

Lecture VI: Sufficient and Ancillary Statistics, and Conjugate Families

Charles B. Moss

May 8, 2013

- I. Returning to the conditional density function at the base of Bayesian estimation

$$p(\theta_A | y, A) = \frac{p(\theta_A | A) p(y | \theta_A, A)}{p(y | A)} \quad (1)$$

a priori becoming

$$p(\theta_A | y^0, A) = \frac{p(\theta_A | A) p(y^0 | \theta_A, A)}{p(y^0 | A)} \quad (2)$$

ex post. The derivation of the posterior distribution $p(\theta_A | y^0, A)$ and the marginal likelihood function $p(y^0 | A)$ depends on the relationship between y^0 and θ_A in each function. This relationship can be developed within the context of *sufficient* and *ancillary* statistics.

- A. **Definition 2.2.1** The vector $s = s(y; A)$ is a *sufficient statistic* in a model with the observables density $p(y | \theta_A, A)$ if

$$p[y | s(y; A), \theta_A, A] = p[y | s(y; A), A] \quad \forall \theta_A \in \Theta_A \quad (3)$$

1. The best way to introduce sufficient statistics is by example. The simplest example, unfortunately, is the normal distribution. Starting with the univariate normal

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{\sum_{i=1}^N (x_i - \mu)^2}{2\sigma^2} \right]. \quad (4)$$

Next, we introduce the sample mean (adding a special form of zero – $\bar{x} - \bar{x}$)

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{\sum_{i=1}^N (x_i - \bar{x} + \bar{x} - \mu)^2}{2\sigma^2} \right]. \quad (5)$$

Next, we want to justify

$$\sum_{i=1}^N (x_i - \bar{x})^2 + \sum_{i=1}^N (\bar{x} - \mu)^2 \quad (6)$$

to do this consider splitting Equation 5 into parts

$$\begin{aligned} \sum_{i=1}^N (x_i - \bar{x} + \bar{x} - \mu)^2 &= \sum_{i=1}^N [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \\ &= \sum_{i=1}^N [(x_i - \bar{x})^2 + (x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2] \end{aligned} \quad (7)$$

$$\begin{aligned} \ni: \sum_{i=1}^N (x_i - \bar{x})(\bar{x} - \mu) &= (\bar{x} - \mu) \sum_{i=1}^N (x_i - \bar{x}) = 0 \\ \Rightarrow \sum_{i=1}^N (x_i - \bar{x})^2 &+ \sum_{i=1}^N (\bar{x} - \mu)^2 \end{aligned}$$

Thus, Equation 4 can be rewritten as

$$f(x | \bar{x}(x), \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{2\sigma^2} \right] \exp \left[-\frac{n(\bar{x} - \mu)^2}{2\sigma^2} \right] \quad (8)$$

There are a couple of ways to look at the result in Equation 8. First, if we assume that $\bar{x} \rightarrow \mu$ (or that the sample mean is a good estimate of the population mean) $\bar{x}$ replaces μ in the sample

$$f(x | \bar{x}(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{2\sigma^2} \right] \quad (9)$$

The second, is the factorization criteria in **Theorem 2.2.2**.

B. Theorem 2.2.1 Ex Post and Ex Ante Equivalence of Sufficiency The vector $s(y; A)$ is a sufficient statistic in a model with observables density $p(y | \theta_A, A)$ if and only if for all $\theta_A \in \Theta_A$ and for all $y \in \Psi$, we have

$$p(\theta_A | y, A) = p(\theta_A | s(y; A), A) \quad (10)$$

1. Again back to the nuts and bolts of sufficiency

$$P[y | s(y; A), \theta_A, A] = p[y | s(y; A), A] \quad (11)$$

implies that all the information in θ_A is included in $s(y; A)$.

2. Working forward

$$\begin{aligned} p(\theta_A | y, A) &= p(\theta_A | y, s, A) \\ &= \frac{p(y | \theta_A, s, A) p(\theta_A | s, A)}{p(y | s, A)} \\ &\Rightarrow \frac{p(y, \theta_A | s, A)}{p(y | s, A)} \end{aligned} \quad (12)$$

3. This proof takes a similar approach based on $p(y|s, \theta_A, A)$

$$p(y|s, \theta_A, A) = \frac{p(\theta_A|y, s, A) p(y|s, A)}{p(\theta_A|y, A)} \quad (13)$$

Note $p(\theta_A|y, s, A) = p(\theta_A|y, A)$ or s and y contain the same information

$$\therefore p(y|s, \theta_A, A) = \frac{p(\theta_A|y, A) p(y|s, A)}{p(\theta_A|y, A)} = p(y|s, A) \quad (14)$$

C. **Theorem 2.2.2 Factorization Criteria** The vector $s(y; A)$ is a sufficient statistic in a model with data density $p(y|\theta_A, A)$ if and only if

$$p(y|\theta_A, A) = p[s(y; A)|\theta_A, A] r(y; A) \quad (15)$$

for some function $r(y; A)$. See Equation 8 for the factorization of the sample mean.

