Consider the relationship between the likelihood function above and the innovation form.

Remember the innovation form:

$$\log L(\phi_1, \sigma^2; y_n, y_{n-1}, \cdots, y_1) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log(1-\phi_1^2) \\ -\frac{1}{2\sigma^2}\Big((\sqrt{1-\phi_1^2}y_1)^2 + \sum_{t=2}^{n}(y_t - \phi_1 y_{t-1})^2\Big)$$

Focus on the last term:

$$(\sqrt{1-\phi_1^2}y_1)^2 + \sum_{t=2}^{n}(y_t - \phi_1 y_{t-1})^2$$
$$= (\sqrt{1-\phi_1^2}y_1, \quad y_2 - \phi_1 y_1, \quad y_3 - \phi_1 y_2, \quad \cdots, \quad y_n - \phi_1 y_{n-1}) \begin{pmatrix} \sqrt{1-\phi_1^2}y_1 \\ y_2 - \phi_1 y_1 \\ y_3 - \phi_1 y_2 \\ \vdots \\ y_n - \phi_1 y_{n-1} \end{pmatrix}$$

$$\begin{pmatrix} \sqrt{1-\phi_1^2}y_1 \\ y_2 - \phi_1 y_1 \\ y_3 - \phi_1 y_2 \\ \vdots \\ y_n - \phi_1 y_{n-1} \end{pmatrix} = \begin{pmatrix} \sqrt{1-\phi_1^2} & 0 & \cdots & \cdots & 0 \\ -\phi_1 & 1 & \ddots & & \vdots \\ 0 & -\phi_1 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & -\phi_1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = Ay$$

Therefore,

$$(\sqrt{1-\phi_1^2}y_1)^2 + \sum_{t=2}^{n}(y_t - \phi_1 y_{t-1})^2 = (Ay)'Ay = y'A'Ay$$

Comaring the exponential parts, the following equality holds.

$$-\frac{1}{2}y'\Omega^{-1}y = -\frac{1}{2\sigma^2}y'A'Ay \qquad \text{i.e.,} \qquad \Omega^{-1} = \frac{1}{\sigma^2}A'A$$

Remember that there exists P such that $\Omega = PP'$ when Ω is a symmetric and positive definite matrix.

In the AR(1) case, $P = \sigma A^{-1}$.

9.5 MLE: Regression Model with AR(1) Error

When the error term is autocorrelated, the regression model is written as:

$$y_t = x_t\beta + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \text{ iid } N(0, \sigma_\epsilon^2).$$

The joint distribution of $u_n, u_{n-1}, \cdots, u_1$ is:

$$\begin{aligned} f_u(u_n, u_{n-1}, \cdots, u_1; \rho, \sigma_\epsilon^2) &= f_u(u_1; \rho, \sigma_\epsilon^2) \prod_{t=2}^{n} f_u(u_t | u_{t-1}, \cdots, u_1; \rho, \sigma_\epsilon^2) \\ &= (2\pi\sigma_\epsilon^2/(1-\rho^2))^{-1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2/(1-\rho^2)} u_1^2\right) \\ &\quad \times (2\pi\sigma_\epsilon^2)^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{n} (u_t - \rho u_{t-1})^2\right). \end{aligned}$$

By transformation of variables from $u_n, u_{n-1}, \cdots, u_1$ to $y_n, y_{n-1}, \cdots, y_1$, the joint distribution of $y_n, y_{n-1}, \cdots, y_1$ is:

$$
\begin{aligned}
& f_y(y_n, y_{n-1}, \cdots, y_1; \rho, \sigma_\epsilon^2, \beta) \\
&\quad = f_u(y_n - x_n\beta, y_{n-1} - x_{n-1}\beta, \cdots, y_1 - x_1\beta; \rho, \sigma_\epsilon^2) \left| \frac{\partial u}{\partial y'} \right| \\
&\quad = (2\pi\sigma_\epsilon^2/(1-\rho^2))^{-1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2/(1-\rho^2)} (y_1 - x_1\beta)^2 \right) \\
&\qquad\qquad \times (2\pi\sigma_\epsilon^2)^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{n} \Big((y_t - \rho y_{t-1}) - (x_t - \rho x_{t-1})\beta \Big)^2 \right) \\
&\quad = (2\pi\sigma_\epsilon^2)^{-1/2} (1-\rho^2)^{1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2} (\sqrt{1-\rho^2}\, y_1 - \sqrt{1-\rho^2}\, x_1\beta)^2 \right) \\
&\qquad\qquad \times (2\pi\sigma_\epsilon^2)^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{n} \Big((y_t - \rho y_{t-1}) - (x_t - \rho x_{t-1})\beta \Big)^2 \right) \\
&\quad = (2\pi\sigma_\epsilon^2)^{-n/2} (1-\rho^2)^{1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2} (y_1^* - x_1^*\beta)^2 \right) \times \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{n} (y_t^* - x_t^*\beta)^2 \right)
\end{aligned}
$$

$$= (2\pi)^{-n/2}(\sigma_\epsilon^2)^{-n/2}(1-\rho^2)^{1/2}\exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^* - x_t^*\beta)^2\right)$$

$$= L(\rho, \sigma_\epsilon^2, \beta; y_n, y_{n-1}, \cdots, y_1),$$

where y_t^* and x_t^* are given by:

$$y_t^* = \begin{cases} \sqrt{1-\rho^2}\,y_t, & \text{for } t = 1, \\ y_t - \rho y_{t-1}, & \text{for } t = 2, 3, \cdots, n, \end{cases}$$

$$x_t^* = \begin{cases} \sqrt{1-\rho^2}\,x_t, & \text{for } t = 1, \\ x_t - \rho x_{t-1}, & \text{for } t = 2, 3, \cdots, n, \end{cases}$$

◎ For maximization, the first derivative of $L(\rho, \sigma_\epsilon^2, \beta; y_n, y_{n-1}, \cdots, y_1)$ with respect to β should be zero.

$$
\begin{aligned}
\tilde{\beta} &= (\sum_{t=1}^{T} x_t^{*\prime} x_t^*)^{-1} (\sum_{t=1}^{T} x_t^{*\prime} y_t^*) \\
&= (X^{*\prime} X^*)^{-1} X^{*\prime} y^*
\end{aligned}
$$

$\Longrightarrow$ This is equivalent to OLS from the regression model: $y^* = X^*\beta + \epsilon$ and $\epsilon \sim N(0, \sigma^2 I_n)$, where $\sigma^2 = \sigma_\epsilon^2/(1-\rho^2)$.

◎ For maximization, the first derivative of $L(\rho, \sigma_\epsilon^2, \beta; y_n, y_{n-1}, \cdots, y_1)$ with respect to σ_ϵ^2 should be zero.

$$\tilde{\sigma}_\epsilon^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t^* - x_t^*\beta)^2 = \frac{1}{n}(y^* - X^*\beta)'(y^* - X^*\beta),$$

where

$$y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix}, \qquad X^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}x_1 \\ x_2 - \rho x_1 \\ \vdots \\ x_n - \rho x_{n-1} \end{pmatrix}.$$

◎ For maximization, the first derivative of $L(\rho, \sigma_\epsilon^2, \beta; y_n, y_{n-1}, \cdots, y_1)$ with respect to ρ should be zero.

$$\max_{\beta,\sigma_\epsilon^2,\rho} L(\rho, \sigma_\epsilon^2, \beta; y) \quad \text{is equivalent to} \quad \max_{\rho} L(\rho, \tilde{\sigma}_\epsilon^2, \tilde{\beta}; y).$$

Note that both $\tilde{\sigma}_\epsilon^2$ and $\tilde{\beta}$ depend only on ρ.

$L(\rho, \tilde{\sigma}_\epsilon^2, \tilde{\beta}; y)$ is called the **concentrated log-likelihood function (集約対数尤度関数)**, which is a function of ρ.

