1. Conditional Mean

May 2, 2010

Junhui Qian

1 Introduction

We discuss conditional mean models in this part of the course. We first introduce linear models (AR, MA, and ARMA), we explore their properties, and we discuss issues of estimation and forecasting. At the end of this part, we also briefly discuss nonlinear models.

2 Linear Models

Consider a time series $X = (X_t)$ and a filtration $(\mathcal{F}_t)$. In this section, we assume that $\mathbb{E}_{t-1}X_t \equiv \mathbb{E}(X_t|\mathcal{F}_{t-1})$ is a linear function of the lags of X and innovations ε. If $\mathbb{E}_{t-1}X_t = c + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}$, $X \sim \text{AR}(p)$. If $\mathbb{E}_{t-1}X_t = \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q}$, $X \sim \text{MA}(q)$. And if $\mathbb{E}_{t-1}X_t = c + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q}$, $X \sim \text{ARMA}(p,q)$.

2.1 Autoregressive Processes

The simplest autoregressive model is the AR(1), the first-order autoregressive model. We say $X = (X_t)$ is a zero-mean AR(1) process if,

$$X_t = \alpha X_{t-1} + \varepsilon_t, \tag{1}$$

where $|\alpha| < 1$ and (ε_t) is a w.n.

Suppose there is a positive deviation from the mean (ie, $X_{t-1} > 0$) at time $t-1$, then X_t would be a fraction of the deviation (ie, αX_{t-1}) plus a shock term (ε_t) representing new information flowing in. As ε_t has a mean of zero, in average X_t would be closer to zero. It

is as if some force is pulling $(X_t - \mu)$ back to zero. This property of AR(1) model is called "mean-reversion".

More generally, $X \sim$AR(p) if

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t, \tag{2}$$

where the coefficients $(\alpha_1, ..., \alpha_p)$ must satisfy that the roots of $1 - a_1 z - \cdots - a_p z^p = 0$ are all outside the unit circle. If $p = 1$, this means that $|a_1| < 1$.

The ACF of AR process decays exponentially (short memory). And the PACF of AR process is truncated.

2.2 Moving Average Processes

(X_t) is called a q-th order moving average process if X_t can be represented as

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} = \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i},$$

where (β_i) are constants. We denote a q-th order moving average process as MA(q).

If we assume $\mathbb{E}\varepsilon_t^2 = \sigma^2$, then the variance of X_t would be

$$\gamma_X(0) = \sigma^2 \left(1 + \sum_{k=0}^{q} \beta_k^2\right).$$

For $q = \infty$, we have an MA(∞) process. In this case we require the coefficients (β_k) to be "square summable", ie, $\sum_k \beta_k^2 < \infty$. To have a well defined long-run variance (hence mean ergodicity), we further require "absolute summability", ie, $\sum_k |\beta_k| < \infty$. Absolute summability implies square summability, but not vice versa.

To show that the MA(∞) representation with square-summable coefficients is well-defined, we need to prove that $\sum_{k=0}^{T\beta_k} \varepsilon_{t-k}$ converges in L^2 to some random variable X_t as $T \to \infty$. It suffices to show that for any $\epsilon > 0$, there exists a large N such that for any

integer $M > N$

$$\mathbb{E}\left(\sum_{k=0}^{M}\beta_k\varepsilon_{t-k} - \sum_{k=0}^{N}\beta_k\varepsilon_{t-k}\right)^2 < \epsilon.$$

This is the Cauchy criterion for L^2 convergence. We have

$$\begin{aligned}\mathbb{E}\left(\sum_{k=0}^{M}\beta_k\varepsilon_{t-k} - \sum_{k=0}^{N}\beta_k\varepsilon_{t-k}\right)^2 &= \mathbb{E}\left(\sum_{k=N+1}^{M}\beta_k\varepsilon_{t-k}\right)^2 \\ &= \sigma^2\left(\sum_{k=0}^{M}\beta_k^2 - \sum_{k=0}^{N}\beta_k^2\right).\end{aligned}$$

The conclusion then follows from the square summmability condition $\sum_k \beta_k^2 < \infty$.

To that absolute summability implies mean ergodicity, we note that

$$\begin{aligned}\sum_{j=0}^{\infty}|\gamma_j| &= \sigma^2\sum_{j=0}^{\infty}|\sum_{k=0}^{\infty}\beta_{j+k}\beta_k| \\ &\leq \sigma^2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}|\beta_{j+k}\beta_k| \\ &= \sigma^2\sum_{k=0}^{\infty}|\beta_k|\sum_{j=0}^{\infty}|\beta_{j+k}|.\end{aligned}$$

2.2.1 Invertibility

For an MA(q) process, if the roots of $\beta(z) = 1 + \beta_1 z + \cdots + \beta_q z^q = 0$ are all outside the unit circle, then the MA(q) process is "invertible" and is equivalent to an AR(∞) process.

