Methods of Economic Investigation II

Lecture 9

Oliver Linton

November 29, 2002

1 Some Fundamental Properties

- We start with univariate time series $\{y_t\}_{t=1}^{T}$.

- There are two main empirical features:

  - trends or absence of trends; stationarity/nonstationarit

  - dependence or independence

- Strong Stationarity. The stochastic process y is said to be strongly stationary if the vectors

$$(y_t, \ldots, y_{t+r}) \text{ and } (y_{t+s}, \ldots, y_{t+s+r})$$

have the same distribution for all t, s, r.

- Weak Stationarity. The stochastic process y is said to be weakly stationary if the vectors

$$(y_t, \ldots, y_{t+r}) \text{ and } (y_{t+s}, \ldots, y_{t+s+r})$$

have the same finite mean and variance for all t, s, r.

  - $E(y_t)$ and $\mathsf{var}(y_t)$ are finite and do not depend on t, and $\mathsf{cov}(y_t, y_s)$ only depends on $|t - s|$.

- Most of what we know is restricted to [weakly] stationary series, but in the last 20 years there have been major advances in the theory of nonstationary time series, see below. There are processes that are weakly stationary but not strongly stationary and, perhaps surprisingly, vice versa.

- Dependence: One measure of dependence is given by the covariogram [or correlogram]

$$\text{cov}(y_t,\ y_{t-s}) = \gamma_s\ ;\quad \rho_s = \frac{\gamma_s}{\gamma_0}.$$

- Note that stationarity was used here in order to assert that these moments only depend on the gap s and not on calendar time t as well.

- For i.i.d. series,

$$\gamma_s = 0 \text{ for all } s \neq 0.$$

For positively (negative) dependent series $\gamma_s > (<)0$. Economics series data often appear to come from positively dependent series.

- Uncorrelated sequences need not be i.i.d.

- A martingale difference sequence [MDS] y_t is one for which

$$E[y_t|y_{t-1},\ldots] = 0.$$

  Such a process has $\gamma_s = 0$ for all $s \neq 0$. However, martingale differences are not necessarily independent and can even be nonstationary [through say the variance].

- Mixing: (Covariance) We say that the process y is mixing if

$$\gamma_s \rightarrow 0 \text{ as } s \rightarrow \infty.$$

- This just says that the dependence [as measured by the covariance] on the past shrinks with horizon. This is an important property that is possessed by many models. The physical analogy with vodka-martini.

- A trivial example of a non-mixing process is

$$y_t = \varepsilon, \; t = \pm 1, \pm 2, \ldots$$

for some random variable ε.

- ARMA Models: The following is a very general class of models called ARMA(p, q) :

$$\begin{aligned} y_t &= \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} \\ &\quad + \varepsilon_t - \delta_1 \varepsilon_{t-1} - \cdots - \delta_q \varepsilon_{t-q}, \end{aligned}$$

  where ε_t is i.i.d., mean zero and variance σ^2.

  – We shall for convenience usually assume that $\mu = 0$.

  – We also assume for convenience that this model holds for $t = 0, \pm 1, \ldots$.

- It is convenient to write this model using lag polynomial notation

$$A(L) y_t = B(L) \varepsilon_t,$$

  where the lag polynomials

$$A(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$$

$$B(L) = 1 - \delta_1 L - \cdots - \delta_q L^q.$$

Here, $Ly_t = y_{t-1}$. The reason for this is to save space and to emphasize the mathematical connection with the theory of polynomials.

- Special case AR(1). Suppose that

$$y_t = \phi y_{t-1} + \varepsilon_t.$$

Here,

$$A(L) = 1 - \phi L.$$

  – We assume $|\phi| < 1$, which is necessary and sufficient for y_t to be a stationary process.

  – Now write

$$y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}.$$

  Continuing we obtain

$$\begin{aligned} y_t &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2 y_{t-2} \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots \\ &= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}, \end{aligned}$$

which is called the MA(∞) representation of the time series;

– this shows that y_t depends on all the past shocks.

- Now we calculate the moments of y_t using the stationarity property. This implies that for all t, s $E(y_t) = E(y_s)$. Therefore,

$$E(y_t) = \phi E(y_{t-1}),$$

which can be phrased as

$$\mu = \phi\mu \Leftrightarrow \mu = 0,$$

where

$$\mu = E(y_t) = E(y_{t-1}).$$

- Furthermore,

$$\mathsf{var}(y_t) = \phi^2 \mathsf{var}(y_{t-1}) + \sigma^2,$$

which implies that

$$\gamma_0 = \frac{\sigma^2}{1-\phi^2},$$

where

$$\gamma_0 = \text{var}(y_t) = \text{var}(y_{t-1}).$$

This last calculation of course requires that $|\phi| < 1$, which we are assuming for stationarity.