D. **Definition 2.2.2** Suppose that $s(y; A)$ is a sufficient statistic in the observables density $p(y|\theta_A, A)$. If there exists partitions $s' = (s'_1, s'_2)$ and $\theta'_A = (\theta'_{A1}, \theta'_{A2})$ such that

$$p(\theta_A|A) = p(\theta_{A1}|A) p(\theta_{A2}|A), \quad (16)$$

$$p(s_1|\theta_A, A) = p(s_1|\theta_{A1}, A). \quad (17)$$

$$p(s_2|s_1, \theta_A, A) = p(s_2|s_1, \theta_{A2}, A), \quad (18)$$

then s_1 is an *ancillary statistic* with respect to θ_{A2} . Building the implications of the ancillary statistics

$$p(\theta_A|y, A) \propto p(\theta_A|s, A) \quad (19)$$

by Equation 16

$$p(\theta_A|y, A) \propto p(\theta_{A1}|A) p(\theta_{A2}|A) \times p(s_1|\theta_A, A) p(s_2|s_1, \theta_A, A). \quad (20)$$

Note in Equation 20 $p(s_1 | \theta_A, A) p(s_2 | s_1, \theta_A, A) = p(s_1, s_2 | \theta_A, A)$.
Rearranging Equation 20

$$p(\theta_A | y, A) \propto [p(\theta_{A1} | A) p(s_1 | \theta_{A1}, A)] [p(\theta_{A2} | A) p(s_2 | s_1, \theta_{A2}, A)] \quad (21)$$

1. "It simplifies inference when the vector of interest ω depends on θ_{A2} and y , but not θ_{A1} , or, alternatively when ω depends on θ_{A1} and y , but not θ_{A2} ."

$$\begin{aligned} p(\theta_{A2} | y, A) &\propto p(\theta_{A2} | A) p(s_2 | s_1, \theta_{A2}, A) \\ \Rightarrow p(\omega | y, A) &= \int_{\Theta_A} p(\omega | y, \theta_A, A) p(\theta_A | y, A) d\nu(\theta_A) \\ &\propto \int_{\Theta_{A2}} p(\omega | y, \theta_{A2}, A) p(\theta_{A2} | A) p(s_2 | s_1, \theta_{A2}, A) d\nu(\theta_{A2}) \end{aligned} \quad (22)$$

"This means that for purposes of learning about ω from the data y^0 it is necessary only to use the prior density of θ_{A2} , $p(\theta_{A2} | A)$ and the conditional density of s_2 , $p(s_2 | s_1, \theta_{A2}, A)$ (Geweke)."

II. Nuisance Parameters

- A. "If $\theta_A = (\theta'_{A1}, \theta'_{A2})$ and $p(\omega | y, \theta_A, A) = p(\omega | y, \theta_{A2}, A)$ but there is no ancillary statistic with respect to θ_{A2} , then θ_{A1} is a vector of nuisance parameters (Geweke)."

 1. Basically, the parameters θ_{A1} affect the probability density function with respect to θ_{A2} . Further, there is no ancillary statistic that can be used to condition θ_{A2} .
 2. "In non-Bayesian econometrics nuisance parameters can be troublesome, because test and related statistics pertaining to $E[h(\omega) | y, \theta_A, A]$ depend on the value of the unknown parameter vector θ_{A2} (Geweke)."

- B. "Nuisance parameters never present any difficulty in Bayesian inference, because they are marginalized in the posterior of ω :"

$$\begin{aligned}
 p(\omega | y^0, A) &= \int_{\Theta_A} p(\omega | y^0, \theta_A, A) p(\theta_A | y^0, A) d\nu(\theta_A) \\
 &= \int_{\Theta_{A2}} p(\omega | y^0, \theta_{A2}, A) p(\theta_{A2} | y^0, A) \\
 &\quad \times \left[\int_{\Theta_{A1}} p(\theta_{A1} | \theta_{A2}, y^0, A) d\nu(\theta_{A1}) \right] d\nu(\theta_{A2}) \\
 &= \int_{\Theta_{A2}} p(\omega | y^0, \theta_{A2}, A) p(\theta_{A2} | y^0, A) d\nu(\theta_{A2}).
 \end{aligned} \tag{23}$$

The Nuisance parameters are integrated out.

