Nonlinear Least Square Estimates and Estimations in ARMA Models

Ching-Kang Ing

National Tsing Hua University

cking@stat.nthu.edu.tw

Overview

Introduction to Nonlinear Least Square Estimates

- Let

$$y_i = g_i(\boldsymbol{\theta_0}) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\theta_0}$ is the true parameter belonging to a parameter space $\boldsymbol{\Theta}$, $g_i(\boldsymbol{\theta_0})$ is specified up to the parameter $\boldsymbol{\theta_0}$ (*i.e.*, $g_i(\cdot)$ is known but $\boldsymbol{\theta}_0$ is unknown) and is independent of ϵ_i, and ϵ_i are i.i.d random noises with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

- The nonlinear LSE $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta_0}$ is given by

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\mathrm{argmin}} S_n(\boldsymbol{\theta}),$$

where $S_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - g_i(\boldsymbol{\theta}))^2$. Moveover, σ^2 can be estimated by

$$\hat{\sigma}_n^2 = \frac{1}{n} S_n(\hat{\boldsymbol{\theta}}_n).$$

- In the following , we shall focus on the asymptotic behaviors of $\hat{\boldsymbol{\theta}}_n$.

Theorem

Assume

(C1) *For any* $\epsilon > 0$ *and* $\delta > 0$, *there exists* $\eta > 0$ *s.t.*

$$\Pr\left(\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} \frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_0))^2 < \eta\right) < \delta,$$

holds for all large n, where $\boldsymbol{\Theta}_\epsilon = \boldsymbol{\Theta} - \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon\}$.

(C2) *For any* $\epsilon, \delta_1 > 0$,

$$\Pr\left(\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} \left|\frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_0))\epsilon_i\right| > \delta_1\right) \to 0, \quad \text{as } n \to \infty.$$

Then, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$.

- **Remark. (C1)** and **(C2)** are easily satisfied in the AR(1) case in which $g_i(\theta) = \theta y_{i-1}$.

Proof. Some key facts (please check by yourself)

(1) $S_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} {\epsilon_i}^2 - 2\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))\epsilon_i + \sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))^2$.

(2) By the definition of $\hat{\boldsymbol{\theta}}_n$,

$$\Pr\left(\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} S_n(\boldsymbol{\theta}) > S_n(\boldsymbol{\theta_0})\right) \leq \Pr(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}\| < \epsilon).$$

Therefore, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ is guaranteed by for any $\epsilon > 0$,

$$\Pr\left(\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} S_n(\boldsymbol{\theta}) \leq S_n(\boldsymbol{\theta_0})\right) \to 0,$$

as $n \to \infty$.

(3) Using key fact (1),

$$\Pr\left(\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} S_n(\boldsymbol{\theta}) \le S_n(\boldsymbol{\theta_0})\right)$$

$$\le \Pr\left(\frac{2}{n}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon}\left|\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))\epsilon_i\right| \ge \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon}\frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))^2\right).$$

Therefore, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ is, in turn, implied by for any $\epsilon > 0$,

$$\Pr\left(\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon}\left|\frac{2}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))\epsilon_i\right| \ge \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon}\frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta_0}))^2\right) \to 0, \tag{1}$$

as $n \to \infty$. In the appendix given at the end of this note, we will show that this is ensured by **(C1)** and **(C2)**, which is the required result. □

Theorem

Assume

(A1) $n^{-1}\sum_{i=1}^{n}\nabla g_i(\boldsymbol{\theta_0})\nabla g_i^{\top}(\boldsymbol{\theta_0}) \xrightarrow{p} \mathbf{C}$, *where* $\mathbf{C}$ *is a positive definite non-random matrix.*

(A2) $n^{-1}\sum_{i=1}^{n}\nabla^2 g_i(\boldsymbol{\theta_0})\epsilon_i \xrightarrow{p} \mathbf{0}$.

(A3) $\{\nabla g_i(\boldsymbol{\theta_0})\varepsilon_i\}$ *is a martingale difference sequence obeying the conditional Lindeberg condition.*

Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{C}^{-1}\sigma^2).$$

Proof. We now give a "heuristic" argument. Again, some key facts,

(1) $\mathbf{0} = \nabla S_n(\hat{\boldsymbol{\theta}}_n)$, provided $S_n(\boldsymbol{\theta})$ is sufficiently smooth and $\boldsymbol{\theta}_{\mathbf{0}}$ is not on the boundary of $\boldsymbol{\Theta}$, where $\nabla S_n(\boldsymbol{\theta}) = (\frac{\partial S_n(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial S_n(\boldsymbol{\theta})}{\partial \theta_k})^\top$, and k is the number of parameters.

(2) By Taylor's theorem and the consistency of $\hat{\boldsymbol{\theta}}_n$,

$$\mathbf{0} \mathrel{\dot\sim} \nabla S_n(\boldsymbol{\theta}_{\mathbf{0}}) + \nabla^2 S_n(\boldsymbol{\theta}_{\mathbf{0}})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}}),$$

where $\nabla^2 S_n(\boldsymbol{\theta}_{\mathbf{0}}) = \left(\frac{\partial^2 S_n(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right)_{1 \le i,j \le k}$.

