When $h(\theta; w_t)$, $t = 1, 2, \cdots, T$, are not serially correlated, the following $\hat{S}_T$ is consistent, i.e.,

$$\hat{S}_T = \frac{1}{T}\sum_{t=1}^{T} h(\hat{\theta}_T; w_t)h(\hat{\theta}_T; w_t)' \longrightarrow S.$$

When $h(\theta; w_t)$, $t = 1, 2, \cdots, T$, are serially correlated,

$$\hat{S}_T = \hat{\Gamma}(0) + \sum_{\tau=1}^{q} k(\frac{\tau}{q+1})(\hat{\Gamma}(\tau) + \hat{\Gamma}(\tau)'),$$

where $\hat{\Gamma}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} h(\hat{\theta}_T; w_t)h(\hat{\theta}_T; w_{t-s})'$.

$k(x) = 1 - x \implies$ Bartlett kernel (Newwey-west estimator),

$k(x) \implies$ Parzen kernel, and etc.

Then, we obtain:

$$\sqrt{T}(\hat{\theta}_T - \theta) \longrightarrow N\left(0, (DS^{-1}D')^{-1}\right),$$

where

$$D = \frac{\partial g(\theta; W_T)}{\partial \theta'}.$$

Note that D is a $r \times k$ matrix.

Let $\hat{D}_T$ be an estimate of D.

The variance estimator of $\hat{\theta}_T$ is given by:

$$\hat{D}_T = \frac{\partial g(\hat{\theta}_T; W_T)}{\partial \theta'}.$$

Asymptotic Normality:

$$\text{Assumption 1}: \quad \hat{\theta}_T \longrightarrow \theta,$$

$$\text{Assumption 2}: \quad \sqrt{T} g(\theta; W_T) \longrightarrow N(0, S).$$

Then, we have the following first-order approximation:

$$\begin{aligned} g(\theta; W_T) &\approx g(\hat{\theta}_T; W_T) + \frac{\partial g(\hat{\theta}_T; W_T)}{\partial \theta'}(\theta - \hat{\theta}_T) \\ &= g(\hat{\theta}_T; W_T) + \hat{D}_T(\theta - \hat{\theta}_T), \end{aligned}$$

where $g(\theta; W_T)$ is linearized around $\theta = \hat{\theta}_T$.

The first-order condition for the minimization problem is:

$$\left(\frac{\partial g(\theta; W_T)}{\partial \theta'}\right)' S^{-1}\Big(g(\theta; W_T)\Big) = 0.$$

Substituting the approximation into the above equation, we obtain the following:

$$\begin{aligned} D'S^{-1}\Big(g(\theta; W_T)\Big) &= D'S^{-1}\Big(g(\hat{\theta}_T; W_T) + \hat{D}_T(\theta - \hat{\theta}_T)\Big) \\ &= D'S^{-1}g(\hat{\theta}_T; W_T) + D'S^{-1}\hat{D}_T(\theta - \hat{\theta}_T). \end{aligned}$$

Therefore,

$$\sqrt{T}(\hat{\theta}_T - \theta) \approx (D'S^{-1}\hat{D}_T)^{-1}D'S^{-1}\sqrt{T}\Big(g(\hat{\theta}_T; W_T) - g(\theta; W_T)\Big).$$

Thus, GMM estimator, $\hat{\theta}_T$, has the following asymptotic distribution:

$$\sqrt{T}(\hat{\theta}_T - \theta) \longrightarrow N\Big(0, (D'S^{-1}D)^{-1}\Big),$$

where $\hat{D}_T \longrightarrow D$ is utilized.

From Assumption 2, we have the following asymptotic distribution:

$$\Big(\sqrt{T}g(\theta; W_T)\Big)' S^{-1} \Big(\sqrt{T}g(\theta; W_T)\Big) \longrightarrow \chi^2(r).$$

When θ is replaced by GMM estimator $\hat{\theta}_T$, we have the following distribution:

$$\left(\sqrt{T} g(\hat{\theta}_T; W_T)\right)' \hat{S}_T^{-1} \left(\sqrt{T} g(\hat{\theta}_T; W_T)\right) \longrightarrow \chi^2(r-k),$$

which is called a test of the overidentifying restrictions.

