8: Auto-Regressive and Moving Average Models

GECO 6281 Advanced Econometrics 1 (Lab)

Patrick Mokre

Fall 2020

Recapitulation

- Discrete Choice Modeling: LOGIT, PROBIT, TOBIT, Heckman
- Dealing with Endogeneity: Instrumental Variables (IV)
- Static Panel Data
- Instrumental Variables in Dynamic Panel Data: Anderson-Hsiao, Arellano-Bond, Arellano-Bover
- Seemingly Unrelated Regression (SUR)
- Auto-Regressive Distributed Lag (ARDL) and Error Correction Modeling (ECM)

Time Series Modeling

Auto-Regressive and Moving Average processes are the basis of dynamic econometric relationships.

They are sometimes, but not always, exchangeable concepts.

Naturally they imply a closer look on the independence assumption of the error terms over time.

The most basic macroeconometric (i.e. time series) processes are *univariate* and *linear*.

$$X_t = f(X_{t-1,X_{t-2},...,\epsilon_t}) = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} ... + \epsilon_t \tag{1}$$

Univariate modeling is often inappropriate in economic application, but a necessary (or at least very useful) building block for understanding time-series econometrics.

White Noise

A process with zero mean, which is homoskedastic and non-autoregressive is called **white noise**.

$$
\begin{aligned}
E[\epsilon_t] &= 0 \\
E[\epsilon_t' \epsilon_t] &= \sigma_\epsilon^2 < \infty \quad \forall t \in T \\
E[\epsilon_s \epsilon_t] &= 0 \quad \forall s \neq t
\end{aligned}
$$

White noise processes can nevertheless change over time in distribution or show non-linear and higher moment dependence.

Martingale Processes

A Martingale is a stochastic sequence where the conditional expectation of the next observation is the current expectation (Fuller, 1996, "*Introduction to statistical time series*", 234f).

$$E(\mid X_n \mid) < \infty$$
$$E(X_{n+1}\ mid X_1, ..., X_n) = X_n$$

A Martingale can also be defined **with respect to another sequence**, ie. Y is a Martingale to X.

$$E(\mid Y_n \mid) < \infty$$
$$E(Y_{n+1} \mid X_a, ..., X_n) = X_n$$

Martingales and White Noise

A process is **white noise** if we cannot predict the next observation.

$$E[Z_t] = 0$$
$$E[Z_t, Z_{t+j}] = 0 \forall j \neq 0$$

Not every white noise process is a Martingale: If you flip a fair coin and it lands on "head" you would not expect the next flip to land on "head" again.

Not every Martingale is white noise: Keep in mind the expected value of a white noise process is always zero.

However, a Martingale is useful to describe the **differences** between observations of a white noise process $\Delta Y_n = Y_n - Y_{n-1}$ when $E[\Delta_t Y \mid Y_1, ..., Y_{n-q}] = 0$. Then we can apply the **central limits theorem for Martingales** for deriving appropriate estimators.

Simple AR(1) Process

$$X_t = \delta + \phi_1 X_{t-1} + \epsilon_t$$

With ϵ_t a **white noise** process and $|\phi_1| < 1$ is a **first-order autoregressive process**.

It can either be intepreted as a stochastic rule for $t \in T$ or as a law of motion for $t \in \mathbb{Z}$. For the latter interpretation one assumes the process has started in the infinite past, which is more convenient and also never true for economic processes.