The log-likelihood function is written as:

$$
\begin{aligned}
\log L(\rho, \tilde{\sigma}_\epsilon^2, \tilde{\beta}; y) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}_\epsilon^2) + \frac{1}{2}\log(1-\rho^2) - \frac{n}{2} \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\Big(\tilde{\sigma}_\epsilon^2(\rho)\Big) + \frac{1}{2}\log(1-\rho^2)
\end{aligned}
$$

For maximization of $\log L$, use Newton-Raphson method, method of scoring or simple grid search

Note that $\tilde{\sigma}_\epsilon^2 = \tilde{\sigma}_\epsilon^2(\rho) = \dfrac{1}{n}(y^* - X^*\tilde{\beta})'(y^* - X^*\tilde{\beta})$ for $\tilde{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$.

Remark: The regression model with AR(1) error is:

$$y_t = x_t\beta + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \text{ iid } N(0, \sigma_\epsilon^2).$$

$$\mathrm{V}(u) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{n-3} \\ \rho^3 & \rho^2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho^2 & \rho & 1 \end{pmatrix} = \sigma^2\Omega, \qquad \text{where } \sigma^2 = \frac{\sigma_\epsilon^2}{1-\rho^2}.$$

where $\mathrm{Cov}(u_i, u_j) = \mathrm{E}(u_i u_j) = \sigma^2\rho^{|i-j|}$, i.e., the ith row and jth column of Ω is $\rho^{|i-j|}$.

The regression model with AR(1) error is: $\qquad y = X\beta + u, \qquad u \sim N(0, \sigma^2\Omega).$

There exists P which satisfies that $\Omega = PP'$, because Ω is a positive definite matrix.

Multiply P^{-1} on both sides from the left.

$$
\begin{aligned}
P^{-1}y = P^{-1}X\beta + P^{-1}u \qquad &\Longrightarrow \qquad y^* = X^*\beta + u^* \text{ and } u^* \sim N(0, \sigma^2 I_n) \\
&\Longrightarrow \qquad \text{Apply OLS.}
\end{aligned}
$$

$$y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\rho & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = P^{-1}y$$

$$X^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}x_1 \\ x_2 - \rho x_1 \\ \vdots \\ x_n - \rho x_{n-1} \end{pmatrix} = P^{-1}X \qquad \Longrightarrow$$

Check $P^{-1}\Omega P^{-1\prime} = aI_n$, where a is constant.

9.6 MLE: Regression Model with Heteroscedastic Errors

In the case where the error term depends on the other exogenous variables, the regression model is written as follows:

$$y_i = x_i\beta + u_i, \qquad u_i \sim \text{ id } N(0, \sigma_i^2), \qquad \sigma_i^2 = (z_i\alpha)^2.$$

The joint distribution of $u_n, u_{n-1}, \cdots, u_1$, denoted by $f_u(\cdot;\cdot)$, is given by:

$$\begin{aligned}
\log f_u(u_n, u_{n-1}, \cdots, u_1; \sigma_1^2, \cdots, \sigma_n^2) &= \sum_{i=1}^{n} \log f_u(u_t; \sigma_i^2) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{u_i}{\sigma_i}\right)^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(z_i\alpha)^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{u_i}{z_i\alpha}\right)^2
\end{aligned}$$

By the transformation of variables from $u_n, u_{n-1}, \cdots, u_1$ to $y_n, y_{n-1}, \cdots, y_1$, the log-likelihood function is:

$$\begin{aligned} L(\alpha,\beta;y_n,y_{n-1},\cdots,y_1) &= \log f_y(y_n,y_{n-1},\cdots,y_1;\alpha,\beta) \\ &= \log f_u(y_n - x_n\beta, y_{n-1} - x_{n-1}\beta,\cdots,y_1 - x_1\beta;\sigma_i^2)\left|\frac{\partial u}{\partial y'}\right| \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(z_i\alpha)^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i\beta}{z_i\alpha}\right)^2 \end{aligned}$$

$\Longrightarrow$ Maximize the above log-likelihood function with respect to β and α.

