

Economics 696, Lecture Note 4: Normal and Multivariate Normal Likelihood Functions

1 Introduction

Our general program is to specify a likelihood function, choose a prior distribution for the parameters of the likelihood function, and calculate the posterior distribution. In this lecture note we will discuss likelihood functions based on normal and multivariate normal densities.

2 Models Based on Normality

2.1 Likelihood Functions

Example 1 (Classical Regression Model)

Suppose that we observe a random variable $Z = (Z_1, \dots, Z_n)$ with $Z_i = (Y_i, X_i)$, where Y_i is 1×1 and X_i is $k \times 1$. P_θ specifies that

$$Y_i | X_1 = x_1, \dots, X_n = x_n, \beta, \sigma \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(x_i' \beta, \sigma^2), \quad i = 1, \dots, n.$$

When $X_i \equiv 1, \forall i$, we have a simple normal model with mean β and variance σ^2 .

By independence, the density of the Y_i 's given X_1, \dots, X_n can be written as

$$f(y_1, \dots, y_n | x_1, \dots, x_n, \beta, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i' \beta)^2\right).$$

We will call this the conditional or *partial likelihood* of (y_1, \dots, y_n) given (x_1, \dots, x_n) . In some cases, we are only interested in parameters related to the distribution of Y given X , not in parameters related to the marginal distribution of X , so focusing on the partial likelihood may be appropriate. The issue of conditioning on regressors will be discussed in more detail below.

Example 2 (Autoregressive Model)

$Z = (Y_0, Y_1, \dots, Y_T)$, and P_θ specifies that

$$Y_0 | \xi \sim F_\xi, \tag{1}$$

where $\{F_\xi : \xi \in \Xi\}$ is some given class of probability distributions, and, for $t = 1, \dots, T$,

$$Y_t | Y_0 = y_0, Y_1 = y_1, \dots, Y_{t-1} = y_{t-1}, \beta, \sigma, \xi \sim \mathcal{N}(\beta_1 + \beta_2 y_{t-1}, \sigma^2). \tag{2}$$

Recall that any joint probability density function can be **sequentially decomposed**:

$$f(y_0, y_1, \dots, y_T) = f_0(y_0) \times f_{1|0}(y_1 | y_0) \times f_{2|0,1}(y_2 | y_0, y_1) \times \dots \times f_{T|0,\dots,T-1}(y_T | y_0, \dots, y_{T-1}),$$

where $f_{t|0,\dots,t-1}$ is the conditional density function of Y_t given Y_0, \dots, Y_{t-1} . Thus (1) and (2) completely specify the joint distribution of Z , and we can write $\theta = (\xi, \beta, \sigma)$.

Suppose that we are only interested in β, σ , not ξ . Using the sequential decomposition, we can write down the partial likelihood function, conditional on y_1 , as

$$f(y_1, \dots, y_T | y_0, \theta) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_t - \beta_1 - \beta_2 y_{t-1})^2\right).$$

Remarkably, this likelihood function has exactly the same form as the likelihood function in the classical regression model, even though the autoregressive model has a very different conditional dependence structure. This similarity in the likelihood functions is important, because it means that the posterior distributions can be calculated in exactly the same way for both models.

Likelihood Function: Suppose that $Z = (Z_1, \dots, Z_n)$, where $Z_i = (Y_i, X_i)$, and suppose the partial likelihood function has the form

$$f_\theta = \prod_{i=1}^n l(y_i | x_i, \theta),$$

where

$$l(y_i | x_i, \theta) = \mathcal{N}(x_i' \beta, \sigma^2).$$

(To see that the autoregressive model fits in this framework, set $X_i \equiv (1, Y_{i-1})$.) Let

$$X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

be the $n \times k$ matrix of regressors and the $n \times 1$ vector of responses, respectively. Recall that a common estimator of β is the least squares estimator

$$b = (X'X)^{-1} X'y$$

and that the typical estimator of σ^2 is given by

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - x_i' b)^2.$$

2.2 Prior and Posterior Distributions: Version 1

The χ^2 Distribution: If

$$w_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for } j = 1, \dots, \nu,$$

then

$$w \equiv \sum_{j=1}^{\nu} w_j^2 \sim \chi_{\nu}^2.$$

The χ^2 distribution has a density function

$$f_{\chi_{\nu}^2}(w) = c \cdot w^{\frac{\nu-2}{2}} \exp\left(-\frac{1}{2}w\right) I_{(0,\infty)}(w) \quad (3)$$

Here c is a constant, which does not depend on w , such that (3) integrates to 1. $I_{(0,\infty)}(w)$ is the indicator function, equal to 1 if $w \in (0, \infty)$ and equal to 0 otherwise.

