STAT 804

Lecture 11

Likelihood Theory

First we review likelihood theory for conditional and full maximum likelihood estimation.

Suppose the data is $X = (Y, Z)$ and write the density of X as

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

Differentiate the identity

$$1 = \int f(y|z,\theta)dy$$

with respect to θ_j (the jth component of θ) and pull the derivative under the integral sign to get

$$\begin{aligned}
0 &= \int \frac{\partial f(y|z,\theta)}{\partial \theta_j} dy \\
&= \int \frac{\partial \log f(y|z,\theta)}{\partial \theta_j} f(y|z,\theta)dy \\
&= \mathrm{E}_\theta(U_{Y|Z;j}(\theta)|Z)
\end{aligned}$$

where $U_{Y|Z;j}(\theta)$ is the jth component of $U_{Y|Z}(\theta)$, the derivative of the log conditional likelihood; $U_{Y|Z}$ is called a conditional score. Since

$$\mathrm{E}_\theta(U_{Y|Z;j}(\theta)|Z) = 0$$

we may take expected values to see that

$$\mathrm{E}_\theta(U_{Y|Z;j}(\theta)) = 0$$

It is also true that the other two scores $U_X(\theta)$ and $U_Z(\theta)$ have mean 0 (when θ is the true value of θ). Differentiate the identity a further time with respect to θ_k to get

$$\begin{aligned}
0 \quad = \quad & \int \frac{\partial^2 \log f(y|z,\theta)}{\partial \theta_j \partial \theta_k} f(y|z,\theta)dy \\
& + \int \frac{\partial \log f(y|z,\theta)}{\partial \theta_j} \frac{\partial \log f(y|z,\theta)}{\partial \theta_k} f(y|z,\theta)dy.
\end{aligned}$$

We define the conditional Fisher information matrix $I_{Y|Z}(\theta)$ to have jkth entry

$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}|Z\right]$$

and get

$$I_{Y|Z}(\theta|Z) = \mathrm{Var}_\theta(U_{Y|Z}(\theta)|Z).$$

The corresponding identities based on f_X and f_Z are

$$I_X(\theta) = \mathrm{Var}_\theta(U_X(\theta))$$

and

$$I_Z(\theta) = \mathrm{Var}_\theta(U_Z(\theta))$$

Now let's look at the model $X_t = \rho X_{t-1} + \epsilon_t$. Putting $Y = (X_1, \ldots, X_{T-1})$ and $Z = X_0$ we find

$$U_{Y|Z}(\rho, \sigma) = \begin{array}{c} \frac{\sum_1^{T-1}(X_t - \rho X_{t-1})X_{t-1}}{\sigma^2} \\ \frac{\sum_1^{T-1}(X_t - \rho X_{t-1})^2}{\sigma^3} - \frac{T-1}{\sigma} \end{array}$$

Differentiating again gives the matrix of second derivatives

$$\left[\begin{array}{cc} -\frac{\sum_1^{T-1} X_{t-1}^2}{\sigma^2} & -2\frac{\sum_1^{T-1}(X_t - \rho X_{t-1})X_{t-1}}{\sigma^3} \\ -2\frac{\sum_1^{T-1}(X_t - \rho X_{t-1})X_{t-1}}{\sigma^3} & -3\frac{\sum_1^{T-1}(X_t - \rho X_{t-1})^2}{\sigma^4} + \frac{T-1}{\sigma^2} \end{array}\right]$$

Taking conditional expectations given X_0 gives

$$I_{Y|Z}(\rho, \sigma) = \left[\begin{array}{cc} \frac{\sum_1^{T-1} \mathrm{E}[X_{t-1}^2|X_0]}{\sigma^2} & 0 \\ 0 & \frac{2(T-1)}{\sigma^2} \end{array}\right]$$

To compute $W_k \equiv \mathrm{E}[X_k^2|X_0]$ write $X_k = \rho X_{k-1} + \epsilon_k$ and get

$$W_k = \rho^2 W_{k-1} + \sigma^2$$

with $W_0 = X_0^2$. You can check carefully that in fact W_k converges to some W_∞ as $k \to \infty$. This W_∞ satisfies $W_\infty = \rho^2 W_\infty + \sigma^2$ which gives

$$W_\infty = \frac{\sigma^2}{1 - \rho^2}$$

It follows that

$$\frac{1}{T} I_{Y|Z}(\rho,\sigma) \to \begin{bmatrix} \frac{1}{1-\rho^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Notice that although the conditional Fisher information might have been expected to depend on X_0 it does not, at least for long series.