$\Longrightarrow$ J test by Hansen (1982)

k linear combinations consisting of a $r \times 1$ vector $g(\hat{\theta}_T; W_T)$ are zeros.

Therefore, the degrees of freedom are $r-k$.

Some Examples:

(a) **OLS:**

Regression Model: $\quad y_t = x_t\beta + \epsilon_t, \quad \mathrm{E}(x_t\epsilon_t) = 0$

$h(\theta; w_t)$ is taken as:

$$h(\theta; w_t) = x_t(y_t - x_t\beta).$$

(b) **IV (Instrumental Variable, 操作変数法):**

Regression Model: $y_t = x_t\beta + \epsilon_t$, $\mathrm{E}(x_t\epsilon_t) \neq 0$, $\mathrm{E}(z_t\epsilon_t) = 0$

$h(\theta; w_t)$ is taken as:

$$h(\theta; w_t) = z_t(y_t - x_t\beta),$$

where z_t is a vector of instrumental variables.

(c) **NLS (Nonlinear Least Squares,** 非線形最小二乗法):

Regression Model: $f(y_t, x_t, \beta) = \epsilon_t$, $\mathrm{E}(x_t\epsilon_t) \neq 0$, $\mathrm{E}(z_t\epsilon_t) = 0$

$h(\theta; w_t)$ is taken as:

$$h(\theta; w_t) = z_t f(y_t, x_t, \beta)$$

where z_t is a vector of instrumental variables.

11 Bayesian Estimation (ベイズ推定)

Greenberg, E. (2013) *Introduction to Bayesian Econometrics* (2nd ed.)

安藤知寛 (2010) 『ベイズ統計モデリング』 (朝倉書店)

豊田秀樹編 (2008) 『マルコフ連鎖モンテカルロ法』 (朝倉書店)

Dey, D.K. and Rao, C.R., (2005) *Handbook of Statistics, Vol.25: Bayesian Thinking: Modeling and Computation*

繁桝・岸野・大森監訳 (2011) 『ベイズ統計分析ハンドブック』 (朝倉書店)

11.1 Introduction

Two Events: A and B

Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Posterior Distribution (事後分布): $f_{\theta|y}(\theta|y)$:

$$f_{\theta|y}(\theta|y) = \frac{f_{y|\theta}(y|\theta)f_\theta(\theta)}{f_y(y)} = \frac{f_{y|\theta}(y|\theta)f_\theta(\theta)}{\int f_{y|\theta}(y|\theta)f_\theta(\theta)\mathrm{d}\theta} \propto f_{y|\theta}(y|\theta)f_\theta(\theta),$$

where $f_\theta(\theta)$ is called the prior distribution (事前分布).

Example 1: Let x be the number of successes in a series of n trials with probability θ of success in each.

That is, x has the binomial probability function, given θ,

$$f_{x|\theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, \cdots, n.$$

θ is assumed to be the beta distribution:

$$f_\theta(\theta) = \frac{1}{B(p,q)}\theta^{p-1}(1-\theta)^{q-1},$$

for $\leq \theta \leq 1$, which corresponds to a prior distribution.

Before applying Bayes' theorem, $f_x(x)$ is given by:

$$\begin{aligned} f_x(x) &= \int f_{x|\theta}(x|\theta) f_\theta(\theta) \mathrm{d}\theta \\ &= \binom{n}{r} \frac{1}{B(p,q)} \int_0^1 \theta^{p+x-1}(1-\theta)^{q+n-x-1} \mathrm{d}\theta \\ &= \binom{n}{r} \frac{B(p+x, q+n-x)}{B(p,q)}. \end{aligned}$$

The posterior distribution of θ is:

$$f_{\theta|x}(\theta|x) = \frac{1}{B(p+x, q+n-x)} \theta^{p+x-1}(1-\theta)^{q+n-x-1},$$

which is also a beta distribution with prameters $p + x$ and $q + n - x$.