We will start by examining a special prior distribution, which leads to very familiar results, and then extend the analysis to a slightly more general prior distribution. Let $\tau = \sigma^{-2}$, and suppose that the prior for β, τ can be written as

$$p(\beta, \tau) = p_1(\tau)p_2(\beta|\tau),$$

where

$$p(\beta|\tau) \propto 1$$

and

$$p(\tau) \propto \frac{1}{\tau}.$$

Then the posterior distribution has the following form:

$$p(\beta, \tau|z) = p(\beta|\tau, z)p(\tau|z),$$

where the conditional posterior distribution of β is multivariate normal:

$$\beta|\tau, z \sim \mathcal{N}(b, \sigma^2(X'X)^{-1});$$

and the marginal posterior distribution of the precision τ is scaled chi-square:

$$\tau|z \sim \frac{\chi_{(n-k)}^2}{(n-k)s^2}.$$

That is, the posterior distribution of τ is the distribution of a chi-square $(n-k)$ random variable divided by $(n-k)s^2$. Since the chi-square distribution with ν degrees of freedom has mean ν , the posterior mean of τ is s^{-2} .

2.3 Prior and Posterior Distributions: Version 2

Gamma Distribution: The gamma distribution $\mathcal{G}(x|\alpha, \beta)$ with shape parameter $\alpha > 0$ and inverse scale parameter $\beta > 0$ has density

$$p_g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) I_{(0, \infty)}(x).$$

Notice that the χ_ν^2 distribution is a special case of the gamma distribution with $\alpha = \nu/2$, $\beta = 1/2$.

Suppose that the prior distribution for β, τ has

$$\tau \sim \mathcal{G}(\tau|a_1, a_2),$$

and

$$\beta|\tau \sim \mathcal{N}(\beta_0, \tau^{-1}\Omega),$$

where β_0 is a given $k \times 1$ vector and Ω is a given $k \times k$ positive-definite symmetric matrix. Notice that letting $a_1 \rightarrow 0$ and $a_2 \rightarrow 0$ gives our earlier prior density for τ , and letting Ω become very “large” (note that it is a variance matrix) gives our earlier prior density for β . Then the posterior distribution is given by

$$\tau|z \sim \mathcal{G}(a_1 + \frac{1}{2}n, a_n)$$

$$\beta|\tau, z \sim \mathcal{N}\left(\tilde{\beta}, \sigma^2(X'X + \Omega^{-1})^{-1}\right),$$

where

$$\begin{aligned} \tilde{\beta} &= (\Omega^{-1} + X'X)^{-1}(\Omega^{-1}\beta_0 + X'y), \\ a_n &= a_2 + \frac{1}{2}(n - k)s^2 + \frac{1}{2}(b - \beta_0)' \Omega^{-1} (X'X + \Omega^{-1})^{-1} X'X (b - \beta_0). \end{aligned}$$

3 Likelihood Functions Based on Multivariate Normality

3.1 Likelihood Function

The observation $Z = (Z_1, \dots, Z_n)$, where $Z_i = (Y_i, X_i)$, $Y_i = (Y_{i1}, \dots, Y_{im})$ is $m \times 1$, $X_i = (X_{i1}, \dots, X_{ik})$ is $k \times 1$, and the partial likelihood function has the form

$$f_\theta = \prod_{i=1}^n l(y_i|x_i, \theta).$$

Here,

$$l(y_i|x_i, \theta) = \mathcal{N}_m(\Pi x_i, \Sigma),$$

that is, each term has the form of the density of a k -dimensional multivariate normal random variable with mean vector Πx_i and variance matrix Σ . The matrix Π can be written

$$\Pi = \begin{pmatrix} \pi'_1 \\ \vdots \\ \pi'_m \end{pmatrix},$$

where π_j is a $k \times 1$ vector, so that $\pi'_j x_i$ is the conditional mean of Y_{ij} given $X_i = x_i$.