For example, consider a zero-mean MA(1) process, $X_t = \varepsilon_t + \beta\varepsilon_{t-1}$. We my write the model as

$$X_t = (1 + \beta L)\varepsilon_t.$$

If $|\beta| < 1$, we may multiply both sides by $(1 + \beta L)^{-1}$ and obtain

$$(1 - \beta L + \beta^2 L^2 - \cdots)X_t = \varepsilon_t,$$

which is in AR(∞) form.

We may also represent an AR process in MA(∞) form. For example, consider an AR(1) model, $(1-\alpha L)X_t = \varepsilon_t$. We may multiply both sides by $(1-\alpha L)^{-1}$ and obtain

$$X_t = (1 + \alpha L + \alpha L^2 + \cdots)\varepsilon_t,$$

which is an invertible MA(∞).

The ACF of MA process is truncated. And the PACF of AR process declines exponentially.

2.3 ARMA Processes

Combine the AR(p) and the MA(p) models, we obtain the celebrated ARMA(p,q) model. Without loss of generality, we consider a zero-mean process $X = (X_t)$. We say that $X \sim$ ARMA(p,q) if X satisfies,

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}. \tag{3}$$

Using the lag operator L, the above may be written as

$$\alpha(L)X_t = \beta(L)\varepsilon_t,$$

where $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$ and $\beta(z) = 1 + \beta_1 z + \cdots + \beta_q z^q$. For an ARMA($p,q$) process to be stationary, we must have that all roots of $\alpha(z) = 0$ are outside the unit circle. In this case, the ARMA process may be written in MA form,

$$X_t = \frac{\beta(L)}{\alpha(L)}\varepsilon_t = \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i},$$

where the coefficients (ϕ_i) may be obtained using "polynomial long division" of $\beta(z)/\alpha(z)$. For obvious reasons, ϕ_i is called the impulse response function.

2.4 The Autocovariance-Generating Function

We define autocovariance-generating function for a covariance-stationary process X as

$$g_X(z) = \sum_{k=-\infty}^{\infty} \gamma_X(k) z^k.$$

When $z = e^{-i\omega}$, where ω is the radian angle, then $s(\omega) = g_X(e^{-i\omega})$ gives the spectrum of X.

For MA(1) process,

$$g_X(z) = \sigma^2(1 + \beta z)(1 + \beta z^{-1}).$$

For AR(1) process,

$$g_X(z) = \frac{\sigma^2}{(1 - \phi z)(1 - \phi z^{-1})}.$$

2.5 Linear Processes

ARMA processes belong to the class of linear process, which is of the following form,

$$X_t = \sum_{k=-\infty}^{\infty} \varphi_k \varepsilon_{t-k}.$$

The study of linear processes is justified by the celebrated Wold's Decomposition Theorem, which states that any zero-mean weak stationary process (X_t) can be represented as

$$X_t = \sum_{k=0}^{\infty} \varphi_k \varepsilon_{t-k} + d_t,$$

where (ε_t) is white noise, $\phi_0 = 0$, $\sum_{k=0}^{\infty} \varphi_k^2 < \infty$, and d_t is perfectly predictable by $(X_{t-1}, X_{t-2}, ...)$. The Wold's theorem is reassuring to practitioners of ARMA models in that although the data generating process may not be ARMA, we may fit the data with an ARMA model nonetheless, which would adequately describe the correlations in data, albeit not optimally.

The sequence (φ_k) is called a filter. And a linear process can be further transformed by a linear filter. The following results give conditions under which the transformation is meaningful.

Results: Let $\{\varphi_k\}$ be sequence of numbers that are absolutely summable, ie, $\sum_{k=-\infty}^{\infty} |\varphi_k| < \infty$, and let $\{\xi_t\}$ be an arbitrary time series, then we have:

(a) If $\sup_t \mathbb{E}|\xi_t| < \infty$, then $\sum_k \varphi_k \xi_{t-k}$ converges almost surely and in mean (L^1).

(b) If $\sup_t \mathbb{E}|\xi_t|^2 < \infty$, then $\sum_k \varphi_k \xi_{t-k}$ converges in L^2.

(c) If (ξ_t) is stationary, then so is $X_t = \sum_k \varphi_k \xi_{t-k}$.