- Finally,

$$\text{cov}(y_t, y_{t-1}) = E(y_t y_{t-1}) = \phi E(y_{t-1}^2) + 0,$$

which implies that

$$\gamma_1 = \sigma^2 \frac{\phi}{1-\phi^2},$$

while

$$\begin{aligned} \text{cov}(y_t,\ y_{t-2}) &= E(y_t y_{t-2}) \\ &= \phi E(y_{t-1} y_{t-2}) \\ &= \phi^2 \frac{\sigma^2}{1-\phi^2}. \end{aligned}$$

- In general

$$\gamma_s = \phi^s \frac{\sigma^2}{1-\phi^2} \ ; \quad \rho_s = \phi^s.$$

Both γ_s and ρ_s decay geometrically towards zero.

- Moving Average MA(1). Suppose that

$$y_t = \varepsilon_t - \delta \varepsilon_{t-1},$$

where ε_t are i.i.d. mean zero with variance σ^2.

  – In this case,

$$E(y_t) = 0 \text{ and } \mathsf{var}(y_t) = \sigma^2(1+\delta^2).$$

  – Furthermore,

$$\begin{aligned}\mathsf{cov}(y_t,\ y_{t-1}) &= E\{(\varepsilon_t - \delta\varepsilon_{t-1})(\varepsilon_{t-1} - \delta\varepsilon_{t-2})\} \\ &= -\delta E(\varepsilon_{t-1}^2) \\ &= -\delta\sigma^2.\end{aligned}$$

Therefore,

$$\rho_1 = \frac{-\delta}{1+\delta^2}\ ,\ \rho_j = 0,\ j = 2, \ldots.$$

- This is a 1–dependent series, i.e., y_t and y_{t-j} are mutually independent for any $j \geq 2$. MA(q) is a q-dependent series.

- Process is automatically stationary for all δ.

- If $|\delta| < 1$, we say that the process is invertible and we can write

$$\sum_{j=0}^{\infty} \delta^j y_{t-j} = \varepsilon_t.$$

- In general ARMA(p, q), we can write

$$A(L)y_t = B(L)\varepsilon_t.$$

  - The stationarity condition for an ARMA(p, q) process is just that we need the roots of the autoregressive polynomial

$$1 - \phi_1 z - \cdots - \phi_p z^p$$

    to be outside unit circle.

  - Likewise the condition for invertibility is that the roots of the moving average polynomial

$$1 - \delta_1 L - \cdots - \delta_q L^q$$

    lie outside the unit circle.

– Assuming these conditions are satisfied we can write this process in two different ways

$$\begin{aligned}\frac{A(L)}{B(L)}y_t &= \sum_{j=0}^{\infty}\beta_j y_{t-j} = \varepsilon_t \\ y_t &= \frac{B(L)}{A(L)}\varepsilon_t = \sum_{j=0}^{\infty}\alpha_j \varepsilon_{t-j}.\end{aligned}$$

* The first line is called the $AR(\infty)$ representation, and expresses y_t in terms of its own past.

* The second line is called the $MA(\infty)$ representation, and expresses y_t in terms of the past history of the random shocks.

2 Estimation

In this section we discuss estimation of the autocovariance function of a stationary time series as well as the parameters of an ARMA model.

- Autocovariance and autocorrelation. Replace population quantities by sample

$$\widehat{\gamma}_s = \frac{1}{T-s} \sum_{t=s+1}^{T} (y_t - \overline{y})(y_{t-s} - \overline{y})$$

$$\widehat{\rho}_s = \widehat{\gamma}_s / \widehat{\gamma}_0.$$

  These sample quantities are often used to describe the actual series properties. Box-Jenkins analysis: ‘identification’ of the process by looking at the correlogram.

  – If the process is MA(q), then $\widehat{\rho}_{q+1}, \widehat{\rho}_{q+2}, \ldots$ should be approximately zero

- Box-Jenkins analysis: 'identification' of the process by looking at the correlogram. In practice, it is hard to identify any but the simplest processes, but the covariance function still has many uses.