(3) By fact (2),

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &\mathrel{\dot\sim} \left(\frac{1}{n}\nabla^2 S_n(\boldsymbol{\theta}_{\mathbf{0}})\right)^{-1}\left(-\frac{1}{n}\nabla S_n(\boldsymbol{\theta}_{\mathbf{0}})\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \nabla g_i(\boldsymbol{\theta}_{\mathbf{0}})\nabla g_i^\top(\boldsymbol{\theta}_{\mathbf{0}}) - \frac{1}{n}\sum_{i=1}^n \nabla^2 g_i(\boldsymbol{\theta}_{\mathbf{0}})\varepsilon_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla g_i(\boldsymbol{\theta}_{\mathbf{0}})\varepsilon_i.\end{aligned}$$

Then the required result follows from the assumptions. □

- **Application to the MA(1) model.** Consider an MA(1) model,

$$y_t = \epsilon_t - \theta_0 \epsilon_{t-1}.$$

- To estimate θ_0, we use the conditional least squares estimate(CLS),

$$\hat{\theta}_0 = \underset{\theta \in [-1+\delta, 1-\delta]}{\text{argmin}} \; S_n(\theta),$$

where δ is an arbitrarily small positive number and $S_n(\theta) = \sum_{t=1}^{n} \epsilon_t^2(\theta)$ with

$$\varepsilon_t(\theta) = \frac{1}{1-\theta B} y_t = y_t - (-\theta y_{t-1} - \theta^2 y_{t-2} - \cdots) = y_t - g_t(\theta).$$

- **(Consistency)** $\hat{\theta}_n \to \theta$, since **(C1)** and **(C2)** are satisfied.
- **(CLT)** According to (1),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, C^{-1}\sigma^2),$$

where

$$C = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(g_t'(\theta_0)\right)^2 = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\varepsilon_t'(\theta_0)\right)^2.$$

Notice that when $X_n \xrightarrow{p} c$, we denote it by $\text{plim}_{n\to\infty} X_n = c$. To compute C, note that

$$\varepsilon_t'(\theta_0) = \frac{By_t}{(1-\theta_0 B)^2} = \frac{y_{t-1}}{(1-\theta_0 B)^2} = \frac{\varepsilon_{t-1}}{1-\theta_0 B} = \sum_{j=0}^{\infty} {\theta_0}^j \varepsilon_{t-1-j},$$

which is an AR(1) process with AR coefficient θ_0. Therefore, $C = \sigma^2/(1-\theta_0^2)$, and consequently, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1-\theta_0^2)$.

- **Remark 1.** Since $y_0, y_{-1} \cdots$ are unobservable in practice, their values are set to 0. This is why the name "conditional" is given. In our asymptotic analysis, we still "pretend" $y_0, y_{-1}, \ldots$ are observed because this can substantially simplify the mathematical derivation. The difference between these two will vanish asymptotically.
- **Reamrk 2.** For the AR(1) model, $y_t = \phi_0 y_{t-1} + \epsilon_t$, we have already shown that the LSE $\hat{\phi}_n$ of ϕ_0 satisfies

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N(0, \sigma^2/\gamma_0) = N(0, 1 - \phi_0^2).$$

- **Application to ARMA(1,1) model.** Consider a ARMA(1,1) model,

$$y_t - \phi_0 y_{t-1} = \epsilon_t - \theta_0 \epsilon_{t-1}.$$

- We have the following facts

(1) $\varepsilon_t(\phi, \theta) = (\frac{1-\phi B}{1-\theta B}) y_t$.
(2) $S_n(\phi, \theta) = \sum_{t=1}^{n} \varepsilon_t^2(\phi, \theta)$.
(3) The CLS estimate of (ϕ_0, θ_0) is given by

$$(\hat{\phi}_n, \hat{\theta}_n) = \operatorname{argmin}_{(\phi,\theta)\in\boldsymbol{\Theta}} S_n(\phi, \theta),$$

where $\boldsymbol{\Theta} = [-1+\delta, 1-\delta] \times [-1+\delta, 1-\delta]$.
(4) The limiting distribution of $(\hat{\phi}_n, \hat{\theta}_n)$ is

$$\sqrt{n}\left[\begin{pmatrix} \hat{\phi}_n \\ \hat{\theta}_n \end{pmatrix} - \begin{pmatrix} \phi_0 \\ \theta_0 \end{pmatrix} \right] \xrightarrow{d} N(\mathbf{0}, \mathbf{C}^{-1}\sigma^2),$$

where $\mathbf{C} = \operatorname{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \nabla \varepsilon_t(\phi_0, \theta_0) \nabla \varepsilon_t^{\top}(\phi_0, \theta_0)$.

- To find $\mathbf{C}$, we note that

$$\frac{\partial \varepsilon_t(\phi_0, \theta_0)}{\partial \phi} = \frac{-y_{t-1}}{1-\theta_0 B} = \frac{-\varepsilon_{t-1}}{1-\phi_0 B} := W_{t-1},$$
$$\frac{\partial \varepsilon_t(\phi_0, \theta_0)}{\partial \theta} = \frac{(1-\phi_0 B) y_{t-1}}{(1-\theta_0 B)^2} = \frac{\varepsilon_{t-1}}{1-\theta_0 B} := S_{t-1},$$

and hence

$$\mathbf{C} = \mathrm{E}\left[\nabla \varepsilon_t(\boldsymbol{\theta_0}) \nabla \varepsilon_t^\top(\boldsymbol{\theta_0})\right] = \begin{pmatrix} \sigma^2/(1-{\phi_0}^2) & -\sigma^2/(1-\phi_0\theta_0) \\ -\sigma^2/(1-\phi_0\theta_0) & \sigma^2/(1-{\theta_0}^2) \end{pmatrix}.$$

Notice that

$$\mathrm{E}(W_{t-1} S_{t-1}) = \mathrm{E}\left[-\left(\sum_{j=0}^{\infty} {\phi_0}^j \varepsilon_{t-1-j}\right)\left(\sum_{j=0}^{\infty} {\theta_0}^j \varepsilon_{t-1-j}\right)\right]$$
$$= -\sum_{j=0}^{\infty} {\phi_0}^j {\theta_0}^j \sigma^2 = -\frac{\sigma^2}{1-\phi_0\theta_0}.$$

- **Application to ARMA(p, q) model.** Consider a ARMA(p, q) model,

$$Z_t - \phi_{1,0} Z_{t-1} - \cdots - \phi_{p,0} Z_{t-p} = \varepsilon_t - \theta_{1,0}\varepsilon_{t-1} - \cdots - \theta_{q,0}\varepsilon_{t-q},$$

- Define

$$\varepsilon_t(\boldsymbol{\eta}) = \frac{1 - \phi_1 B - \cdots - \phi_p B^p}{1 - \theta_1 B - \cdots - \theta_q B^q} Z_t,$$

where $\boldsymbol{\eta} = (\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q)^\top$. The CLS estimate of $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{\eta}}_n = \underset{\boldsymbol{\eta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} S_n(\boldsymbol{\eta}),$$

with $S_n(\boldsymbol{\eta}) = \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\eta})$ and $\boldsymbol{\Theta}$ is the stataionary region of the ARMA(p, q) model.