Simple AR(1) Process: Properties

The mean of the infinite-past AR(1) process $\mu = E[X_t]$ is not dependent on t, and ϵ_t is white noise, i.e. $E[\epsilon_t] = 0$.

$$\begin{aligned} E[X_t] &= E[\delta + \phi_1 X_{t-1} + \epsilon_t] \\ \mu &= \delta + \phi_1 \mu \\ \mu &= \frac{\delta}{1 - \phi_1} \end{aligned} \tag{2}$$

Let $x_t = X_t - \mu$ (**mean difference**) and assume $V[x_t] = V[x_{t-1}]$ (**homoskedasticity**).

$$\begin{aligned} V[x_t] &= V[\phi_1 x_{t-1}] + V[\epsilon_t] \\ &= \phi_1^2 V[x_t] + \sigma_\epsilon^2 \\ V[x_t] &= \frac{\sigma_\epsilon^2}{1 - \phi_1^2} \end{aligned} \tag{3}$$

Simple MA(1) Process

$$X_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \tag{4}$$

Let ϵ_t be a **white noise** process and $\theta \in [-1, 1]$ (usually).

The statistical properties are simple:

$$E[X_t] = \mu$$
$$V[X_t] = (1 + \theta^2)\sigma_\epsilon^2$$

AR(1) as an infinite-order MA process

Apply continued substitution to the mean difference formulation of the AR(1) process $x_t = X - t - \mu$.

$$
\begin{aligned}
x_t &= \phi_1 x_{t-1} + \epsilon_t \\
&= \phi_1^k x_{t-k} + \sum_{j=0}^{k-1} \phi_1^j \epsilon_{t-j} \\
&= \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j} \qquad (5)
\end{aligned}
$$

Reminder: Stationarity

A process is called **stationary** if and only if its statistical properties are the same over time. In other words, the joint distribution of $(X_t, X_{t-1}, ..., X_{t-s})$ and $(X_{t-j}, X_{t-j-1}, ..., X_{t-j-s})$ are the same $\forall t, j, s$.

For many econometrical applications it is sufficient to show **weak stationarity** (covariance stationarity), i.e. the first two moments of the joint distribution are unaffected by time shifts.

This is also simpler to reject or not reject than equality in all statistical properties.

$$
\begin{aligned}
E[X_t] &= \mu < \infty \quad \forall t \in T \\
V[X_t] &= \sigma_X^2 = \gamma(0) = \gamma_0 < \infty \quad \forall t \in T \\
Cov[X_t, X_{t-k}] &= \gamma(k) = \gamma_k < \infty \quad \forall t, s \in T
\end{aligned}
$$

Auto-Covariance and Auto-Correlation

If a process is covariance stationary, one can define the auto-covariance function of order k $\gamma(k) = Cov(X_t, X_{t-k}) = \gamma_k$, where γ is a function $\gamma : \mathbb{Z} \to \mathbb{R}$.

Correspondingly, one can define the auto-correlation function (ACF) $\rho : \mathbb{Z} \to [-1, 1]$. $\rho(k) = \frac{Cov(X_t, X_{t-k})}{V[Y_t]} = \frac{\gamma_k}{\gamma_0}$.

A stationary AR(1) process is described by its expected value, its variance and its auto-correlation function, which capture the most important aspects of its dynamics over time.

$$
\begin{aligned}
X_t &= \delta + \phi X_{t-1} + \epsilon_t \\
\rho_k &= \phi^k
\end{aligned} \tag{6}
$$

For the MA(1) process, the ACF plays a similarily important role.

$$
\begin{aligned}
X_t &= \mu + \epsilon_t + \theta \epsilon_{t-1} \\
\rho_1 &= \frac{\theta}{1 + \theta^2} && (7) \\
\rho_k &= 0 \quad \forall k > 1 && (8)
\end{aligned}
$$

Stationarity of AR

- White Noise is stationary ($\equiv \phi = 0$)
- AR(1) with $|\phi| \geq 1$ are not stationary
- AR(1) with $|\phi| < 1$ are not generally stationary for arbitrary starting values, but become more stationary over time (**asymptotical stationarity**)
- AR(1) with $|\phi| < 1$ are stationary if the process started in the infinite past, or if the first distribution is drawn from a specific distribution.

