = [\boldsymbol{I}_{i,j}]$$

where

$$\boldsymbol{I}_{i,j} = E\left[\frac{\partial}{\partial \theta_i} \log(f(X|\boldsymbol{\theta})) \frac{\partial}{\partial \theta_j} \log(f(X|\boldsymbol{\theta})) \right].$$

Definition 3.3. An estimator $\boldsymbol{T}_n$ of $\boldsymbol{\theta}$ is **asymptotically efficient** if

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})).$$

Following Lehmann (1999, p. 511), if $\boldsymbol{T}_n$ is asymptotically efficient and if the estimator $\boldsymbol{W}_n$ satisfies

$$\sqrt{n}(\boldsymbol{W}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\theta}))$$

where $\boldsymbol{J}(\boldsymbol{\theta})$ and $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ are continuous functions of $\boldsymbol{\theta}$, then under regularity conditions, $\boldsymbol{J}(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})$ is a positive semidefinite matrix, and $\boldsymbol{T}_n$ is "better" than $\boldsymbol{W}_n$.

Definition 3.4. Assume that $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$. Then

$$\boldsymbol{I}(\boldsymbol{\eta}) = \boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\theta})) = [\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}]^{-1}.$$

Notice that this definition agrees with the multivariate delta method if

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \boldsymbol{I}^{-1}(\boldsymbol{\theta})$.

Now suppose that $X_1, ..., X_n$ are iid random variables from a k-parameter REF

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right] \tag{3.1}$$

with natural parameterization

$$f(x|\boldsymbol{\eta}) = h(x)b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^{k} \eta_i t_i(x)\right]. \tag{3.2}$$

Then the complete minimal sufficient statistic is

$$\overline{\boldsymbol{T}}_n = \frac{1}{n}(\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i))^T.$$

Let $\boldsymbol{\mu}_T = (E(t_1(X), ..., E(t_k(X)))^T$. From Theorem 1.31, for $\boldsymbol{\eta} \in \Omega$,

$$E(t_i(X)) = \frac{-\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})),$$

and

$$\mathrm{Cov}(\mathrm{t_i(X)}, \mathrm{t_j(X)}) \equiv \sigma_{\mathrm{i,j}} = \frac{-\partial^2}{\partial \eta_{\mathrm{i}} \partial \eta_{\mathrm{j}}} \log(\mathrm{b}(\boldsymbol{\eta})).$$

Theorem 3.3. If the random variable X is a kP–REF with pmf or pdf (3.2), then the information matrix

$$\boldsymbol{I}(\boldsymbol{\eta}) = [\boldsymbol{I}_{i,j}]$$

where

$$\boldsymbol{I}_{i,j} = E\left[\frac{\partial}{\partial \eta_i} \log(f(X|\boldsymbol{\eta})) \frac{\partial}{\partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right] = -E\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right].$$

Several authors, including Barndorff–Nielsen (1982), have noted that the multivariate CLT can be used to show that $\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. The fact that $\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\eta})$ appears in Lehmann (1983, p. 127).

Theorem 3.4. If $X_1, ..., X_n$ are iid from a k-parameter regular exponential family, then

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\eta})).$$

Proof. By the multivariate central limit theorem,

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = [\sigma_{i,j}]$. Hence the result follows if $\sigma_{i,j} = \boldsymbol{I}_{i,j}$. Since

$$\log(f(x|\boldsymbol{\eta})) = \log(h(x)) + \log(b(\boldsymbol{\eta})) + \sum_{l=1}^{k} \eta_l t_l(x),$$

$$\frac{\partial}{\partial \eta_i} \log(f(x|\boldsymbol{\eta})) = \frac{\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})) + t_i(X).$$

Hence

$$-\boldsymbol{I}_{i,j} = E\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\boldsymbol{\eta})) = -\sigma_{i,j}. \quad \square$$

To obtain standard results, use the multivariate delta method, assume that both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are $k \times 1$ vectors, and assume that $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$ is a one to one mapping so that the inverse mapping is $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$. If $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular, then

$$\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})} \tag{3.3}$$

(see Searle 1982, p. 339), and

$$\boldsymbol{I}(\boldsymbol{\eta}) = [\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}]^{-1} = [\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})}]^T \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})} \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}. \tag{3.4}$$

Compare Lehmann (1999, p. 500) and Lehmann (1983, p. 127).