- The limiting distribution of $\hat{\boldsymbol{\eta}}_n$ is

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}^{-1}),$$

where $\mathbf{C} = \mathrm{E}[\nabla \varepsilon_t(\boldsymbol{\eta}_0) \nabla \varepsilon_t^\top(\boldsymbol{\eta}_0)]$.

Prediction

- To predict Z_{n+1}, we use

$$\hat{Z}_{n+1}(\hat{\boldsymbol{\eta}}_n) = Z_{n+1} - \varepsilon_{n+1}(\hat{\boldsymbol{\eta}}_n).$$

- To predict Z_{n+2}, we use

$$\hat{Z}_{n+2}(\hat{\boldsymbol{\eta}}_n) = Z_{n+1} - \varepsilon^*_{n+2}(\hat{\boldsymbol{\eta}}_n),$$

  where $\varepsilon^*_{n+2}(\hat{\boldsymbol{\eta}}_n)$ is $\varepsilon_{n+2}(\hat{\boldsymbol{\eta}}_n)$ with Z_{n+1} therein replaced by $\hat{Z}_{n+1}(\hat{\boldsymbol{\eta}}_n)$, called plug-in method.

- For the mean squared error, we focus on one-step prediction. Note that

$$\begin{aligned} \mathrm{E}[Z_{n+1} - \hat{Z}_{n+1}(\hat{\boldsymbol{\eta}}_n)]^2 &= \mathrm{E}[\epsilon_{n+1}(\boldsymbol{\eta}_0) + (\varepsilon_{n+1}(\hat{\boldsymbol{\eta}}_n) - \varepsilon_{n+1}(\boldsymbol{\eta}_0))]^2 \\ &= \sigma^2 + \mathrm{E}[\varepsilon_{n+1}(\hat{\boldsymbol{\eta}}_n) - \varepsilon_{n+1}(\boldsymbol{\eta}_0)]^2, \end{aligned}$$

and

$$\begin{aligned} &n\mathrm{E}\big(\varepsilon_{n+1}(\hat{\boldsymbol{\eta}}_n) - \varepsilon_{n+1}(\boldsymbol{\eta}_0)\big)^2 \dot{\sim} \mathrm{E}\big(\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\big)^2 \\ &= \mathrm{E}\big(\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)^\top \nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\big). \end{aligned} \tag{2}$$

- Discussions:
  (1) $\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)$ and $\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)$ are asymptotically independent. This is similar to AR(p) models which we've discussed before.
  (2) Therefore, the (2) is asymptotically equivalent to

$$\mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)^\top \mathrm{E}\big(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\big)\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\right).$$

- Discussions (continue):

  (3) $\mathrm{E}\left(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\right) = \mathbf{C}$ because by our hypothesis ensures

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum\nabla\varepsilon_t(\boldsymbol{\eta}_0)\nabla\varepsilon_t^\top(\boldsymbol{\eta}_0) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum\nabla g_t(\boldsymbol{\eta}_0)\nabla g_t^\top(\boldsymbol{\eta}_0) = \mathbf{C},$$

  and $\nabla\varepsilon_t(\boldsymbol{\eta}_0)$ is stationary.

  (4) By (3) and CLT for $\hat{\boldsymbol{\eta}}_n$, we have

$$\frac{\sqrt{n}(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)^\top\mathrm{E}\left(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\right)\sqrt{n}(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)}{\sigma^2}\xrightarrow{d}\chi^2(p+q),$$

  and hence

$$\mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)^\top\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\nabla\varepsilon_{n+1}^\top(\boldsymbol{\eta}_0)\sqrt{n}(\hat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0)\right)\dot\sim(p+q)\sigma^2.$$

  Consequently,

$$\mathrm{E}[Z_{n+1}-\hat{Z}_{n+1}(\hat{\boldsymbol{\eta}}_n)]^2\dot\sim\sigma^2+\frac{(p+q)\sigma^2}{n}.$$

Model Selection

- Consider the following two information criteria,

$$\mathrm{AIC} = \log \hat{\sigma}_n^2(p,q) + \frac{2(p+q)}{n},$$
$$\mathrm{BIC} = \log \hat{\sigma}_n^2(p,q) + \frac{C_n(p+q)}{n},$$

  where $C_n \to \infty$, $C_n/n \to 0$, and $\hat{\sigma}_n^2(p,q) = n^{-1}\sum_{t=1}^{n} \epsilon_t^2(\hat{\boldsymbol{\eta}}_n)$.
- It can be shown that BIC is model selection consistent in ARMA(p,q) model; see E. J. Hannan (1980, Annals of Statistics).

ARMA Models with Non-Zero Mean

- **Application to AR(1) model.** Consider a AR(1) model,

$$Z_t - \mu_0 = \phi_0(Z_{t-1} - \mu_0) + \epsilon_t.$$

- Define

$$\varepsilon_t(\mu, \phi) = (1 - \phi B)(Z_t - \mu) = (1 - \phi B)Z_t - (1 - \phi)\mu.$$

- We have two estimation methods:
  (1) Minimize $S_n(\mu, \phi) = \sum \varepsilon_t^2(\mu, \phi)$.
  (2) Estimate μ by $\bar{Z}$ and then treat $Z_t - \bar{Z}$ as a zero-mean process whose estimation problems have been discussed previously.

- For method (1), we have

  (1) $\frac{\partial}{\partial \mu}\varepsilon_t(\mu_0, \phi_0) = -(1-\phi_0)$.