Example 3 (Multivariate Regression Model)

Suppose that $Y_i = (Y_{i1}, \dots, Y_{im})'$ is $m \times 1$ and $X_i = (X_{i1}, \dots, X_{ik})'$ is $k \times 1$, and P_θ specifies that

$$\begin{aligned} y_{i1} &= \pi_{11}x_{i1} + \dots + \pi_{1k}x_{ik} + \epsilon_{i1} \\ &\vdots \\ y_{im} &= \pi_{m1}x_{i1} + \dots + \pi_{mk}x_{ik} + \epsilon_{im} \end{aligned}$$

and that $\epsilon \equiv (\epsilon_{i1}, \dots, \epsilon_{im})'$ is distributed i.i.d. normal:

$$\epsilon | X_1 = x_1, \dots, X_n = x_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma).$$

Equivalently, we can write

$$Y_i | X_1 = x_1, \dots, X_n = x_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\Pi x_i, \Sigma),$$

where Π is a $m \times k$ matrix of coefficients.

Example 4 (Vector Autoregression)

Suppose that we observe Y_0, \dots, Y_T , where $Y_t = (Y_{t1}, \dots, Y_{tm})'$, and that

$$Y_t | Y_0 = y_0, \dots, Y_{t-1} = y_{t-1}, \Pi, \Sigma \sim \mathcal{N}(\Pi y_{t-1}, \Sigma), \quad t = 1, \dots, T.$$

This fits into our framework with $X_t = (Y_{t-1})'$.

3.2 Prior Distribution:

Wishart Distribution: If v_i is distributed i.i.d. $\mathcal{N}_m(0, B)$ for $i = 1, \dots, \nu$, then

$$V \equiv \sum_{i=1}^{\nu} v_i v_i' \sim \mathcal{W}(\nu, B).$$

The density function of the Wishart distribution is

$$f_{\mathcal{W}(\nu, B)}(V) = c|V|^{(\nu-m-1)/2} \exp \left[-\frac{1}{2} \text{tr}(B^{-1}V) \right].$$

The prior distribution for π and $D = \Sigma^{-1}$ is given by:

$$p(\pi, D) = p_\pi(\pi)p_D(D),$$

where

$$p(\pi) \propto 1,$$

$$p(D) = f_{\mathcal{W}(r, Q)}(D).$$

So the prior for the mean parameters π is “diffuse.”

3.3 Posterior Distribution:

It will be useful to introduce some additional notation. Let

$$\mathbf{X} = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}$$

be the $n \times k$ matrix of regressors, and let $\hat{\pi}_j$ be the least-squares coefficients in a regression of the j th component of the Y_i on \mathbf{X} :

$$\hat{\pi}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_j,$$

where

$$\mathbf{y}_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix}, \quad j = 1, \dots, m.$$

Let

$$\hat{\Pi} = \begin{pmatrix} \hat{\pi}'_1 \\ \vdots \\ \hat{\pi}'_m \end{pmatrix},$$

be the $m \times k$ matrix of least-squares coefficients arranged in the same format as Π , and define the least-squares residuals as

$$e_i = y_i - \hat{\Pi}x_i.$$

So e_i is $m \times 1$. Then form

$$S = \sum_{i=1}^n e_i e_i'.$$

Let

$$\pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix}, \quad \hat{\pi} = \begin{pmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_m \end{pmatrix}.$$

So π and $\hat{\pi}$ are $km \times 1$ vectors. We can write the posterior distribution as:

$$\begin{aligned} p(\pi, D|z) &= p(\pi|D, z)p(D|z), \\ p(\pi|D, z) &= \mathcal{N}(\hat{\pi}, D^{-1} \otimes (\mathbf{X}'\mathbf{X})^{-1}), \\ p(D|z) &= \mathcal{W}(n + r - k, (S + Q^{-1})^{-1}). \end{aligned}$$

A “diffuse” prior for D has $r = 0$, $Q^{-1} = 0$, so that

$$p(D) \propto |D|^{-(m+1)/2}.$$

In this case,

$$p(D|z) = \mathcal{W}(n - k, S^{-1}).$$

4 Monte Carlo Simulation

In general, we want to use the posterior distribution to simulate posterior loss:

$$E(L(\theta, a)|z) = \int L(\theta, a)p(\theta|z)d\theta. \quad (4)$$

This is the expectation of $L(\theta, a)$ with respect to the posterior distribution of θ given z , which has density $p(\theta|z)$. How do we calculate this expectation? One simple way is by *Monte Carlo Integration*. Suppose we can take J independent draws from $p(\theta|z)$:

$$\theta^{(j)} \stackrel{\text{i.i.d.}}{\sim} p(\theta|z), \quad j = 1, \dots, J.$$

Then we can form the sample analog to Equation 4:

$$\frac{1}{J} \sum_{j=1}^J L(\theta^{(j)}, a).$$

By the weak law of large numbers, as $J \rightarrow \infty$,

$$\frac{1}{J} \sum_{j=1}^J L(\theta^{(j)}, a) \xrightarrow{P} E(L(\theta, a)|z).$$