The posterior mean and variance are:

$$\mathrm{E}(\theta|x) = \frac{p+x}{p+q+n}, \qquad \mathrm{V}(\theta|x) = \frac{(p+x)(q+n-x)}{(p+q+n)^2(p+q+n+1)}.$$

Example 2: $x|\theta \sim N(\theta, v)$, where v is known.

$\theta \sim N(m, w)$, where m and w are known. $\Longrightarrow$ prior dist.

Then, the posterior distribution of θ is:

$$\theta | x \;\sim\; N\Big(\frac{wx + vm}{w + v}, \frac{vw}{w + v}\Big).$$

Example 3: $x_1,\ x_2,\ \cdots,\ x_n$ are mutually independently and identically distributed as $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

$$\begin{aligned} f_{x|\theta}(x|\theta) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\big(s^2 + n(\overline{x} - \mu)^2\big)\Big), \end{aligned}$$

where $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$ and $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

The prior density is:

$$f_\theta(\theta) = k(a, b, w)\sigma^{b+3} \exp\Big(-\frac{1}{2\sigma^2}\Big(a + \frac{(\mu - m)^2}{w}\Big)\Big),$$

where $k(a, b, w) = \dfrac{a^{b/2} 2^{-(b+1)/2} (\pi w)^{-1/2}}{\Gamma(\frac{1}{2}b)}$ is a constant.

The posterior density is:

$$f_{\theta|x}(\theta|x) = k(a_1, b_1, w_1)\sigma^{-(b_1+3)} \exp\left(-\frac{1}{2\sigma^2}\left(a_1 + \frac{(\mu - m_1)^2}{w_1}\right)\right),$$

where $\quad w_1 = \dfrac{w}{1+nw}, \quad m_1 = \dfrac{m + nw\overline{x}}{1+nw}, \quad b_1 = b + n, \quad a_1 = a + s^2 + \dfrac{n(\overline{x} - m)^2}{1+nw}.$

Inference on μ: The posterior density of μ is:

$$f(\mu|x) = \int_0^\infty f(\theta|x)\mathrm{d}\sigma^2 = k_\mu(t_1, b_1)\left(1 + \frac{(\mu - m_1)^2}{b_1 t_1}\right)^{-(b_1+1)/2},$$

where $\quad t_1 = \dfrac{w_1 a_1}{b_1} \quad$ and $\quad k_\mu(t_1, b_1) = \dfrac{1}{\sqrt{t_1 k_1} B(\frac{1}{2}, \frac{1}{2}b_1)}.$

Thus, $\dfrac{\mu - m_1}{\sqrt{t_1}}$ has a t distribution with b_1 degrees of freedom.

Inference of σ^2: The posterior density of σ^2 is:

$$f(\sigma^2|x) = \int_{-\infty}^{\infty} f(\theta|x)\mathrm{d}\mu = k_{\sigma^2}(a_1, b_1)\sigma^{-(b_1+2)} \exp\left(-\frac{a_1}{2\sigma^2}\right),$$

where $\quad k_{\sigma^2}(a_1, b_1) = \dfrac{(\frac{1}{2}a_1)^{b_1/2}}{\Gamma(\frac{1}{2}b_1)}.$

Thus, $\dfrac{a_1}{\sigma^2}$ is chi-squared with b_1 degrees of freedom.