  (2) $\frac{\partial}{\partial \phi}\varepsilon_t(\mu_0, \phi_0) = -(Z_{t-1}-\mu_0)$.

  (3) Notice that

$$\mathbf{C} = \mathrm{E}\left[\begin{pmatrix} -(1-\phi_0) \\ -(Z_{t-1}-\mu_0) \end{pmatrix}\begin{pmatrix} -(1-\phi_0) \\ -(Z_{t-1}-\mu_0) \end{pmatrix}^\top\right] = \begin{pmatrix} (1-\phi_0)^2 & 0 \\ 0 & \sigma^2/(1-\phi_0^2) \end{pmatrix}.$$

  Hence, the limiting distribution of $(\hat{\mu}_n, \hat{\phi}_n)$ is

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu}_n \\ \hat{\phi}_n \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \phi_0 \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2/(1-\phi_0)^2 & 0 \\ 0 & 1-\phi_0^2 \end{pmatrix}\right).$$

- For method (2), we have

$$\sqrt{n}(\bar{Z} - \mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \mu_0),$$

and

$$\begin{aligned} E\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \mu_0)\Big)^2 \to \sum_{j=-\infty}^{\infty} \gamma_j &= \gamma_0(1 + 2\rho_1 + 2\rho_2 + \cdots) \\ &= \gamma_0\Big(\frac{2}{1-\phi_0} - 1\Big) = \frac{\sigma^2}{(1-\phi_0)^2}. \end{aligned}$$

as $n \to \infty$.

- This means $\bar{Z}$ is as good as (asymptotically equivalent to) $\hat{\mu}_n$ obtained in method (1). Moreover, it can be shown that the LSE of ϕ_0 based on $Z_t - \bar{Z}$ is as good as $\hat{\phi}_n$.

- **Application to MA(1) model.** Consider a MA(1) model,

$$Z_t = \mu_0 + \epsilon_t - \theta_0 \epsilon_{t-1}.$$

- Define

$$\varepsilon_t(\mu, \theta) = \frac{Z_t - \mu}{1 - \theta B}.$$

- Again, we have two estimation methods:
  (1) Minimize $S_n(\mu, \theta) = \sum \varepsilon_t^2(\mu, \theta)$.
  (2) Estimate μ by $\bar{Z}$ and then treat $Z_t - \bar{Z}$ as a zero-mean process.

- For method (1), we have
  (1) $\frac{\partial}{\partial\mu}\varepsilon_t(\mu_0, \theta_0) = -1/(1-\theta_0)$.
  (2) $\frac{\partial}{\partial\theta}\varepsilon_t(\mu_0, \theta_0) = \varepsilon_{t-1}/(1-\theta_0 B)$.
  (3) Since

$$\mathbf{C} = \begin{pmatrix} 1/(1-\theta_0)^2 & 0 \\ 0 & \sigma^2/(1-\theta_0^2) \end{pmatrix},$$

  the limiting distribution of $(\hat{\mu}_n, \hat{\theta}_n)$ is

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu}_n \\ \hat{\theta}_n \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \theta_0 \end{pmatrix} \right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2(1-\theta_0)^2 & 0 \\ 0 & 1-\theta_0^2 \end{pmatrix} \right).$$

- For method (2), we have

$$\mathrm{E}\left(\sqrt{n}(\bar{Z} - \mu_0)\right)^2 \to \sigma^2(1 - \theta_0)^2,$$

  as $n \to \infty$.
- This means that method (1) and method (2) are asymptotically equivalent.

- **Exercise.** The application to ARMA(1,1) model is left as an exercise.

Prediction

- In the following, we focus on the AR(1) model with non-zero mean, *i.e.*,

$$Z_t - \mu_0 = \phi_0(Z_{t-1} - \mu_0) + \epsilon_t.$$

- To predict Z_{n+1}, we use

$$\begin{aligned}
\hat{Z}_{n+1} &= Z_{n+1} - \epsilon_{n+1}(\hat{\mu}_n, \hat{\phi}_n) \\
&= Z_{n+1} - (1 - \hat{\phi}_n B)Z_{n+1} + (1 - \hat{\phi}_n)\hat{\mu}_n \\
&= \hat{\phi}_n(Z_n - \hat{\mu}_n) + \hat{\mu}_n.
\end{aligned}$$

- The one-step mean squared error is

$$\begin{aligned} \mathrm{E}(Z_{n+1} - \hat{Z}_{n+1})^2 &= \mathrm{E}(\epsilon_{n+1}(\hat{\mu}_n, \hat{\phi}_n))^2 \\ &= \mathrm{E}[\epsilon_{n+1} + (\epsilon_{n+1}(\hat{\mu}_n, \hat{\phi}_n) - \epsilon_{n+1}(\mu_0, \phi_0))]^2 \\ &\dot{\sim} \sigma^2 + \frac{1}{n}\mathrm{E}[\nabla\epsilon_{n+1}^\top(\eta_0)\sqrt{n}(\hat{\eta}_n - \eta_0)]^2 \\ &\dot{\sim} \sigma^2\left(1 + \frac{2}{n}\right). \end{aligned}$$

where $\eta_0 = (\mu_0, \phi_0)$ and $\hat{\eta}_n = (\hat{\mu}_n, \hat{\phi}_n)$.

- What if we use $\bar{Z}$ to predict Z_{n+1}? Let $S_t = Z_t - \mu_0$, then the one-step mean squared error becomes

$$\begin{aligned} \mathrm{E}(Z_{n+1} - \bar{Z})^2 &= \mathrm{E}(S_{n+1} - (\bar{Z} - \mu_0))^2 \\ &= \mathrm{E}\left(S_{n+1} - \frac{1}{n}\sum_{t=1}^{n} S_t\right)^2 \dot{\sim} \frac{\sigma^2}{1-\phi_0^2}\left(1 + \frac{1}{n}\right). \end{aligned}$$

Regression Models with Time Series Error

- Consider the following model

$$y_t = \beta_{0,0} + \beta_{1,0}x_{t1} + \cdots + \beta_{r,0}x_{tr} + \epsilon_t := f_t + \epsilon_t,$$

where $\epsilon_t \sim \text{ARMA}(p,q)$, *i.e.*,

$$(1 - \phi_{1,0}B - \cdots - \phi_{p,0}B^p)\varepsilon_t = (1 - \theta_{1,0}B - \cdots - \theta_{q,0}B^q)\delta_t,$$

with δ_t's are i.i.d. random variables with mean 0 and variance σ^2.