Example 5 Return to the classical regression model as discussed in Example 1. If we use the first prior distribution (which we will sometimes call the “conventional diffuse prior”), then the posterior distribution is given by

$$\tau|z \sim \frac{\chi^2_{(n-k)}}{(n-k)s^2}.$$

$$\beta|\tau, z \sim \mathcal{N}(b, \sigma^2(X'X)^{-1});$$

We can form draws for (τ, β) in the following way: First, calculate s^2 and b . Then, for each $j = 1, \dots, J$,

1. Draw $n - k$ independent standard normal random variables $w_1^{(j)}, \dots, w_{n-k}^{(j)}$, and form

$$w^{(j)} = \sum_{l=1}^{n-k} (w_l^{(j)})^2.$$

2. Form

$$\tau^{(j)} = \frac{w^{(j)}}{(n-k)s^2}.$$

3. Draw

$$\beta^{(j)} \sim \mathcal{N}(b, [\tau^{(j)}(X'X)]^{-1}).$$

To do this, calculate the Cholesky factor A such that

$$A'A = [\tau^{(j)}(X'X)]^{-1}.$$

(In Matlab, use the `chol` command.) Let $v^{(j)}$ be a $k \times 1$ vector of independent standard normal random variables, and form

$$\beta^{(j)} = b + A'v^{(j)}.$$

Most statistics packages provide routines to draw from various distributions such as the standard normal. If a routine for the χ^2 distribution is available, step 1 above can be simplified by simply drawing $w^{(j)}$ from the $\chi^2(n-k)$ distribution. Likewise, if a routine for the multivariate normal distribution is available, step 3 is straightforward.

Each $(\beta^{(j)}, \tau^{(j)})$ pair will be an independent draw from the posterior distribution of β, τ . If the loss function is given by $L(\beta, \tau, a)$, then we can approximate its expectation via

$$\frac{1}{J} \sum_{j=1}^J L(\beta^{(j)}, \tau^{(j)}, a).$$

5 The Role of Conditioning

So far in this lecture note we have assumed that it is appropriate to focus on inference based on the partial likelihood, rather than the full joint likelihood implied by P_θ . In the classical and multivariate regression models, this meant doing the analysis conditional on the regressors X . In the autoregressive models, this meant conditioning on the initial observation Y_0 .

In this section we discuss when this sort of conditional analysis can be justified within the decision-theoretic framework we have set up. Suppose that we have an experiment in which we observe a random variable $Z = (Y, X)$. Here Y and X could each be vector-valued. Suppose that Z is distributed P_θ , with associated likelihood function $f(z|\theta)$. Assume that $\Theta = \Theta_1 \times \Theta_2$, and $\theta = (\theta_1, \theta_2)$, where $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, and that the likelihood can be written as

$$f(z|\theta) = f(y, x|\theta) = f(y|x, \theta_1)f(x|\theta_2).$$

Notice that $f(y|x, \theta_1)$ is a partial likelihood function, and that X is *ancillary* with respect to θ_1 : its sampling distribution does not depend on θ_1 . Suppose that the loss function depends on θ only through θ_1 :

$$L((\theta_1, \theta_2), a) = L(\theta_1, a).$$

Finally, suppose that the prior density for θ has θ_1 and θ_2 independent:

$$p(\theta_1, \theta_2) = p_1(\theta_1)p_2(\theta_2).$$

Consider the posterior expected loss in this problem. It can be written as

$$\begin{aligned} & \int L(\theta, a) f(z|\theta) p(\theta) d\theta \\ &= \int \int L(\theta_1, a) f(y|x, \theta_1) f(x|\theta_2) p_1(\theta_1) p_2(\theta_2) d\theta_1 d\theta_2 \\ &= \int \left[\int L(\theta_1, a) f(y|x, \theta_1) p_1(\theta_1) d\theta_1 \right] f(x|\theta_2) p_2(\theta_2) d\theta_2 \\ &= \int L(\theta_1, a) f(y|x, \theta_1) p_1(\theta_1) d\theta_1 \times \int f(x|\theta_2) p_2(\theta_2) d\theta_2. \end{aligned}$$

Since the second term is constant in a , the Bayes rule can be calculated simply as

$$d_0(z) = \arg \min_{a \in \mathcal{A}} \int L(\theta_1, a) f(y|x, \theta_1) p_1(\theta_1) d\theta_1.$$