5.4 AR(p) model: Augmented Dickey-Fuller (ADF) Test

Consider the case where the error term is serially correlated.

Consider the following AR(p) model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t, \qquad \epsilon_t \sim \text{iid}(0, \sigma_\epsilon^2),$$

which is rewritten as:

$$\phi(L) y_t = \epsilon_t.$$

When the above model has a unit root, we have $\phi(1) = 0$, i.e., $\phi_1 + \phi_2 + \cdots + \phi_p = 1$.

The above AR(p) model is written as:

$$y_t = \rho y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + +\delta_{p-1} \Delta y_{t-p+1} + \epsilon_t,$$

where $\rho = \phi_1 + \phi_2 + \cdots + \phi_p$ and $\delta_j = -(\phi_{j+1} + \phi_{j+2} + \cdots + \phi_p)$.

The null and alternative hypotheses are:

$$H_0: \ \rho = 1 \text{ (Unit root)},$$

$$H_1: \ \rho < 1 \text{ (Stationary)}.$$

Use the t test, where we have the same asymptotic distributions.

We can utilize the same tables as before.

Choose p by AIC or SBIC.

Use $N(0,1)$ to test $H_0: \ \delta_j = 0$ against $H_1: \ \delta_j \neq 0$ for $j = 1, 2, \cdots, p-1$.

Reference

Kurozumi (2008) “Economic Time Series Analysis and Unit Root Tests: Development and Perspective,” *Japan Statistical Society*, Vol.38, Series J, No.1, pp.39 – 57.

Download the above paper from:

6 Bayesian Estimation

Bayes' procedure is briefly discussed (see Zellner (1971), Bernardo and Smith (1994), O'Hagan (1994), Hogg and Craig (1995) and so on for further discussion).

6.1 Elements of Bayesian Inference

When we have the random sample $(X_1, X_2, \cdots, X_n)$, consider estimating the unknown parameter θ. The maximum likelihood estimator is introduced for estimation of the parameter. Suppose that X_1, X_2, $\cdots$, X_n are mutually independently distributed and X_i has a probability density function $f(x;\theta)$, where θ is the unknown parameter to be estimated. The joint density of X_1, X_2, $\cdots$, X_n is given by:

$$f(x_1, x_2, \cdots, x_3; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

which is called the likelihood function, denoted by $l(\theta) = f(x_1, x_2, \cdots, x_n; \theta)$.

In Bayes' estimation, the parameter is taken as a random variable, say Θ, where a prior information on Θ is taken into account for estimation. The joint density function (or the likelihood function) is regarded as the conditional density function of X_1, X_2, $\cdots$, X_n given $\Theta = \theta$. Therefore, we write the likelihood function as the conditional density $f(x_1, x_2, \cdots, x_n|\theta)$. The probability density function of Θ is called the **prior probability density function** and given by $f_\theta(\theta)$. The conditional probability density function, $f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)$, have to be obtained, which is represented as:

$$\begin{aligned} f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n) &= \frac{f(x_1, x_2, \cdots, x_n|\theta) f_\theta(\theta)}{\int f(x_1, x_2, \cdots, x_n|\theta) f_\theta(\theta)\, \mathrm{d}\theta} \\ &\propto f(x_1, x_2, \cdots, x_n|\theta) f_\theta(\theta). \end{aligned}$$

The relationship in the first equality is known as Bayes' formula. The conditional probability density function of Θ given $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, i.e., $f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)$, is called the **posterior probability density function**, which is proportional to the product of the likelihood function and the prior density function.

6.1.1 Bayesian Point Estimate

Thus, the Bayesian approach yields the posterior probability density function for Θ. To obtain a point estimate of Θ, we introduce a loss function, denoted by $L(\Theta, \hat{\theta})$, where $\hat{\theta}$ indicates a point estimate depending on $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$. Since Θ is considered to be random, $L(\Theta, \hat{\theta})$ is also random. One solution which yields point estimates is to find the value of Θ that minimizes the mathematical expectation of the **loss function**, i.e.,

$$\min_{\hat{\theta}} \mathrm{E}\Big(L(\Theta, \hat{\theta})\Big) = \min_{\hat{\theta}} \int L(\theta, \hat{\theta}) f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)\, \mathrm{d}\theta,$$

where the absolute value of $\mathrm{E}\Big(L(\Theta,\hat{\theta})\Big)$ is assumed to be finite.