- The CLS of $\boldsymbol{\eta}_0 = (\beta_{0,0}, \ldots, \beta_{r,0}, \phi_{1,0}, \ldots, \phi_{p,0}, \theta_{1,0}, \ldots, \theta_{q,0})^\top$ is

$$\hat{\boldsymbol{\eta}}_n = \underset{\boldsymbol{\eta}\in\boldsymbol{\Theta}}{\text{argmin}} S_n(\boldsymbol{\eta}) = \underset{\boldsymbol{\eta}\in\boldsymbol{\Theta}}{\text{argmin}} \sum \delta_t^2(\boldsymbol{\eta}),$$

where

$$\delta_t(\boldsymbol{\eta}) = \frac{\Phi_p(B)}{\Theta_p(B)}\varepsilon_t(\boldsymbol{\beta}),$$

with $\varepsilon_t(\boldsymbol{\beta}) = (y_t - \beta_0 - \cdots - \beta_r x_{t,r})$ for $t \geq 1$ and $\varepsilon_t(\boldsymbol{\beta}) = 0$ for $t \leq 0$.

- In the following, we will focus on the following three scenarios:
  (a) $\varepsilon_t \sim$ AR(1),
  (b) $\varepsilon_t \sim$ MA(1),
  (c) $\varepsilon_t \sim$ ARMA(1,1).

- **Case 1**: $f_t = \beta_{0,0}$, *i.e.*,

$$y_t = \beta_{0,0} + \varepsilon_t.$$

  In this case, it reduces to the case that ARMA models with non-zero mean, which we have already shown their properties in the previous section.

- **Case 2**: $f_t = \beta_{0,0} + \beta_{1,0} t$, *i.e.*,

$$y_t = \beta_{0,0} + \beta_{1,0} t + \varepsilon_t.$$

We first consider the scenario (a), $\epsilon_t \sim \text{AR}(1)$. It is easy to see that

(1) $\delta_t(\boldsymbol{\eta}) = (1 - \phi_1 B)(y_t - \beta_0 - \beta_1 t)$.

(2) $\nabla \delta_t(\boldsymbol{\eta}_0) = \big(-(1 - \phi_{1,0}), -(1 - \phi_{1,0} B)t, -\varepsilon_{t-1} \big)^\top$.

(3) $\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0 \dot{\sim} -\big(\sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0) \nabla \delta_t^\top(\boldsymbol{\eta}_0) \big)^{-1} \sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0) \delta_t$.

- **Remark 1.** Since in this case $n^{-1} \sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0) \nabla \delta_t^\top(\boldsymbol{\eta}_0)$ does not converge to a constant matrix, the previous theory on the CLS estimate, which require $n^{-1} \sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0) \nabla \delta_t^\top(\boldsymbol{\eta}_0)$ have a constant limit, is no longer applicable here. So, the analysis has to "redo from scratch".

- **Remark 2.** Fact (3) follows from the fact that

$$0 = \nabla S_n(\hat{\boldsymbol{\eta}}_n) \dot{\sim} \nabla S_n(\boldsymbol{\eta}_0) + (\nabla^2 S_n(\boldsymbol{\eta}))^{-1}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0),$$

with $\nabla S_n(\boldsymbol{\eta}_0) = 2\sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0)\delta_t$ and

$$\nabla^2 S_n(\boldsymbol{\eta}_0) = 2\sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0)\nabla \delta_t^\top(\boldsymbol{\eta}_0) + 2\sum_{t=1}^n (\nabla^2 \delta(\boldsymbol{\eta}_0))\delta_t.$$

Notice that $2\sum_{t=1}^n (\nabla^2 \delta(\boldsymbol{\eta}_0))\delta_t$ is asymptotically negligible compared to $\sum_{t=1}^n \nabla \delta_t(\boldsymbol{\eta}_0)\nabla \delta_t^\top(\boldsymbol{\eta}_0)$.

- Let $\mathbf{D} = \text{diag}(1, 1/n, 1)$, it follows that

$$\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0 \dot{\sim} \left[\sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix}^{\top} \right]^{-1} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix} \delta_t$$

$$\dot{\sim} \mathbf{D} \left[\mathbf{D} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix}^{\top} \mathbf{D} \right]^{-1} \mathbf{D} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t \\ \epsilon_{t-1} \end{pmatrix} \delta_t$$

$$\dot{\sim} \mathbf{D} \left[\sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix}^{\top} \right]^{-1} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix} \delta_t$$

and hence

$$\sqrt{n}\,\mathbf{D}^{-1} (\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \dot{\sim} \left[\frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix}^{\top} \right]^{-1}$$

$$\cdot \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix} \delta_t \xrightarrow{d} N(0, \mathbf{C}^{-1}\sigma^2).$$

- To compute **C**, notice that

$$
\begin{aligned}
\mathbf{C} =& \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix} \begin{pmatrix} 1-\phi_{1,0} \\ (1-\phi_{1,0}B)t/n \\ \epsilon_{t-1} \end{pmatrix}^{\top} \\
=& \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} (1-\phi_{1,0})^2 & (1-\phi_{1,0})(1-\phi_{1,0}B)t/n & (1-\phi_{1,0})\epsilon_{t-1} \\ - & \left((1-\phi_{1,0}B)t/n\right)^2 & (1-\phi_{1,0}B)t\epsilon_{t-1}/n \\ - & - & \epsilon_{t-1}^2 \end{pmatrix} \\
\dot{\sim}& \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} (1-\phi_{1,0})^2 & (1-\phi_{1,0})^2 t/n & (1-\phi_{1,0})\epsilon_{t-1} \\ - & \left((1-\phi_{1,0})t/n\right)^2 & (1-\phi_{1,0})t\epsilon_{t-1}/n \\ - & - & \epsilon_{t-1}^2 \end{pmatrix}.
\end{aligned}
$$

The last step holds because $t \,\dot{\sim}\, t-1$.