Lag Operator

The lag operator simplifies the relationship between two observations in a time series. Most of the time, it can be treated like a constant. It leaves constants unaffected.

$$\begin{aligned} Lx_t &= x_{t-1} \\ L^2 x_t &= L(Lx_t) = x_{t-2} \\ L^p y_t &= y_{t-p} \end{aligned} \tag{9}$$

ARMA models can be written in terms of the lag operator.

$$\begin{aligned} x_t &= \phi y_{t-1} + \epsilon_t \\ (1 - \phi L)x_t &= \epsilon_t \end{aligned} \tag{10}$$

More general $\phi(L)$ is a lag polynomial for an AR process with $\phi(L) = 1 - \phi_1 L - ... - \phi_p L^p$. The polynomial for an MA process is written as $\theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$.

The roots of the characteristical lag polynomial is the subject of most **unit root tests**.

Invertibility

Note how the AR polynomial has minus signs, while the MA polynomial has plus signs. If a $\theta^{-1}(L)$ exists, one can rewrite the MA polynomial as an infinite-order AR process.

$$\begin{aligned} x_t &= \theta(L)\epsilon_t \\ \theta^{-1}(L)x_t &= \epsilon_t \end{aligned} \tag{11}$$

For the infinite AR representation to exist for an MA process, θ has to be invertible, i.e. $|\theta| < 1$.

Generally, one can find find values ϕ such that

$$(1 - \theta_1 L - \theta_2 L^2) = (1 - \phi_1 L)(1 - \phi_2 L)$$

For the second-order case, ϕ are found by solving $\phi_1 + \phi_2 = \theta_1$ and $-\phi_1\phi_2 = \theta_2$. This gives the **characteristical equation** with roots z. for the process to be **invertible**, both roots need to be larger than 0. $|z| = 1$ is called a unit root.

$$(1 - \phi_1 z)(1 - \phi_2 z) = 0 \tag{12}$$

Wold Theorem

Wold's (1938) theorem states that any trend stationary time series can be decomposed into the sum of deterministic, and thus predictable, and purely stochastic (non-predictable) elements.

$$y_t = d_t + \sum_{i}^{\infty} \alpha_i \epsilon_{t-i} \tag{13}$$

where $\alpha_0 = 1$ and $\sum_{i=0}^{\infty} \alpha_i^2 < K < \infty$. d_t is deterministic while ϵ_t is serially uncorrelated with finite expected value and variance.

Wold's theorem allows for the **linear estimation** of time series processes.

Every ARMA process can be **forecast** (approximated) by a **Wold polynomial** $\psi(L)$.

$$\psi(L) = \frac{\theta(L)}{\phi(L)}$$

$$x_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

$$\approx \frac{\theta(L)}{\phi(L)} u_t$$

Forecasting: Terminology

Many use **prediction** and **forecast** synonymously. Hoewer, *Clements and Hendry* argue that predictability is a theoretical notion, while forecastability describes when predictability can be exploited in practice.

Three types of forecast:

1 *Magical*: no recognizable relationship between cause and effect in theory or model

2 *Subjective*: Judgemental forecast, for example forecasting recession probability by averaging “business climate surveys”.

3 *Objective*: Purely econometric forecasts.

Forecasting: Single-Exponential Smoothing/ARIMA(0,1,1) 1

ARMA and ARIMA models translate into some popular forecasting methods (e.g. single exponential smoothing or Holt-Winter forecasting).

Consider single exponential smoothing (a filtering method).

$$\hat{x}_t = \alpha x_t + (1-\alpha)\hat{x}_{t-1} \quad \alpha \in (0,1) \tag{14}$$
$$= \alpha \sum_{j=0}^{t-1} (1-\alpha)^j x_{t-j} + (1-\alpha)^t \hat{x}_0$$
$$\hat{x}_t = \hat{x}_{t-1} + \alpha(x_t - \hat{x}_{t-1}) \quad \text{Error Correction Form} \tag{15}$$

The out-of-sample forecast is then performes by **flat extrapolation**, i.e. $\hat{x}_N(1) = \hat{x}_N(2) = ...$