Large Sample Theory for Conditional Likelihood:

We have data $X = (Y, Z)$ and study the conditional likelihood, score Fisher information and mle: $\ell_{Y|Z}(\theta)$, $U_{Y|Z}(\theta)$, $\mathcal{I}_{Y|Z}(\theta)$ and $\hat\theta$. In general standard maximum likelihood theory may be expected to apply to these conditional objects:

1. $P(\ell_{Y|Z}(\theta_0) > \ell_{Y|Z}(\theta)) \to 1$ as the "sample size" (often measured by the Fisher information) tends to infinity.

2. $\mathrm{E}_\theta\left[U_{Y|Z}(\theta)|Z\right] = 0$

3. $\hat\theta$ is consistent (converges to the true value as the Fisher information converges to infinity).

4. The usual Bartlett identities hold. For example:

$$\mathcal{I}_{Y|Z}(\theta) \equiv \mathrm{Var}\left[U_{Y|Z}(\theta)|Z\right] = -\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta} U_{Y|Z}(\theta)|Z\right]$$

5. The error in the mle has approximately the form

$$\hat\theta - \theta \approx \left(\mathcal{I}_{Y|Z}(\theta)\right)^{-1} U_{Y|Z}(\theta)$$

6. The mle is approximately normal:

$$\left(\mathcal{I}_{Y|Z}(\theta)\right)^{1/2}\left(\hat\theta - \theta\right) \approx MVN(0, I)$$

(where I is the identity matrix).

7. The conditional Fisher information can be estimated by the observed information:

$$\left(\mathcal{I}_{Y|Z}(\theta)\right)^{-1}\left(-\frac{\partial}{\partial\theta} U_{Y|Z}(\hat\theta)\right) \to I$$

8. The log-likelihood ratio is approximately χ^2:

$$2(\ell_{Y|Z}(\hat\theta) - \ell_{Y|Z}(\theta_0)) \Rightarrow \chi^2_p$$

In the previous lecture I showed you 2) and 4) in this list. Today we look at 5), 6) and 7) in the context of the $AR(1)$ model $X_t = \rho X_{t-1} + \epsilon_t$.
Non Gaussian series.

The fitting methods we have studied are based on the likelihood for a normal fit. However, the estimates work reasonably well even if the errors are not normal.

Example: AR(1) fit. We fit $X_t - \mu = \rho(X_{t-1} - \mu) + \epsilon_t$ using $\hat\mu = \bar X$ which is consistent for non-Gaussian errors. (In fact

$$(1-\rho)\sum_0^{T-1} X_t + \rho X_{T-1} - X_0 = (T-1)(1-\rho)\mu + \sum_0^{T-1} \epsilon_t - \epsilon_0;$$

divide by T and apply the law of large numbers to $\bar\epsilon$ to see that $\bar X$ is consistent.)

Here is an outline of the logic of what follows. We will assume that the errors are iid mean 0, variance σ^2 and finite fourth moment $\mu_4 = \mathrm{E}(\epsilon_t^4)$. We will **not** assume that the errors have a normal distribution.

1. The estimates of ρ and σ are consistent.

2. The score function satisfies

$$T^{-1/2} U(\theta_0) \Rightarrow MVN(0, B)$$

where

$$B = \begin{bmatrix} \frac{1}{1-\rho^2} & 0 \\ 0 & \frac{\mu_4 - \sigma^4}{\sigma^6} \end{bmatrix}$$

3. The matrix of second derivatives satisfies

$$\lim_{T\to\infty} -\frac{1}{T}\frac{\partial U}{\partial \theta} = \lim_{T\to\infty} -\frac{1}{T}\mathrm{E}\left(\frac{\partial U}{\partial \theta}\right) = A$$

where

$$A = \begin{bmatrix} \frac{1}{1-\rho^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