Now we specify the loss function as: $L(\Theta,\hat{\theta}) = (\Theta-\hat{\theta})'A(\Theta-\hat{\theta})$, which is called the **quadratic loss function**, where A is a known nonstochastic positive definite symmetric matrix. Then, the solution gives us the posterior mean of Θ, i.e.,

$$\hat{\theta} = \mathrm{E}(\Theta) = \int \theta f_{\theta|x}(\theta|x_1,x_2,\cdots,x_n)\,\mathrm{d}\theta.$$

An alternative loss function is given by: $L(\Theta,\hat{\theta}) = |\Theta-\hat{\theta}|$, which is called the **absolute error loss function**, where both Θ and $\hat{\theta}$ are assumed to be scalars. Then, the median of the posterior probability density function is an optimal point estimate of θ, i.e.,

$$\hat{\theta} = \text{median of the posterior probability density function.}$$

We have shown two Bayesian point estimates. Hereafter, the quadratic loss function is adopted for estimation. That is, the posterior mean of Θ is taken as the point estimate.

6.1.2 Bayesian Interval for Parameter

Given that the posterior probability density function $f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)$ has been obtained, it is possible to compute the probability that the parameter Θ lies in a particular subregion, R, of the parameter space. That is, we may compute the following probability:

$$P(\Theta \in R) = \int_R f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)\,\mathrm{d}\theta.$$

When the above probability is set to be $1 - \alpha$, it is possible to find the region that satisfies $P(\Theta \in R) = 1 - \alpha$, which region is not necessarily unique. In the case where Θ is a scalar, one possibility to determine the unique region $R = \{\Theta|a < \Theta < b\}$ is to obtain a and b by minimizing the distance $b - a$ subject to $\int_a^b f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n)\,\mathrm{d}\theta = 1 - \alpha$. By solving this minimization problem, determining a and b such that $\int_a^b f_{\theta|x}(\theta|x_1,\ x_2,\ \cdots,\ x_n)\,\mathrm{d}\theta = 1 - \alpha$ and $f_{\theta|x}(a|x_1,\ x_2,\ \cdots,\ x_n) = f_{\theta|x}(b|x_1,\ x_2,\ \cdots,\ x_n)$ leads to the shortest interval with probability $1 - \alpha$.

6.1.3 Prior Probability Density Function

We discuss a little bit about the prior probability density function. In the case where we know any information about the parameter θ beforehand, the plausible estimate of the parameter might be obtained if the parameter θ is estimated by including the prior information. For example, if we know that Θ is normally distributed with mean θ_0 and variance Σ_0, the prior density $f_\theta(\theta)$ is given by $N(\theta_0, \Sigma_0)$, where θ_0 and Σ_0 are known. On the contrary, we have the case where we do not know any prior information about the parameter θ. In this case, we may take the prior density as:

$$f_\theta(\theta) \propto \text{constant},$$

where the prior density of Θ is assumed to be uniform, which prior is called the **improper prior**, the **noninformative prior**, the **flat prior** or the **diffuse prior**. Then, the posterior density is given by:

$$f_{\theta|x}(\theta|x_1, x_2, \cdots, x_n) \propto f(x_1, x_2, \cdots, x_n|\theta).$$

That is, the posterior density function is proportional to the likelihood function.

Example: Suppose that X_1, X_2, $\cdots$, X_n are mutually independently, identically and normally distributed with mean μ and variance σ^2. Then, the likelihood function is given by:

$$\begin{aligned} f(x_1, x_2, \cdots, x_n|\theta) &= \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Bigl(-\frac{1}{2\sigma^2}(x_i-\mu)^2\Bigr) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Bigl(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Bigr), \end{aligned}$$

where θ indicates μ in this case. For simplicity of discussion, we assume that σ^2 is known. Therefore, we focus on μ.