Forecasting: Single-Exponential Smoothing/ARIMA(0,1,1) 2

SES is equivalent to basing forecasts on ARIMA(0,1,1), i.e. MA(1) on the first difference.

$$\begin{aligned}
x_t &= x_{t-1} + \epsilon_t + \theta\epsilon_{t-1} \\
\hat{x}_t &= x_{t-1} + \theta(x_{t-1} + \hat{x}_{t-1}) \\
&\text{Define } \alpha - 1 = \theta \\
\hat{x}_t &= \alpha x_{t-1} + (1-\alpha)\hat{x}_{t-1}
\end{aligned} \tag{16}$$

Forecasting: Gardner & McKenzie's approach to Holt's linear trend method

Define trend (or slope) T and local level L equations, as in Holt:

$$L_t = \alpha x_t + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$$

The resulting forecast in h steps is given by

$$\hat{x}_N(h) = L_n + hT_n$$
$$\text{s.t. } \hat{x}_{t-1}(1) = L_{t-1} + T_{t-1} = \hat{x}_t \quad (17)$$
$$(18)$$

Gardner & McKenzie's ARIMA(1,1,2) re-formulation:

$$\hat{x}_N(h) = L_n + (\sum_j^h \phi^j)T_n \quad (19)$$

Forecasting: Direct ARMA Forecasting

Suppose an ARMA(2,2) with known parameters ϕ, θ generates a data vector $(x_1, ..., x_T)$. This gives us the conditional expected value for the first forecast. Let $\mathbb{I}_t$ be the available information set at time t.

$$E[X_{t+1} \mid \mathbb{I}_t] = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \theta_1 \epsilon_1 + \theta_2 \epsilon_2 + E(\epsilon_t \mid \mathbb{I}_t)$$

Note that ϵ is a white noise process ($\Rightarrow E[\epsilon] = 0$).

$$\hat{X}_t(1) = \phi_1 X_t + \phi_2 X_{t-1} + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1}$$

In practice, parameters are estimated, p and q are determined empirically, and ϵ_t must be estimated, and these estimates are plugged in.

Forecasts for time periods further in the future will be derived iteratively.

Forecasting: Model Selection 1

The selection problem can be decomposed in two questions: the **family of processes** (e.g. ARMA) and the **specifiation of the process** (e.g. $p = 2, q = 0$).

Often, at least the first question will be answered with some respect to economic theory.

Sometimes, model-free forecasting is preferred. Totally model- and parameter-free forecasts usually out-perform parametric forecasts, however at the danger of overfitting (e.g. Gaussian processes, k-nearest-neighbor machine learning, ...).

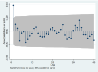
However, model-free forecasting provides little value for economic questions.

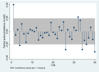
Forecasting: Model Selection 2

Box and Jenkins propose model selection upon visual selection of observational ACF and partial ACF plots.

The sample autocorrelation function ρ_k should be insignificant for $k > q$ in MA(q) processes.

For AR(p) models, the sample ACF should decay geometrically, which is of little help for model selection. However, the partial auto-correlation function should be significant only for lags $k \leq q$. Colloquially, partial correlation refers to correlation **conditional** on the relationship with larger lags.





Forecasting: Model Selection

A different (but not mutually exclusive approach) is applying information criteria and goodness-of-fit statistics.

Note: It is functionally different to apply information criteria to the estimation, or to apply goodness-of-fit to predictions (conditional on information you already have).

- log AIC: $= 2\frac{k}{T} - 2log(\hat{L}) = log(\sigma^2) + 2\frac{p+q+1}{T}$.
- log BIC: $= log(T)k + 2log(\hat{L}) = log(\sigma^2) + \frac{p+q+1}{T}log(T)$.
- Mean Squared Error MSE: $= \frac{\sum_t^T e_t^2}{T-p-q}$.