3 Estimation

We may use OLS or Method of Moments (Yule-Walker) to estimate the AR coefficients in AR and ARMA models. However, MLE is most often used, especially for models with MA components.

3.1 Gaussian Error

We assume that $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. The joint likelihood function of $X \sim$ ARMA(p, q) is given by

$$p(\theta|X_1, X_2, ..., X_T) = p(X_1, X_2, ..., X_T) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2) \cdots p(X_T|X_1, ..., X_{T-1}), \tag{4}$$

where θ is the parameter vector $\theta = (c, (\alpha_i, i = 1, ..., p), (\beta_i, i = 1, ..., q), \sigma^2)'$.

The conditional distribution $X_t|(X_1, ..., X_{t-1})$ is normal with mean $c + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$ and variance σ^2.

ε_t is not observable, but can be calculated given parameters and starting values of X and ε. We have $\varepsilon_t = \alpha(L)X_t - c - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q}$. We may choose the initial values

for X, ie, $(X_0, X_{-1}, ..., X_{1-p})$, to be the sample mean of X; and we may simply set the initial values of ε to be zero. As $T \to \infty$, the choice of initial values plays an insignificant role. For the same reason, we may delete $p(X_1)$ from (4). It is in this sense that we should call the above estimation procedure as conditional MLE. Unconditional MLE for ARMA models is more involved and will be covered later in the course.

3.2 Non-Gaussian Error

If the assumption that $\varepsilon_t \sim N(0, \sigma^2)$ does not hold, we may still pretend that it does and use the above strategy to estimate the parameters. An estimator that maximizes a misspecified likelihood function is called a "quasi-maximum likelihood estimator". QMLE often gives consistent estimates of the model parameters. However, if the residuals are not Gaussian, standard errors that are calculated under the Gaussian assumption are incorrect in general. For more details on QMLE, we refer to Heyde (1997).

For a positive process, we may use the following transformation to produce a Gaussian-like process,

$$X_t^\lambda = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} \text{ for } \lambda \neq 0, \\ \log(X_t) \text{ for } \lambda = 0. \end{cases}$$

This transformation is due to Box and Cox (1964). The parameter λ may be estimated along with other parameters associated with the process.

4 Diagnostics

After estimating a model, it is necessary to check whether the model is valid, in the sense that whether the model adequately captures the correlations in data. For linear models we have introduced so far, the model diagnostics boils down to checking the white noise assumption of the error ε_t. In the minimum, the estimated residuals should display no serial

correlation. Formally, we test the following hypothesis,

$$\mathrm{H}_0: \ \rho_k = 0 \ \forall k, \quad \mathrm{H}_1: \ \text{otherwise.}$$

In practice, we may employ portmanteau tests, the most famous of which is the Ljung-Box test. The test statistic is

$$Q = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k},$$

where m is commonly chosen to be some integer around $\log(T)$. The statistic is asymptotically distributed as χ_m^2.

If Gaussian error is assumed, we also need to check it.

5 Prediction

5.1 Principle

Given past information, we usually rely on estimating conditional expectation to give a prediction. It turns out that conditional expectation is a predictor that gives the least mean squared error.

Consider the problem of predicting Y_{t+1} given X_t, where X_t may be a vector that contains past observations of Y. The conditional expectation $\mathbb{E}(Y_{t+1}|X_t)$ solves the following minimization problem,

$$\min_{\hat{Y}_{t+1|t}} \left(Y_{t+1} - \hat{Y}_{t+1|t}\right)^2.$$

The quantity to be minimized can be called a quadratic loss function.

To see this, we consider an arbitrary predictor $\hat{Y}_{t+1|t} = g(X_t)$, where g is an arbitrary nonlinear function. It is straightforward to show that the MSE for this predictor satisfies

$$\mathbb{E}[Y_{t+1} - g(X_t)]^2 = \mathbb{E}[Y_{t+1} - \mathbb{E}(Y_{t+1}|X_t)]^2 + \mathbb{E}[\mathbb{E}(Y_{t+1}|X_t) - g(X_t)]^2.$$

It is now clear that, to have a minimal MSE, $g(X_t) = \mathbb{E}(Y_{t+1}|X_t)$.