Now, we consider two prior density functions for μ. One is noninformative and another is normal, i.e.,

(i) **Noninformative Prior:** $f_\theta(\mu) \propto$ constant, where μ is uniformly distributed.

(ii) **Normal Prior:** $f_\theta(\mu) = (2\pi\sigma_0^2)^{-1/2} \exp\Big(-\dfrac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big)$, where μ_0 and σ_0 are assumed to be known.

For each prior density, we obtain the posterior distributions as follows:

(i) When the prior density is noninformative, the posterior density function is:

$$
\begin{aligned}
& f_{\theta|x}(\mu|x_1, x_2, \cdots, x_n) \\
& \quad \propto f(x_1, x_2, \cdots, x_n|\mu) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big) \\
& \quad = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{1}{2\sigma^2}n(\overline{x} - \mu)^2\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= (2\pi\sigma^2)^{-n/2}\exp\Bigl(-\frac{(n-1)s^2}{2\sigma^2} - \frac{1}{2\sigma^2/n}(\mu - \overline{x})^2\Bigr) \\
&\propto \exp\Bigl(-\frac{1}{2\sigma^2/n}(\mu - \overline{x})^2\Bigr)
\end{aligned}
$$

where $\overline{x} = \sum_{i=1}^{n} x_i/n$ and $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$. Thus, the posterior density of μ represents the normal distribution with mean $\overline{x}$ and variance σ^2/n. Since under the quadratic loss function the point estimate of μ is given by the posterior mean, $\overline{x}$ gives us Bayes' point estimate. The Bayesian interval estimate of μ is: $(\overline{x} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{x} + z_{\alpha/2}\sigma/\sqrt{n})$, because from the posterior density function we have $P\Bigl(\Bigl|\dfrac{\mu - \overline{x}}{\sigma/\sqrt{n}}\Bigr| < z_{\alpha/2}\Bigr) = 1 - \alpha$.

(ii) When the prior density is normal, the posterior density function is given by:

$$
\begin{aligned}
&f_{\theta|x}(\mu|x_1, x_2, \cdots, x_n) \\
&\quad \propto f(x_1, x_2, \cdots, x_n|\mu) f_\theta(\mu)
\end{aligned}
$$

$$
\begin{aligned}
&= (2\pi\sigma^2)^{-n/2} \exp\Bigl(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Bigr) \\
&\quad \times(2\pi\sigma_0^2)^{-1/2} \exp\Bigl(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Bigr) \\
&\propto \exp\Bigl(-\frac{1}{2\sigma^2/n}(\mu-\overline{x})^2\Bigr)\times\exp\Bigl(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Bigr) \\
&\propto \exp\left(-\frac{1}{2}\Bigl(\frac{\sigma_0^2+\sigma^2/n}{\sigma_0^2\sigma^2/n}\Bigr)\Bigl(\mu-\frac{\overline{x}\sigma_0^2+\mu_0\sigma^2/n}{\sigma_0^2+\sigma^2/n}\Bigr)^2\right),
\end{aligned}
$$

which indicates that the posterior density of μ is a normal distribution with mean $\frac{\overline{x}\sigma_0^2+\mu_0\sigma^2/n}{\sigma_0^2+\sigma^2/n}$ and variance $\Bigl(\frac{\sigma_0^2+\sigma^2/n}{\sigma_0^2\sigma^2/n}\Bigr)^{-1}$. The posterior mean is rewritten as:

$$
\frac{\overline{x}\sigma_0^2+\mu_0\sigma^2/n}{\sigma_0^2+\sigma^2/n} = \overline{x}w+\mu_0(1-w),
$$

where $w=\sigma_0^2/(\sigma_0^2+\sigma^2/n)$. $\overline{x}$ is a maximum likelihood estimate of μ and μ_0 is a prior mean of μ. Thus, the posterior mean is the weighted average of $\overline{x}$ and

μ_0. As $w \longrightarrow 1$, i.e., as $\sigma_0^2 \longrightarrow \infty$, the posterior mean approaches $\overline{x}$, which is equivalent to the posterior mean with the noninformative prior.