For example, suppose that $\boldsymbol{\mu}_T$ and $\boldsymbol{\eta}$ are $k \times 1$ vectors, and

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta}))$$

where $\boldsymbol{\mu}_T = \boldsymbol{g}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)$. Also assume that $\overline{\boldsymbol{T}}_n = \boldsymbol{g}(\hat{\boldsymbol{\eta}})$ and $\hat{\boldsymbol{\eta}} = \boldsymbol{g}^{-1}(\overline{\boldsymbol{T}}_n)$. Then by the multivariate delta method and Theorem 3.4,

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) = \sqrt{n}(\boldsymbol{g}(\hat{\boldsymbol{\eta}}) - \boldsymbol{g}(\boldsymbol{\eta})) \xrightarrow{D} N_k[\mathbf{0}, \boldsymbol{I}(\boldsymbol{\eta})] = N_k[\mathbf{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})} \boldsymbol{I}^{-1}(\boldsymbol{\eta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})}].$$

Hence

$$\boldsymbol{I}(\boldsymbol{\eta}) = \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})} \boldsymbol{I}^{-1}(\boldsymbol{\eta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})}.$$

Similarly,

$$\sqrt{n}(\boldsymbol{g}^{-1}(\overline{\boldsymbol{T}}_n) - \boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)) = \sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_k[\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta})] =$$

$$N_k[\mathbf{0}, \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)} \boldsymbol{I}(\boldsymbol{\eta}) \boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}].$$

Thus

$$\boldsymbol{I}^{-1}(\boldsymbol{\eta}) = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)} \boldsymbol{I}(\boldsymbol{\eta}) \boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)} \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})} \boldsymbol{I}^{-1}(\boldsymbol{\eta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})} \boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}$$

as expected by Equation (3.4). Typically $\hat{\boldsymbol{\theta}}$ is a function of the sufficient statistic $\boldsymbol{T}_n$ and is the unique MLE of $\boldsymbol{\theta}$. Replacing $\boldsymbol{\eta}$ by $\boldsymbol{\theta}$ in the above discussion shows that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$ is equivalent to $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}))$ provided that $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular.

3.2 More Multivariate Results

Definition 3.5. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 3.5. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 3.6. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$ and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 3.7: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic function $c_n(\boldsymbol{t})$ and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $c(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \mathrm{c_n}(\boldsymbol{t}) \to \mathrm{c}(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 3.8: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Application: Proof of the MCLT Theorem 3.1. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance $\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T \sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})$. The right hand side has

distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\square$

Theorem 3.9. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 3.6. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 3.10. Let W_n, X_n, Y_n and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Recall that the smallest integer function $\lceil x \rceil$ rounds up, e.g. $\lceil 7.7 \rceil = 8$.

Definition 3.7. The *sample ρ quantile* $\hat{y}_{n,\rho} = \hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. The *population quantile* $y_\rho = \xi_\rho = Q(\rho) = \inf\{y : F(y) \geq \rho\}$.

There are many other ways to define sample quantiles, and the different estimators tend to be asymptotically equivalent. If the inverse F^{-1} of the cdf exists, then $Q(u) = F^{-1}(u)$. $Q(u) \leq x$ iff $u \leq F(x)$. $F(y_\rho) = P(Y \leq y_\rho) \geq \rho$ and $P(Y \geq y_\rho) \geq 1 - \rho$. Let the observed data be $Y_1, ..., Y_n$, and let $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$. Then $\hat{Q}(\rho) = \inf\{y : \hat{F}(y) \geq \rho\} = \hat{y}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. (An alternative definition of the population quantile that is often used is that y_ρ is any real number satisfying $P(Y \leq y_\rho) \geq \rho$ and $P(Y \geq y_\rho) \geq 1 - \rho$. Then y_ρ is not necessarily unique. Definition 3.7 makes the population quantile

unique. The regularity conditions in Theorem 3.11 make y_ρ unique if the alternative definition is used.)