4. If $\mathcal{I}$ is the (conditional) Fisher information then

$$\lim \lim_{T\to\infty} \frac{1}{T}\mathcal{I} = A$$

5. We can expand $U(\hat\theta)$ about θ_0 and get

$$T^{1/2}(\hat\theta - \theta) = \left[\frac{1}{T} I(\theta_0)\right]^{-1} \left[T^{-1/2}U(\theta_0)\right] + \text{negligible remainder}$$

6. So

$$T^{1/2}(\hat\theta - \theta) \approx MVN(0, A^{-1}BA^{-1}) = MVN(0,\Sigma)$$

   where

$$\Sigma = A^{-1}BA^{-1} = \left[\begin{array}{cc} 1-\rho^2 & 0 \\ 0 & \frac{\mu_4 - \sigma^4}{4\sigma^2} \end{array}\right]$$

7. So $T^{1/2}(\hat\rho - \rho) \Rightarrow N(0, 1-\rho^2)$ even for non-normal errors.

8. On the other hand the estimate of σ has a limiting distribution which will be different for non-normal errors (because it depends on μ_4 which is $3\sigma^4$ for normal errors and something else in general for non-normal errors).

Here are details.

Consistency: One of our many nearly equivalent estimates of ρ is

$$\hat\rho = \frac{\sum(X_t - \bar X)(X_{t-1} - \bar X)}{\sum(X_t - \bar X)^2}$$

Divide both top and bottom by T. You need essentially to prove

$$T^{-1}\sum(X_t - \mu)(X_{t-1} - \mu) \to C(1)$$

and

$$T^{-1}\sum(X_t - \mu)^2 \to C(0)$$

Each of these is correct and hinges on the fact that these linear processes are ergodic — long time averages converge to expected values. For these particular averages it is possible to compute means and variances and prove that the mean squared error converges to 0.

Score function: asymptotic normality

The score function is

$$U(\rho,\sigma) = \left[\begin{array}{c} \frac{\sum X_{t-1}(X_t-\rho X_{t-1})}{\sigma^2} \\ \frac{\sum (X_t-\rho X_{t-1})^2}{\sigma^3} - \frac{T-1}{\sigma} \end{array}\right]$$

If ρ and σ are the true values of the parameters then

$$U(\rho,\sigma) = \left[\begin{array}{c} \frac{\sum X_{t-1}\epsilon_t}{\sigma^2} \\ \frac{\sum \epsilon_t^2}{\sigma^3} - \frac{T-1}{\sigma} \end{array}\right]$$

I claim that $T^{-1/2}U(\rho,\sigma) \Rightarrow MVN(0,B)$. This is proved by the martingale central limit theorem. Technically you fix an $a \in R^2$ and study $T^{-1/2}a^tU(\rho,\sigma)$, proving that the limit is $N(0,a^tBa)$. I do here only the special cases $a=(1,0)^t$ and $a=(0,1)^t$. The second of these is simply

$$T^{-1/2}\sum(\epsilon_i^2-\sigma^2)/\sigma^3$$

which converges by the usual CLT to $N(0,(\mu_4-\sigma^4)/\sigma^6)$. For $a=(1,0)^t$ the claim is that

$$T^{-1/2}\sum X_{t-1}\epsilon_t \Rightarrow N(0,C(0)\sigma^2)$$

because $C(0)=\sigma^2/(1-\rho^2)$.

To prove this assertion we define for each T a martingale $M_{T,k}$ for $k = 1,\ldots,T$ where

$$M_{T,k} = \sum_1^k D_{T,i}$$

with

$$D_{T,i} = T^{-1/2}X_{i-1}\epsilon_i$$

The martingale property is that

$$\mathrm{E}(M_{T,k+1}|\epsilon_k,\epsilon_{k-1},\ldots) = M_{T,k}$$

The martingale central limit theorem (Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application.* New York: Academic Press.) states that

$$M_{T,T} \Rightarrow N(0,b)$$

provided that

$$\sum_k D_{T,k}^2 \to b$$

and provided that an analogue of Lindeberg's condition holds. Here I check only the former condition:

$$\sum_k D_{T,k}^2 = \frac{1}{T}\sum X_{t-1}^2\epsilon_t^2 \to \mathrm{E}(X_0^2\epsilon_1^2) = C(0)\sigma^2$$