- Now, we list some key facts to simply this matrix

  (1) $\sum_{t=1}^{n} t \sim \int_1^n t\, dt \sim n^2/2$.
  (2) $\sum_{t=1}^{n} t^2 \sim \int_1^n t^2\, dt \sim n^3/3$.
  (3) $\text{Var}(n^{-2} \sum_{t=1}^{n} t\epsilon_{t-1}) = O(1/n) \to 0$.

- To show fact (3), remember that $\lambda_{\max}(B^\top AB) \leq \lambda_{\max}(A)\lambda_{\max}(B^\top B)$, provided A is symmetric and positive definite. Then the required result follows from the assumption that $\lambda_{\max}(\mathbf{R})$ is bounded and

$$\text{Var}\left(n^{-2}\sum_{t=1}^{n} t\epsilon_{t-1}\right) = n^{-4}\begin{pmatrix}1\\ \vdots\\ n\end{pmatrix}^\top E\left[\begin{pmatrix}\epsilon_0\\ \vdots\\ \epsilon_{n-1}\end{pmatrix}\begin{pmatrix}\epsilon_0\\ \vdots\\ \epsilon_{n-1}\end{pmatrix}^\top\right]\begin{pmatrix}1\\ \vdots\\ n\end{pmatrix}$$

$$\leq n^{-4}\lambda_{\max}(\mathbf{R})\sum_{t=1}^{n} t^2.$$

- Therefore, we have

$$\mathbf{C} = \begin{pmatrix} (1-\phi_{1,0})^2 & (1-\phi_{1,0})^2/2 & 0 \\ (1-\phi_{1,0})^2/2 & (1-\phi_{1,0})^2/3 & 0 \\ 0 & 0 & \sigma^2/(1-\phi_{1,0}^2) \end{pmatrix},$$

and hence

$$\begin{pmatrix} n^{0.5} & 0 & 0 \\ 0 & n^{1.5} & 0 \\ 0 & 0 & n^{0.5} \end{pmatrix} (\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\lambda & -4\lambda & 0 \\ -6\lambda & 12\lambda & 0 \\ 0 & 0 & 1-\phi_{1,0}^2 \end{pmatrix} \right),$$

where $\lambda = \lim_{n\to\infty} \mathrm{Var}(n^{-0.5} \sum_{t=1}^{n} \epsilon_t) = \sigma^2/(1-\phi_{1,0})^2$.

- We next show that the LSE $(\tilde{\beta}_0, \tilde{\beta}_1)$ (method (2) for the regression part) is asymptotically equivalent to $(\hat{\beta}_0, \hat{\beta}_1)$. Note first that

$$
\begin{aligned}
&\mathrm{diag}(\sqrt{n}, n^{1.5}) \left[\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} - \begin{pmatrix} \beta_{0,0} \\ \beta_{1,0} \end{pmatrix} \right] \\
&= \mathrm{diag}(\sqrt{n}, n^{1.5}) \left(\sum_{t=1}^{n} \begin{pmatrix} 1 \\ t \end{pmatrix} \begin{pmatrix} 1 \\ t \end{pmatrix}^{\top} \right)^{-1} \sum_{t=1}^{n} \begin{pmatrix} 1 \\ t \end{pmatrix} \left[y_t - \begin{pmatrix} 1 \\ t \end{pmatrix}^{\top} \begin{pmatrix} \beta_{0,0} \\ \beta_{1,0} \end{pmatrix} \right] \\
&= \left(\frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} 1 \\ t/n \end{pmatrix} \begin{pmatrix} 1 \\ t/n \end{pmatrix}^{\top} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \begin{pmatrix} 1 \\ t/n \end{pmatrix} \epsilon_t \\
&\to \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}^{-1} \begin{pmatrix} n^{-0.5} \sum_{t=1}^{n} \epsilon_t \\ n^{-1.5} \sum_{t=1}^{n} t\epsilon_t \end{pmatrix}.
\end{aligned}
$$

- For the first element,

$$\mathrm{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\epsilon_t\right)^2 = n\mathrm{Var}(\bar{\epsilon}) \to \sum_{j=-\infty}^{\infty}\gamma_j = \frac{\sigma^2}{(1-\phi_{1,0})^2}.$$

- For the second element,

$$\mathrm{E}\left(\frac{1}{n^{1.5}}\sum_{t=1}^{n}t\epsilon_t\right)^2 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}ij\gamma_{i-j} = \frac{\gamma_0}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}ij\rho_{i-j}.$$

Using the fact that $\rho_j = \phi_{1,0}^{|j|}$, we have

$$\mathrm{E}\left(\frac{1}{n^{1.5}}\sum_{t=1}^{n}t\epsilon_t\right)^2 = \frac{\gamma_0}{n^3}\sum_{i=1}^{n}i\left(\underbrace{\sum_{j=1}^{i}\phi_{1,0}^{i-j}}_{(S_1)} + \underbrace{\sum_{j=i+1}^{n}j\phi_{1,0}^{j-i}}_{(S_2)}\right)$$

.