11.2 Inference

Posterior Distribution (事後分布): $f_{\theta|y}(\theta|y)$

11.2.1 Point Estimate

Posterior Mean (事後平均):

$$\bar{\theta} = \int_{-\infty}^{\infty} \theta f_{\theta|y}(\theta|y) \mathrm{d}\theta.$$

Posterior Mode (事後モード):

$$\hat{\theta} = \mathrm{argmax}_{\theta}\ f_{\theta|x}(\theta|y).$$

Posterior Median (事後メディアン):

$$\tilde{\theta} \text{ such that } \int_{-\infty}^{\tilde{\theta}} f_{\theta|y}(\theta|y) \mathrm{d}\theta = 0.5.$$

11.2.2 Interval Estimate

$$\int_R f_{\theta|y}(\theta|y)\mathrm{d}\theta = 1 - \alpha,$$

where R is called confidence interval.

Bayesian confidence interval (ベイズ信頼区間) or credible interval (信用区間):

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha.$$

θ_L and θ_U lead to lower and upper bounds.

(θ_L, θ_U) is called Bayesian confidence interval or credible interval.

Highest posterior density interval (最高事後密度区間):

$$f_{\theta|y}(\theta_0|y) \geq f_{\theta|y}(\theta_1|y), \qquad \text{for } \theta_0 \in R \text{ and } \theta_1 \notin R.$$

11.2.3 Marginal Likelihood (周辺尤度)

Marginal Likelihood $\Longrightarrow$ Fitness of the Model:

$$f_y(y) = \int f_{y|\theta}(y|\theta) f_\theta(\theta)\mathrm{d}\theta,$$

which corresponds to the denominator in the posterior distribution.

11.3 Example: Linear Regression

Regression Model:

$$y = X\beta + u, \qquad u \sim N(0, \sigma^2 I_n),$$

where y and u are $n \times 1$ vectors, X is an $n \times k$ matrix and β is a $k \times 1$ vector.

Likelihood Function: $\theta = (\beta, \sigma^2)$

$$f_{y|\theta}(y|\theta) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\Big)$$

Prior Distributions:

$$f_\theta(\beta, \sigma^2) = f_{\beta|\sigma^2}(\beta|\sigma^2) f_{\sigma^2}(\sigma^2),$$

where

$$f_{\beta|\sigma^2}(\beta|\sigma^2) = N(\beta_0, \sigma^2 A^{-1}) = (2\pi\sigma^2)^{-k/2}|A|^{1/2} \exp\Big(-\frac{1}{2\sigma^2}(\beta - \beta_0)'A(\beta - \beta_0)\Big),$$

$$f_{\sigma^2}(\sigma^2) = IG\Big(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\Big) = \frac{(\lambda_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}(\sigma^2)^{-\nu_0/2-1} \exp\Big(-\frac{\lambda_0}{2\sigma^2}\Big).$$

β_0, A, ν_0 and λ_0 are called the hyper-parameters.

Note that $Y \sim IG(a, b)$ for $X \sim G(a, b)$ and $Y = \dfrac{1}{X}$.

The posterior distribution of β and σ^2 is:

$$\begin{aligned}
f_{\theta|y}(\beta, \sigma^2|y) &\propto f_{y|\theta}(y|\beta, \sigma^2) f_{\beta|\sigma^2}(\beta|\sigma^2) f_{\sigma^2}(\sigma^2) \\
&= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\Big) \\
&\qquad \times (2\pi\sigma^2)^{-k/2}|A|^{1/2} \exp\Big(-\frac{1}{2\sigma^2}(\beta - \beta_0)'A(\beta - \beta_0)\Big) \\
&\qquad \times \frac{(\lambda_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}(\sigma^2)^{-\nu_0/2-1} \exp\Big(-\frac{\lambda_0}{2\sigma^2}\Big) \\
&\propto (\sigma^2)^{-(n+k+\nu_0)/2-1} \exp\Big(-\frac{(y - X\beta)'(y - X\beta) + (\beta - \beta_0)'A(\beta - \beta_0) + \lambda_0}{2\sigma^2}\Big) \\
&\propto |\sigma^2\hat{A}|^{-1/2} \exp\Big(-\frac{(\beta - \hat{\beta})'\hat{A}^{-1}(\beta - \hat{\beta})}{2\sigma^2}\Big) \times (\sigma^2)^{-\hat{\nu}/2-1} \exp\Big(-\frac{\hat{\lambda}}{2\sigma^2}\Big) \\
&\propto f_{\beta|\sigma^2,y}(\beta|\sigma^2, y) \times f_{\sigma^2|y}(\sigma^2|y) = N(\hat{\beta}, \sigma^2\hat{A}) \times IG(\frac{\hat{\nu}}{2}, \frac{\hat{\lambda}}{2})
\end{aligned}$$