10 Asymptotic Theory

1. **Definition: Convergence in Distribution (分布収束)**

   A series of random variables X_1, X_2, $\cdots$, X_n, $\cdots$ have distribution functions F_1, F_2, $\cdots$, respectively.

   If

   $$\lim_{n\to\infty} F_n = F,$$

   then we say that a series of random variables X_1, X_2, $\cdots$ converges to F in distribution.

2. **Consistency (一致性):**

   (a) **Definition: Convergence in Probability (確率収束)**

   Let $\{Z_n :\ n = 1, 2, \cdots\}$ be a series of random variables.

If the following holds,

$$\lim_{n\to\infty} P(|Z_n - \theta| < \epsilon) = 1,$$

for any positive ϵ, then we say that Z_n converges to θ in probability.

θ is called a **probability limit (確率極限)** of Z_n.

$$\text{plim}\, Z_n = \theta.$$

(b) Let $\hat{\theta}_n$ be an estimator of parameter θ.

If $\hat{\theta}_n$ converges to θ in probability, we say that $\hat{\theta}_n$ is a consistent estimator of θ.

3. **Markov's Inequality:** A General Case of **Chebyshev's Inequality:**

For $g(X) \geq 0$,

$$P(g(X) \geq k) \leq \frac{\mathrm{E}(g(X))}{k},$$

where k is a positive constant. See *Introduction Mathematical Statistics* (8th ed.), p.79 for the proof.

4. **Example:** For a random variable X, set $g(X) = (X-\mu)'(X-\mu)$, $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \Sigma$.

   Then, we have the following inequality:

$$P((X-\mu)'(X-\mu) \geq k) \leq \frac{\mathrm{tr}(\Sigma)}{k}.$$

   Note as follows:

$$\begin{aligned}\mathrm{E}((X-\mu)'(X-\mu)) &= \mathrm{E}\Big(\mathrm{tr}((X-\mu)'(X-\mu))\Big) = \mathrm{E}\Big(\mathrm{tr}((X-\mu)(X-\mu)')\Big) \\ &= \mathrm{tr}\Big(\mathrm{E}((X-\mu)(X-\mu)')\Big) = \mathrm{tr}(\Sigma).\end{aligned}$$

5. **Example 1 (Univariate Case):**

Suppose that $X_i \sim (\mu, \sigma^2)$, $i = 1, 2, \cdots, n$.

Then, the sample average $\overline{X}$ is a consistent estimator of μ.

Proof:

Note that $g(\overline{X}) = (\overline{X} - \mu)^2$, $\epsilon^2 = k$, $\mathrm{E}(g(\overline{X})) = \mathrm{V}(\overline{X}) = \dfrac{\sigma^2}{n}$.

Use Chebyshev's inequality.

If $n \longrightarrow \infty$,

$$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0, \qquad \text{for any } \epsilon.$$

That is. for any ϵ,

$$\lim_{n\to\infty} P(|\overline{X} - \mu| < \epsilon) = 1.$$

$\Longrightarrow$ **Chebyshev's inequality**

6. **Example 2 (Multivariate Case):**

Suppose that $X_i \sim (\mu, \Sigma)$, $i = 1, 2, \cdots, n$.

Then, the sample average $\overline{X}$ is a consistent estimator of μ.

Proof:

Note that $g(\overline{X}) = (\overline{X} - \mu)'(\overline{X} - \mu)$, $\epsilon^2 = k$, $\mathrm{E}(g(\overline{X})) = \mathrm{tr}\Big(\mathrm{V}(\overline{X})\Big) = \mathrm{tr}\Big(\frac{1}{n}\Sigma\Big)$.

Use Chebyshev's inequality.