- We first simply (S_1),

$$(S_1) = \sum_{j=1}^{i} \phi_{1,0}^{i-j} = \phi_{1,0}^{i-1} + 2\phi_{1,0}^{i-2} + \cdots + i\phi_{1,0}^{0}$$

$$= i\phi_{1,0}^{0} + (i-1)\phi_{1,0}^{1} + \cdots + (i-(i-1))\phi_{1,0}^{i-1}$$

$$= i\frac{1-\phi_{1,0}^{i}}{1-\phi_{1,0}} - \sum_{j=1}^{i-1} j\phi_{1,0}^{j} \dot{\sim} \frac{i}{1-\phi_{1,0}} - \sum_{j=1}^{i-1} j\phi_{1,0}^{j}.$$

- Similarly, (S_2) can be simplified as

$$(S_2) = \sum_{j=i+1}^{n} j\phi_{1,0}^{j-i} = (i+1)\phi_{1,0}^{1} + (i+2)\phi_{1,0}^{2} + \cdots + (i+(n-i))\phi_{1,0}^{n-i}$$

$$= \sum_{j=1}^{n-i} j\phi_{1,0}^{j} + i\phi_{1,0}\frac{1-\phi_{1,0}^{n-i}}{1-\phi_{1,0}} \dot{\sim} \frac{i\phi_{1,0}}{1-\phi_{1,0}}.$$

- It follows that

$$\mathrm{E}\left(\frac{1}{n^{1.5}}\sum_{t=1}^{n} t\epsilon_t\right)^2 \dot{\sim} \frac{\gamma_0}{n^3}\sum_{i=1}^{n} i^2 \frac{1+\phi_{1,0}}{1-\phi_{1,0}} \to \frac{1}{3}\frac{\sigma^2}{(1-\phi_{1,0})^2}.$$

- Similarly, we have

$$\mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} j\epsilon_i\epsilon_j\right) \to \frac{1}{2}\frac{\sigma^2}{(1-\phi_{1,0})^2}.$$

- As a result,

$$\begin{pmatrix} \sqrt{n} & 0 \\ 0 & n^{1.5} \end{pmatrix}\left[\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} - \begin{pmatrix} \beta_{0,0} \\ \beta_{1,0} \end{pmatrix}\right] \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\sigma^2}{(1-\phi_{1,0})^2}\begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix}\right),$$

  meaning that $(\tilde{\beta}_0, \tilde{\beta}_1)^\top$ is asymptotically equivalent to $(\hat{\beta}_0, \hat{\beta}_1)^\top$.
- In fact, we can further the LSE $\tilde{\phi}_1$ of $\phi_{1,0}$ based on the detrended series $y_t - \tilde{\beta}_0 - \tilde{\beta}_1 t$ is asymptotically equivalent to $\hat{\phi}_1$.

- We now turn to the scenario (b), $\epsilon_t \sim$ MA(1).
- Main result.

$$\begin{pmatrix} n^{0.5} & 0 & 0 \\ 0 & n^{1.5} & 0 \\ 0 & 0 & n^{0.5} \end{pmatrix} (\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\psi & -4\psi & 0 \\ -6\psi & 12\psi & 0 \\ 0 & 0 & 1-\theta_{1,0}^2 \end{pmatrix} \right),$$

where $\psi = \lim_{n\to\infty} \mathrm{Var}(n^{-0.5} \sum_{t=1}^{n} \epsilon_t) = (1-\theta_{1,0})^2 \sigma^2$.

- Some details:

(1) $\delta_t(\boldsymbol{\eta}) = \frac{y_t - \beta_0 - \beta_1 t}{1 - \theta_1 B}$.

(2) $\nabla \delta_t(\boldsymbol{\eta}_0) = (-1/(1-\theta_{1,0}), -t/(1-\theta_{1,0}B), \delta_t/(1-\theta_{1,0}B))^\top$.

(3) $-t/(1-\theta_{1,0}B) \dot{\sim} -t/(1-\theta_{1,0})$ as t is large.

- Some details (continue):

  (4) Let $\mathbf{D} = \mathrm{diag}(1, n^{-1}, 1)$, then

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{D}\nabla\delta_t(\boldsymbol{\eta}_0)\nabla\delta_t^{\top}(\boldsymbol{\eta}_0)\mathbf{D} \xrightarrow{p} \mathbf{C},$$

  where

$$\mathbf{C} = \begin{pmatrix} 1/(1-\theta_{1,0})^2 & 1/(2(1-\theta_{1,0})^2) & 0 \\ 1/(2(1-\theta_{1,0})^2) & 1/(3(1-\theta_{1,0})^2) & 0 \\ 0 & 0 & \sigma^2/(1-\theta_{1,0}^2) \end{pmatrix}.$$

  (5) It follows that

$$\mathbf{C}^{-1}\sigma^2 = \begin{pmatrix} 4\psi & -4\psi & 0 \\ -6\psi & 12\psi & 0 \\ 0 & 0 & 1-\theta_{1,0}^2 \end{pmatrix},$$

  where $\psi = (1-\theta_{1,0})^2\sigma^2$.

- It can be shown that the LSE $(\tilde{\beta}_0, \tilde{\beta}_1)^\top$ of $(\beta_0, \beta_1)^\top$ is asymptotically equivalent to $(\hat{\beta}_0, \hat{\beta}_1)^\top$, and the CLS of $\theta_{1,0}$ based on detrended time series $y_t - \tilde{\beta}_0 - \tilde{\beta}_1 t$, denoted by $\tilde{\theta}_1$, is asymptotically equivalent to $\hat{\theta}_1$.

- **Exercise.** The scenario (c), $\epsilon_t \sim \text{ARMA}(1,1)$, is left as an exercise.

Prediction

- To predict y_{n+1},we use
$$\hat{y}_{n+1}(\hat{\boldsymbol{\eta}}_n) = y_{n+1} - \delta_{n+1}(\hat{\boldsymbol{\eta}}_n).$$
- To predict y_{n+2}, we use
$$\hat{y}_{n+2}(\hat{\boldsymbol{\eta}}_n) = y_{n+2} - \delta^*_{n+2}(\hat{\boldsymbol{\eta}}_n),$$
where $\delta^*_{n+2}(\hat{\boldsymbol{\eta}}_{\mathbf{n}})$ is $\delta_{n+2}(\hat{\boldsymbol{\eta}}_n)$ with y_{n+1} therein replaced by $\hat{y}_{n+1}(\hat{\boldsymbol{\eta}}_n)$.
- The predictor of y_{n+h}, $h \geq 3$, can be obtained recursively using the same manner.