where

$$\hat{\beta} = (X'X + A)^{-1}(X'X\hat{\beta}_{OLS} + A\beta_0), \qquad \hat{\beta}_{OLS} = (X'X)^{-1}X'y,$$

$$\hat{A} = (X'X + A)^{-1}, \qquad \hat{\nu} = \nu_0 + n,$$

$$\hat{\lambda} = \lambda_0 + (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta_0 - \hat{\beta}_{OLS})'((X'X)^{-1} + A^{-1})^{-1}(\beta_0 - \hat{\beta}_{OLS}).$$

The marginal posterior distribution of β is:

$$f_{\beta|y}(\beta|y) = \int f_{\theta|y}(\beta, \sigma^2|y)\mathrm{d}\sigma^2 = \int f_{\beta|\sigma^2,y}(\beta|\sigma^2, y) f_{\sigma^2|y}(\sigma^2|y)\mathrm{d}\sigma^2$$
$$\propto \left(1 + \frac{1}{\hat{\nu}}(\beta - \hat{\beta})'\left(\frac{\hat{\lambda}}{\hat{\nu}}\hat{A}\right)^{-1}(\beta - \hat{\beta})\right)^{-(\hat{\nu}+k)/2},$$

which is a k-dimensional t distribution with parameters $\hat{\beta}$, $\dfrac{\hat{\lambda}}{\hat{\nu}}\hat{A}$ and $\hat{\nu}$.

Note that the k-dimensional t distribution with parameters μ, Σ and ν is given by:

$$f(x) = \frac{\Gamma(\frac{\nu+k}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{k/2}}|\Sigma|^{-1/2}\left(1 + \frac{1}{\nu}(x - \mu)'\Sigma^{-1}(x - \mu)\right)^{-(\nu+k)/2}.$$

The marginal likelihood is:

$$f_y(y) = \frac{f_{y|\theta}(y|\theta) f_\theta(\theta)}{f_{\theta|y}(\theta|y)} = \frac{|\hat{A}|^{1/2}|A|^{1/2}(\lambda_0/2)^{\nu_0/2}\Gamma(\hat{\nu}/2)}{\pi^{n/2}\Gamma(\nu_0/2)(\hat{\lambda}/2)^{\hat{\nu}/2}},$$

which is utilized for model selection.

In general, how do we evaluate $f_{\theta|y}(\theta|y)$, $\mathrm{E}(\theta|y)$, $f_y(y)$ and so on?

11.4 On Prior Distribution

11.4.1 Non-informative Prior

$$f_\theta(\theta) = \text{ const.}$$

In this case, the posterior distribution is:

$$f_{\theta|y}(\theta|y) \propto f_{y|\theta}(y|\theta),$$

which is proportional to the likelihood function.

However, we have the case where the integration of prior diverges, i.e.,

$$\int f_\theta(\theta)\mathrm{d}\theta = \infty.$$

In this case, $f_\theta(\theta)$ is called an improper prior.

11.4.2 Jeffreys' Prior

$$f_\theta(\theta) \propto |J(\theta)|^{1/2},$$

where

$$J(\theta) = -\int \frac{\partial^2 \log f_{y|\theta}(y|\theta)}{\partial\theta\partial\theta'} f_{y|\theta}(y|\theta)\mathrm{d}y = -\mathrm{E}\left(\frac{\partial^2 \log f_{y|\theta}(y|\theta)}{\partial\theta\partial\theta'}\right),$$

which is Fisher's information matrix.