Theorem 3.11: Serfling (1980, p. 80). Let $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$. Suppose that F has a density f that is positive and continuous in neighborhoods of $\xi_{\rho_1}, ..., \xi_{\rho_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\rho_1}, ..., \hat{\xi}_{n,\rho_k})^T - (\xi_{\rho_1}, ..., \xi_{\rho_k})^T] \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})$ and

$$\sigma_{ij} = \frac{\rho_i(1-\rho_j)}{f(\xi_{\rho_i})f(\xi_{\rho_j})}$$

for $i \leq j$ and $\sigma_{ij} = \sigma_{ji}$ for $i > j$.

Theorem 3.12: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous and does not depend on n, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 3.13. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1}\boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 3.14. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 3.15. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$. □

Example 3.2. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by Theorem 3.7. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1) \phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1) \phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 3.12.

Remark 3.1. In the above example, we can show $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ instead of assuming $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. See Ferguson (1996, p. 42).

3.3 Summary

3.4 Complements

Suppose $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$. In analysis, the fact that

$$\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}$$

is a corollary of the inverse mapping theorem (or of the inverse function theorem). See Apostol (1957, p. 146), Bickel and Doksum (2007, p. 517), Marsden and Hoffman (1993, p. 393) and Wade (2000, p. 353).

3.5 Problems

3.1. Many multiple linear regression estimators $\hat{\boldsymbol{\beta}}$ satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{3.5}$$

when

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}, \tag{3.6}$$

and when the errors e_i are iid with a cdf F and a unimodal pdf f that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

$$V(OLS, F) = V(e_i) = \sigma^2 \ \text{ while } \ \mathrm{V(L_1, F)} = \frac{1}{4[\mathrm{f(0)}]^2}.$$

In the multiple linear regression model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{3.7}$$

for $i = 1, \ldots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.8}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

a) What is the ijth element of the matrix

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}?$$

b) Suppose $x_{k,1} = 1$ and that $x_{k,j} \sim X_j$ are iid with $E(X_j) = 0$ and $V(X_j) = 1$ for $k = 1, ..., n$ and $j = 2, ..., p$. Assume that X_i and X_j are independent for $i \neq j$, $i > 1$ and $j > 1$. (Often $x_{k,j} \sim N(0, 1)$ in simulations.) Then what is $\boldsymbol{W}^{-1}$ for model (3.7)?

c) Suppose $p = 2$ and $Y_i = \alpha + \beta X_i + e_i$. Show

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2} & \frac{-\sum X_i}{n\sum(X_i - \overline{X})^2} \\ \frac{-\sum X_i}{n\sum(X_i - \overline{X})^2} & \frac{n}{n\sum(X_i - \overline{X})^2} \end{bmatrix}.$$

d) Under the conditions of c), let $S_x^2 = \sum(X_i - \overline{X})^2/n$. Show that

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} = \begin{bmatrix} \frac{\frac{1}{n}\sum X_i^2}{S_x^2} & \frac{-\overline{X}}{S_x^2} \\ \frac{-\overline{X}}{S_x^2} & \frac{1}{S_x^2} \end{bmatrix}.$$

e) If the X_i are iid with variance $V(X)$ then $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{W}$. What is $\boldsymbol{W}$?

f) Now suppose that n is divisible by 5 and the $n/5$ of X_i are at 0.1, $n/5$ at 0.3, $n/5$ at 0.5, $n/5$ at 0.7 and $n/5$ at 0.9. (Hence if $n = 100$, 20 of the X_i are at 0.1, 0.3, 0.5, 0.7 and 0.9.)

Find $\sum X_i^2/n$, $\overline{X}$ and S_x^2. (Your answers should not depend on n.)

g) Under the conditions of f), estimate $V(\hat{\alpha})$ and $V(\hat{\beta})$ if L_1 is used and if the e_i are iid $N(0, 0.01)$.

Hint: Estimate $\boldsymbol{W}$ with $n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $V(\hat{\boldsymbol{\beta}}, F) = V(L_1, F) = \frac{1}{4[f(0)]^2}$. Hence

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \approx N_2\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \frac{1}{n}\frac{1}{4[f(0)]^2}\begin{pmatrix} \frac{\frac{1}{n}\sum X_i^2}{S_x^2} & \frac{-\overline{X}}{S_x^2} \\ \frac{-\overline{X}}{S_x^2} & \frac{1}{S_x^2} \end{pmatrix}\right].$$

You should get an answer like $0.0648/n$.