III. Conjugate Prior Distribution

- A. "The densities $p(\theta_A | A)$ and $p(Y_T | \theta_A, A)$ together represents a belief regarding the observables Y_T (Geweke)." Essentially, the sample distribution and the prior distribution for the parameters of the sample distribution (i.e., the prior distribution) forms a complete distribution for the observable variables.
- B. There are several considerations in choosing these pairs of distributions including the characteristics of the observed variables (i.e., index numbers could be hypothesized to follow a Gamma distribution while the regression variable for the income elasticity of demand may be better modelled as a normal variable), but the pairing (e.g., the combination of the sampling distribution and the prior) may be largely determined by the tractability of joint distribution

$$p(\theta_A, Y_T^0, A) \propto p(\theta_A | A) p(Y_T^0 | \theta_A, A) \tag{24}$$

- C. **Definition 2.3.1** Suppose that the observables density $p(Y_T | \theta_A, A)$ has the $r \times 1$ sufficient statistic vector $s_T = s_T(Y_T; A)$, that r is fixed as T varies, and $(s_T)_+ = T$. Denote the corresponding likelihood function

$$\begin{aligned}
 L(\theta_A; s_T^0, A) &= L(\theta_A; s_T(Y_T^0, A), A) \\
 &= L(\theta_A; Y_T^0, A) \propto p(Y_T^0 | \theta_A, A)
 \end{aligned} \tag{25}$$

Then the *conjugate family of prior densities* with respect to $p(Y_T^0 | \theta_A, A)$ is $\{ p(\theta_A | \gamma_A, A), \gamma_A \in \Gamma_A \}$, where

$$p(\theta_A | \gamma_A, A) \propto L(\theta_A; \gamma_A, A) = k(\theta_A | \gamma_A, A) \quad (26)$$

and

$$\Gamma_A = \left\{ \gamma_A : \int_{\Theta_A} k(\theta_A | \gamma_A, A) d\nu(\theta_A) < \infty \right\}. \quad (27)$$

D. **Definition 2.3.2** In the data density $p(Y_T | \theta_A, A)$ and $\theta'_A = (\theta'_{A1}, \theta'_{A2})$ fix $\theta_{A2} = \theta_{A2}^0$. Suppose that the data density $p(Y_T | \theta_{A1}, \theta_{A2}^0, A)$ has the $r^* \times 1$ sufficient statistic vector $s_T^* = s_T^*(Y_T; \theta_{A2}^0, A)$, that r^* is fixed as T varies, and $(s_T^*)_1 = T$. Denote the corresponding partial likelihood function

$$L(\theta_{A1}; \theta_{A2}^0, s_T^0, A) = L(\theta_{a1}; \theta_{A2}^0, Y_T^0, A) \propto p(Y_T^0 | \theta_{A1}, \theta_{A2}^0, A) \quad (28)$$

Then the *conditionally conjugate family of prior densities with respect to* $p(Y_T, \theta_{A1}, \theta_{A2}^0, A)$ is $\{ p(\theta_{a1} | \gamma_A^*, \theta_{A2}^0, A), \gamma_A^* \in \Gamma_A^* \}$ where

$$p(\theta_{A1} | \gamma_A^*, \theta_{A2}^0, A) \propto L(\theta_{A1}; \theta_{A2}^0, \gamma_A^*, A) = k(\theta_{A1} | \gamma_A^*, \theta_{A2}^0, A) \quad (29)$$

and

$$\Gamma_A^* = \left\{ \gamma_A^* : \int_{\Theta_{A1}} k(\theta_{A1} | \gamma_A^*, \theta_{A2}^0, A) d\nu(\theta_{A1}) < \infty \right\} \quad (30)$$

E. **Definition 2.3.3** The *exponential family of distributions* consists of all distributions with the observables density

$$p(Y_T | \theta_A, A) = [g(\theta_A)]^T \left[\prod_{t=1}^T f(y_t) \right] \exp \left\{ \sum_{t=1}^T c_i \phi_i(\theta_A) \left[\sum_{t=1}^T h_t(y_t) \right] \right\}, \quad (31)$$

where $\Theta \supset \theta_A$, $f(\cdot)$ and $\{c_i, \phi(\cdot), h_i(\cdot)\}_{i=1}^t$ are all specified.