5.2 Practice

To illustrate how prediction is done in practice, we discuss a simple case. Suppose the data generating process is ARMA(1,1),

$$X_t = aX_{t-1} + \varepsilon_t + b\varepsilon_{t-1}, \quad t = 1, \ldots, T,$$

where $\varepsilon_t \sim$ i.i.d. with $\text{var}(\varepsilon_t) = \sigma^2$. Let $\hat{a}$ and $\hat{b}$ be consistent parameter estimates, and let $\hat{\varepsilon}_t$ be estimated residuals. The one-step-ahead forecast would estimate the following conditional mean,

$$\mathbb{E}(X_{T+1}|X_T, \ldots, X_1) = aX_T + b\varepsilon_T.$$

Obviously, the desired forecast is

$$\hat{X}_{T+1} = \hat{a}X_T + \hat{b}\hat{\varepsilon}_T.$$

Assume that T is large, the forecast error may be approximated by ε_{T+1}. Hence the variance of forecast error is simply σ^2. If we further assume Gaussian error, we may construct interval forecast. The two-step-ahead forecast would estimate

$$\mathbb{E}(X_{T+2}|X_T, \ldots, X_1) = a\mathbb{E}(X_{T+1}|X_T, \ldots, X_1).$$

The desired forecast is thus

$$\hat{X}_{T+2} = \hat{a}\hat{X}_{T+1}.$$

Assume T is large, the forecast error may be approximated by $\varepsilon_{T+2} + (a+b)\varepsilon_{T+1}$. Hence the variance of forecast error is simply $\left(1 + (a+b)^2\right)\sigma^2$.

6 Nonlinear Models

6.1 Threshold AR

First consider a simple TAR(1) model:

$$X_t = \begin{cases} a_1 X_{t-1} + \varepsilon_t, & \text{if } X_{t-1} < 0 \\ a_2 X_{t-1} + \varepsilon_t, & \text{if } X_{t-1} \geq 0, \end{cases}$$

where $a_1 < 1$, $a_2 < 1$, and $a_1 a_2 < 1$. TAR can be viewed as a regime switching model. Here in this example we have two regimes, and the entry and exit of regimes are determined by an observable signal which is X_{t-1}.

This model captures asymmetries in the dynamics of time series. For example, if $a_1 = -1.5$ and $a_2 = 0.5$, then there are one regime that tend to reverse forcefully and another one that tend to stay.

The sufficient and necessary condition for X_t to be ergodic is that

$$a_1 < 1, \quad a_2 < 1, \quad \text{and} \quad a_1 a_2 < 1.$$

The TAR(1) model can be extended in several directions. We may increase the number of regimes and lags, and we may choose other trigger signal than X_{t-1}.

6.2 Smooth-Transition AR

The entry and the exit of regimes in TAR models are sudden. Sometimes it is more reasonable to assume gradual transitions into other regimes. Consider the following smooth-transition AR (STAR) model,

$$X_t = c_1 + a_1 X_{t-1} + F\left(\frac{X_{t-1} - \ell}{s}\right)(c_2 + a_2 X_{t-1}) + \varepsilon_t,$$

where F is a cumulative distribution function. The typical choices of F include logistic function ($F(x) = 1/(1 + \exp(-x))$) and cdf of $N(0, 1)$.

It is clear that as $X_{t-1} \to -\infty$, we have

$$X_t = c_1 + a_2 X_{t-1} + \varepsilon_t.$$

And as $X_{t-1} \to \infty$, we have

$$X_t = (c_1 + c_2) + (a_1 + a_2) X_{t-1} + \varepsilon_t.$$

In both TAR and STAR, regime switching is based on observable signal. At any time point, we are sure about the current and past regimes. In markov switching AR (MSAR) models, only an unobservable probability law of the regimes is assumed. We will discuss MSAR later in the course.

6.3 Estimation and Forecasting

MLE is the usual choice for estimating nonlinear models. Given a model like TAR or STAR, it is not difficult to write down its likelihood function. However, it is much more difficult to make forecasts based on an estimated nonlinear model. It is difficult to obtain explicit forms of $\mathbb{E}(X_{T+\ell}|X_T, \ldots, X_1)$, which is essential for the usual ARMA-based forecasting. It is thus common to use bootstrap to make forecasts. We may do the following:

(1) Draw with replacement $\varepsilon_{T+1}, \ldots, \varepsilon_{T+\ell}$ from $(\hat{\varepsilon}_t)$.

(2) Compute $\hat{X}^{(i)}_{T+1}, \ldots, \hat{X}^{(i)}_{T+\ell}$ recursively.

(3) Repeat (1) and (2) and obtain B realizations of $\hat{X}^{(i)}_{T+\ell}$.

(4) The point forecast would be $B^{-1} \sum_{i=1}^{B} \hat{X}^{(i)}_{T+\ell}$. Forecasts of intervals and distribution can also be made.

References

Heyde,C.C., Quasi-Likelihood and Its Application. A General Approach to Optimal Parameter Estimation Springer, 1997