- For the mean squared error, we focus on one-step prediction, with $f_t = \beta_{0,0} + \beta_{1,0} t$ and ϵ_t is an AR(1) model. Note that

$$
\begin{aligned}
\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\hat{\boldsymbol{\eta}}_n))^2 &= \mathrm{E}(\delta_{n+1}(\hat{\boldsymbol{\eta}}_n) - \delta_{n+1}(\boldsymbol{\eta}_0) + \delta_{n+1}(\boldsymbol{\eta}_0))^2 \\
&= \sigma^2 + \frac{1}{n}\mathrm{E}\left[n\left(\delta_{n+1}(\hat{\boldsymbol{\eta}}_n) - \delta_{n+1}(\boldsymbol{\eta}_0)\right)^2\right],
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{E}\left[n\left(\delta_{n+1}(\hat{\boldsymbol{\eta}}_n) - \delta_{n+1}(\boldsymbol{\eta}_0)\right)^2\right] \dot{\sim} \mathrm{E}\left(\nabla\delta_{n+1}^\top(\eta_0)\sqrt{n}(\hat{\eta}_n - \eta_0)\right)^2 \\
&= \mathrm{E}\left[\begin{pmatrix} -(1-\phi_{1,0}) \\ -(1-\phi_{1,0}B)(n+1)/n \\ -\epsilon_n \end{pmatrix}^\top \mathbf{D}_1\left(\hat{\eta}_n - \eta_0\right)\right]^2 \\
&\dot{\sim} \mathrm{E}\left[\begin{pmatrix} -(1-\phi_{1,0}) \\ -(1-\phi_{1,0}) \\ -\epsilon_n \end{pmatrix}^\top \mathbf{D}_1\left(\hat{\eta}_n - \eta_0\right)\right]^2
\end{aligned}
$$

where $\mathbf{D}_1 = \mathrm{diag}(n^{0.5}, n^{1.5}, n^{0.5})$.

- Notice that
  (1) $\mathbf{D}_1(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)$ and $(-(1-\phi_{1,0}), -(1-\phi_{1,0}), \epsilon_n)$ are asymptotically independent.
  (2) $\mathbf{D}_1(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4\lambda & -6\lambda & 0 \\ -6\lambda & 12\lambda & 0 \\ 0 & 0 & 1-\phi_{1,0}^2 \end{pmatrix}$$

  with $\lambda = \sigma^2/(1-\phi_{1,0})^2$.
  (3) If $\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$, then $\mathrm{E}(\mathbf{X}^\top \mathbf{A} \mathbf{X}) = \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\Sigma)$.

- Therefore, we obtain

$$
\begin{aligned}
& n\mathrm{E}(\delta_{n+1}(\hat{\boldsymbol{\eta}}_n) - \delta_{n+1}(\boldsymbol{\eta}_0))^2 \\
& \dot{\sim} \operatorname{tr}\left[\begin{pmatrix} (1-\phi_{1,0})^2 & (1-\phi_{1,0})^2 & 0 \\ (1-\phi_{1,0})^2 & (1-\phi_{1,0})^2 & 0 \\ 0 & 0 & \sigma^2/(1-\phi_{1,0}^2) \end{pmatrix} \boldsymbol{\Sigma}\right] \\
& \dot{\sim} \operatorname{tr}\left[\sigma^2 \begin{pmatrix} \lambda^{-1} & \lambda^{-1} & 0 \\ \lambda^{-1} & \lambda^{-1} & 0 \\ 0 & 0 & (1-\phi_{1,0}^2)^{-1} \end{pmatrix} \begin{pmatrix} 4\lambda & -6\lambda & 0 \\ -6\lambda & 12\lambda & 0 \\ 0 & 0 & 1-\phi_{1,0}^2 \end{pmatrix}\right] \\
& = 5\sigma^2.
\end{aligned}
$$

- As a result,

$$
\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\hat{\boldsymbol{\eta}}_n))^2 = \sigma^2 + \frac{5\sigma^2}{n}.
$$

- Since $5 \neq 3$ (the number of estimated parameters), this mean squared prediction error in nonstationary time series is quite different from that in the stationary case. We will go back to this point when discussing unit root models.

Proof of (1). Define W_n as

$$W_n = \left\{ \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\xi} \left| \frac{2}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_0))\epsilon_i \right| \geq \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\xi} \frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_0))^2 \right\}.$$

Then,

$$\Pr(W_n) \leq \Pr\left(W_n \cap \left\{ \min_{\theta\in\boldsymbol{\Theta}_\xi} \frac{1}{n}\sum_{i=1}^{n}(g_i(\theta) - g_i(\theta_0))^2 \geq \eta \right\}\right) + \Pr\left(\left\{ \min_{\theta\in\boldsymbol{\Theta}_\xi} \frac{1}{n}\sum_{i=1}^{n}(g_i(\theta) - g_i(\theta_0))^2 < \eta \right\}\right).$$

By **(C1)**, we obtain

$$\Pr(W_n) \leq \Pr\left(W_n \cap \left\{ \min_{\theta\in\boldsymbol{\Theta}_\xi} \frac{1}{n}\sum_{i=1}^{n}(g_i(\theta) - g_i(\theta_0))^2 \geq \eta \right\}\right) + \delta.$$

Moreover, by definition of W_n, we have

$$\Pr(W_n) \leq \Pr\left(\max_{\theta \in \mathbf{\Theta}_\xi} \left| \frac{2}{n} \sum_{i=1}^{n} (g_i(\theta) - g_i(\theta_0))\epsilon_i \right| \geq \eta \right) + \delta.$$

Since the first term on the RHS goes to zero by **(C2)**, we have the required result. □