6.2 Sampling Methods

6.2.1 Gibbs Sampling

The Gibbs sampler shows how to generate random draws from the unconditional densities under the situation that we can generate random draws from two conditional densities.

Geman and Geman (1984), Tanner and Wong (1987), Gelfand, Hills, Racine-Poon and Smith (1990), Gelfand and Smith (1990), Carlin and Polson (1991), Zeger and Karim (1991), Casella and George (1992), Gamerman (1997) and so on developed the Gibbs sampling theory. Carlin, Polson and Stoffer (1992), Carter and Kohn (1994, 1996) and Geweke and Tanizaki (1999, 2001) applied the Gibbs sampler to the nonlinear and/or non-Gaussian state-space models. There are numerous other applications of the Gibbs sampler. The Gibbs sampling theory is concisely described as follows.

We can deal with more than two random variables, but we consider two random variables X and Y in order to make things easier. Two conditional density functions, $f_{x|y}(x|y)$ and $f_{y|x}(y|x)$, are assumed to be known, which denote the conditional distribution function of X given Y and that of Y given X, respectively. Suppose that we can easily generate random draws of X from $f_{x|y}(x|y)$ and those of Y from $f_{y|x}(y|x)$. However, consider the case where it is not easy to generate random draws from the joint density of X and Y, denoted by $f_{xy}(x, y)$. In order to have the random draws of (X, Y) from the joint density $f_{xy}(x, y)$, we take the following procedure:

(i) Take the initial value of X as x_{-M}.

(ii) Given x_{i-1}, generate a random draw of Y, i.e., y_i, from $f(y|x_{i-1})$.

- (iii) Given y_i, generate a random draw of X, i.e., x_i, from $f(x|y_i)$.
- (iv) Repeat the procedure for $i = -M+1, -M+2, \cdots, 1$.

From the convergence theory of the Gibbs sampler, as M goes to infinity, we can regard x_1 and y_1 as random draws from $f_{xy}(x, y)$, which is a joint density function of X and Y. M denotes the **burn-in period**, and the first M random draws, (x_i, y_i) for $i = -M+1, -M+2, \cdots, 0$, are excluded from further consideration. When we want N random draws from $f_{xy}(x, y)$, Step (iv) should be replaced by Step (iv)', which is as follows.

- (iv)' Repeat the procedure for $i = -M+1, -M+2, \cdots, N$.

As in the Metropolis-Hastings algorithm, the algorithm shown in Steps (i) – (iii) and (iv)' is formulated as follows:

$$f_i(u) = \int f^*(u|v) f_{i-1}(v)\, \mathrm{d}v.$$

For convergence of the Gibbs sampler, we need to have the invariant distribution $f(u)$ which satisfies $f_i(u) = f_{i-1}(u) = f(u)$. If we have the reversibility condition shown in equation (3), i.e.,

$$f^*(v|u)f(u) = f^*(u|v)f(v),$$

the random draws based on the Gibbs sampler converge to those from the invariant distribution, which implies that there exists the invariant distribution $f(u)$. Therefore, in the Gibbs sampling algorithm, we have to find the transition distribution, i.e., $f^*(u|v)$. Here, we consider that both u and v are bivariate vectors. That is, $f^*(u|v)$ and $f_i(u)$ denote the bivariate distributions. x_i and y_i are generated from $f_i(u)$ through $f^*(u|v)$, given $f_{i-1}(v)$. Note that $u = (u_1, u_2) = (x_i, y_i)$ is taken while $v = (v_1, v_2) = (x_{i-1}, y_{i-1})$ is set. The transition distribution in the Gibbs sampler is taken as:

$$f^*(u|v) = f_{y|x}(u_2|u_1)f_{x|y}(u_1|v_2)$$

Thus, we can choose $f^*(u|v)$ as shown above. Then, as i goes to infinity, (x_i, y_i) tends in distribution to a random vector whose joint density is $f_{xy}(x, y)$. See, for example, Geman and Geman (1984) and Smith and Roberts (1993).