If $n \longrightarrow \infty$,

$$P((\overline{X} - \mu)'(\overline{X} - \mu) \geq k) = P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\mathrm{tr}(\Sigma)}{n\epsilon^2} \longrightarrow 0, \quad \text{for any positive } \epsilon.$$

That is. for any positive ϵ, $\lim_{n\to\infty} P((\overline{X} - \mu)'(\overline{X} - \mu) < k) = 1$.

Note that $|\overline{X} - \mu| = \sqrt{(\overline{X} - \mu)'(\overline{X} - \mu)}$, which is the distance between X and μ.

$\Longrightarrow$ **Chebyshev's inequality**

7. **Some Formulas:**

Let X_n and Y_n be the random variables which satisfy plim $X_n = c$ and plim $Y_n = d$. Then,

(a) plim $(X_n + Y_n) = c + d$

(b) plim $X_n Y_n = cd$

(c) plim $X_n / Y_n = c/d$ for $d \neq 0$

(d) plim $g(X_n) = g(c)$ for a function $g(\cdot)$

8. **Some Formulas II:**

Let X_n and Y_n be the random variables which satisfy $X_n \longrightarrow c$ (converfgence in probability) and $Y_n \longrightarrow Y$ (convergence in distribution). Then,

$$X_n Y_n \longrightarrow cY$$

(a) cY is distributed with mean $c\mathrm{E}(Y)$ and variance $c^2\mathrm{V}(Y)$.

(b) In the multivariate case, cY is distributed with mean $c\mathrm{E}(Y)$ and variance $c\mathrm{V}(Y)c'$, where c, Y, $\mathrm{E}(Y)$ and $\mathrm{V}(\mathrm{Y})$ are $m\times k$, $k\times 1$, $k\times 1$ and $k\times k$ vectors or matrices.

9. **Central Limit Theorem (中心極限定理)**

Univariate Case: $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed as mean μ and variance σ^2.

Define $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Then,

$$\frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{V}(\overline{X})}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0, 1),$$

which implies

$$\sqrt{n}(\overline{X} - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \longrightarrow N(0, \sigma^2).$$

Multivariate Case: $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed as mean μ and variance Σ.

Define $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Then,

$$\sqrt{n}(\overline{X} - \mu) \longrightarrow N(0, \Sigma),$$

i.e.,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \longrightarrow N(0, \Sigma)$$

10. **Central Limit Theorem (Generalization)**

$X_1, X_2, \cdots, X_n$ are mutually independently distributed as mean μ and variance Σ_i.

Define $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Then,

$$\sqrt{n}(\overline{X} - \mu) \longrightarrow N(0, \Sigma),$$

i.e.,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \longrightarrow N(0, \Sigma),$$

where

$$\Sigma = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\Sigma_i\right).$$

11. **Definition:** Let $\hat{\theta}_n$ be a consistent estimator of θ.

    Suppose that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges to $N(0, \Sigma)$ in distribution.

    Then, we say that $\hat{\theta}_n$ has an **asymptotic distribution (漸近分布)**: $N(\theta, \Sigma/n)$.

10.1 MLE: Asymptotic Properties

1. $X_1, X_2, \cdots, X_n$ are random variables with density function $f(x; \theta)$.

   Let $\hat{\theta}_n$ be a maximum likelihood estimator of θ.

   Then, under some **regularity conditions**. $\hat{\theta}_n$ is a consistent estimator of θ and the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is given by:

$$\sqrt{n}(\hat{\theta} - \theta) \longrightarrow N\left(0, \lim_{n\to\infty}\left(\frac{I(\theta)}{n}\right)^{-1}\right)$$

2. **Regularity Conditions:**

   (a) The domain of X_i does not depend on θ.

   (b) There exists at least third-order derivative of $f(x;\theta)$ with respect to θ, and their derivatives are finite.

3. Thus, MLE is

   (i) consistent,

   (ii) asymptotically normal, and

   (iii) asymptotically efficient.

Proof: The log-likelihood function is given by:

$$\log L(\theta) = \log \prod_{i=1}^{n} f(X_i;\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$$