(by the ergodic theorem or you could compute means and variances).
Second derivative matrix and Fisher information: the matrix of negative second derivatives is

$$-\frac{\partial U}{\partial \theta} = \left[\begin{array}{cc} \frac{\sum X_{t-1}^2}{\sigma^2} & 2\frac{\sum X_{t-1}(X_t-\rho X_{t-1})}{\sigma^3} \\ 2\frac{\sum X_{t-1}(X_t-\rho X_{t-1})}{\sigma^3} & 3\frac{\sum (X_t-\rho X_{t-1})^2}{\sigma^4} - \frac{T-1}{\sigma^2} \end{array}\right]$$

If you evaluate at the true parameter value and divide by T the matrix and the expected value of the matrix converge to

$$A = \left[\begin{array}{cc} \frac{C(0)}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{array}\right]$$

(Again this uses the ergodic theorem or a variance calculation.)
Taylor expansion: In the next step we are supposed to prove that a random vector has a MVN limit. The usual tactic to prove this uses the so called Cramér-Wold device — you prove that each linear combination of the entries in the vector has a univariate normal limit. Then $U(\hat\rho, \hat\sigma) = 0$ and Taylor's theorem is that

$$0 = U(\hat\rho, \hat\sigma) = U(\rho, \sigma) + \left[\frac{\partial U(\theta)}{\partial \theta}\right](\hat\theta - \theta) + R$$

(Here we are using $\theta^t = (\rho, \sigma)$ and R is a remainder term — a random variable with the property that

$$P(||R||/||U(\theta)|| > \eta) \to 0$$

for each $\eta > 0$.) Multiply through by

$$\left[\frac{\partial U(\theta)}{\partial \theta}\right]^{-1}$$

and get

$$T^{1/2}(\hat\theta - \theta) = \left[-T^{-1}\frac{\partial U(\theta)}{\partial\theta}\right]^{-1}(T^{-1/2}U(\rho,\sigma) + T^{-1/2}R)$$

It is possible with care to prove that

$$\left[-T^{-1}\frac{\partial U(\theta)}{\partial\theta}\right]^{-1}(T^{-1/2}R) \to 0$$

Asymptotic normality: This is a consequence of Slutsky's theorem applied to the Taylor expansion and the results above for U and I. According to Slutsky's theorem the asymptotic distribution of $T^{1/2}(\hat\theta - \theta)$ is the same as that of

$$A^{-1}(T^{-1/2}U(\rho,\sigma))$$

which converges in distribution to $MVN(0, A^{-1}B(A^{-1})^t)$. Now since $C(0) = \sigma^2/(1-\rho^2)$

$$A^{-1}B(A^{-1})^t = \begin{bmatrix} 1-\rho^2 & 0 \\ 0 & \frac{\mu_4-\sigma^4}{4\sigma^4} \end{bmatrix}$$

Behaviour of $\hat\rho$: pick off the first component and find

$$T^{1/2}(\hat\rho - \rho) \Rightarrow N(0, 1-\rho^2)$$

Notice that this answer is the same for normal and non-normal errors.
Behaviour of $\hat\sigma$: on the other hand

$$T^{1/2}(\hat\sigma - \sigma) \Rightarrow N(0, (\mu_4 - \sigma^4)/(4\sigma^2))$$

which has μ_4 in it and will match the normal theory limit if and only if $\mu_4 = 3\sigma^4$.
More general models: For an ARMA(p,q) model the parameter vector is

$$\theta = (a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma)^t \,.$$

In general the matrices B and A are of the form

$$B = \begin{bmatrix} B_1 & 0 \\ 0 & \frac{\mu_4-\sigma^4}{\sigma^6} \end{bmatrix}$$

and

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

where $A_1 = B_1$ and A_1 is a function of the parameters $a_1, \ldots, a_p, b_1, \ldots, b_q$ only and is the same for both normal and non-normal data.