Furthermore, under the condition that there exists the invariant distribution, the basic result of the Gibbs sampler is as follows:

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i, y_i) \longrightarrow \mathrm{E}(g(x, y)) = \iint g(x, y) f_{xy}(x, y)\,\mathrm{d}x\,\mathrm{d}y, \quad \text{as } N \longrightarrow \infty,$$

where $g(\cdot, \cdot)$ is a function.

The Gibbs sampler is a powerful tool in a Bayesian framework. Based on the conditional densities, we can generate random draws from the joint density.

Remark 1: We have considered the bivariate case, but it is easily extended to the multivariate cases. That is, it is possible to take multi-dimensional vectors for x and

y. Taking an example, as for the tri-variate random vector (X, Y, Z), if we generate the ith random draws from $f_{x|yz}(x|y_{i-1}, z_{i-1})$, $f_{y|xz}(y|x_i, z_{i-1})$ and $f_{z|xy}(z|x_i, y_i)$, sequentially, we can obtain the random draws from $f_{xyz}(x, y, z)$.

Remark 2: Let X, Y and Z be the random variables. Take an example of the case where X is highly correlated with Y. If we generate random draws from $f_{x|yz}(x|y, z)$, $f_{y|xz}(y|x, z)$ and $f_{z|xy}(z|x, y)$, it is known that convergence of the Gibbs sampler is slow. In this case, without separating X and Y, random number generation from $f(x, y|z)$ and $f(z|x, y)$ yields better random draws from the joint density $f(x, y, z)$.

6.2.2 Metropolis-Hastings Algorithm

This section is based on Geweke and Tanizaki (2003), where three sampling distributions are compared with respect to precision of the random draws from the target density $f(x)$.

The **Metropolis-Hastings algorithm** is also one of the sampling methods to generate random draws from any target density $f(x)$, utilizing sampling density $f_*(x)$, even in the case where it is not easy to generate random draws from the target density.

Let us define the acceptance probability by:

$$\omega(x_{i-1}, x^*) = \min\Big(\frac{q(x^*)}{q(x_{i-1})}, 1\Big) = \min\Big(\frac{f(x^*)/f_*(x^*)}{f(x_{i-1})/f_*(x_{i-1})}, 1\Big),$$

where $q(\cdot)$ is defined as equation (**??**). By the Metropolis-Hastings algorithm, a random draw from $f(x)$ is generated in the following way:

(i) Take the initial value of x as x_{-M}.

(ii) Generate x^* from $f_*(x)$ and compute $\omega(x_{i-1}, x^*)$ given x_{i-1}.

(iii) Set $x_i = x^*$ with probability $\omega(x_{i-1}, x^*)$ and $x_i = x_{i-1}$ otherwise.

(iv) Repeat Steps (ii) and (iii) for $i = -M+1, -M+2, \cdots, 1$.

In the above algorithm, x_1 is taken as a random draw from $f(x)$. When we want more random draws (say, N), we replace Step (iv) by Step (iv)', which is represented as follows:

(iv)' Repeat Steps (ii) and (iii) for $i = -M + 1, -M + 2, \cdots, N$.

When we implement Step (iv)', we can obtain a series of random draws x_{-M}, x_{-M+1}, $\cdots$, x_0, x_1, x_2, $\cdots$, x_N, where x_{-M}, x_{-M+1}, $\cdots$, x_0 are discarded from further consideration. The last N random draws are taken as the random draws generated from the target density $f(x)$. Thus, N denotes the number of random draws. M is sometimes called the **burn-in period**.

We can justify the above algorithm given by Steps (i) – (iv) as follows. We show that x_i is the random draw generated from the target density $f(x)$ under the assumption x_{i-1} is generated from $f(x)$. Let U be the uniform random variable between zero and one, X be the random variable which has the density function $f(x)$ and x^* be