- Estimation of ARMA parameters $\theta = (\phi, \delta)$. In the AR case, can think of the process as satisfying the linear regression

$$y_t = x_t'\phi + \varepsilon_t,$$

where $x_t = (y_{t-1}, \ldots, y_{t-p})'$.

- The error term satisfies

$$E[\varepsilon_t | x_t] = 0$$

and even ε_t is independent of x_t.

- However,

$$E[\varepsilon_t | x_1, \ldots, x_T] = E[\varepsilon_t | y_1, \ldots, y_T] \neq 0.$$

- If there are MA terms, then error term is $\varepsilon_t - \delta_1\varepsilon_{t-1} - \cdots - \delta_q\varepsilon_{t-q}$ which can be correlated with x_t.

- Estimation of ARMA parameters ϕ. Can 'invert' the autocovariance/autocorrelation function to compute an estimate of ϕ. For example in the AR(1) case, the parameter ρ is precisely the first order autocorrelation. In the MA(1) case, can show that the parameter θ satisfies a quadratic equation in which the coefficients are the autocorrelation function at the first two lags. A popular estimation method is the Likelihood under normality.

- A popular estimation criterion is the Gaussian Likelihood. Suppose that

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix} \sim N(0,\ \sigma^2 I),$$

then

$$\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \sim N(0,\ \Sigma)$$

for some matrix Σ.

– for an AR(1) process

$$\Sigma = \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ & & & \ddots & \vdots \\ & & & \ddots & 1 \end{bmatrix},$$

– for an MA(1) process

$$\Sigma = \sigma^2(1+\delta^2)\begin{bmatrix} 1 & \frac{-\delta}{1+\delta^2} & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}.$$

- In either case, the log likelihood function is

$$\ell = \frac{-T}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}y'\Sigma^{-1}y.$$

Maximize with respect to all the parameters.

- Distribution theory. The MLE is consistent and asymptotically normal provided the process is stationary and invertible.

$$T^{1/2}(\widehat{\theta}-\theta) \xrightarrow{D} N(0, I_{\theta\theta}^{-1}),$$

where $I_{\theta\theta}$ is the information matrix. In the special case of AR(1) process

$$n^{1/2}(\phi - \phi) \xrightarrow{D} N(0, 1 - \phi^2).$$

- In practice, $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ can be tough to find. We seek a helpful approach to computing the likelihood and an approximation to it, which is even easier to work with.

- The Prediction error decomposition is just a factorization of the joint density into the product of a conditional density and a marginal density,

$$f(x, z) = f(x|z)f(z).$$

We use this repeatedly and take logs to give

$$\begin{aligned} \ell(y_1, \ldots, y_T;\ \theta) &= \sum_{t=p+1}^{T} \ell(y_t|y_{t-1}, \cdots, y_1) \\ &\quad + \ell(y_1, \ldots, y_p). \end{aligned}$$

- This writes the log likelihood in terms of conditional distributions and a single marginal distribution. In AR cases the distribution of $y_t|y_{t-1}, \ldots, y_1$ is easy to find:

$$y_t|y_{t-1}, \ldots, y_1 \sim N(\phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}, \sigma^2).$$

- In the AR(1) case:
  - The conditional distribution of $y_t | y_{t-1}, \ldots, y_1$ is $N(\phi y_{t-1}, \sigma^2)$ so that

$$\ell_{t|t-1} \sim -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(y_t - \phi_1 y_{t-1})^2.$$

  - The marginal distribution of y_1 is $N(0, \sigma^2/(1-\phi^2))$, which means that

$$\ell(y_1) = -\frac{1}{2} \log \frac{\sigma^2}{1-\phi^2} - \frac{(1-\phi^2}{2\sigma^2} y_1^2.$$

  - Therefore, the full likelihood in the AR(1) case is

$$\begin{aligned} \ell \;\; = \;\; & -\frac{T-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^{T} (y_t - \phi y_{t-1})^2 \\ & -\frac{1}{2} \log \frac{\sigma^2}{1-\phi^2} - \frac{1-\phi^2}{2\sigma^2} y_1^2. \end{aligned}$$

- Often it is argued that $\ell(y_1)$ is small relative to $\sum_{t=2}^{T} \ell(y_t|y_{t-1}, \ldots, y_1)$, in which case we use

$$-\frac{T-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2.$$

- This criterion is equivalent to the least squares criterion, and has unique maximum

$$\widehat{\phi} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}.$$

This estimator is just the OLS of y_t on y_{t-1} [but using the reduced sample].

- The full MLE will be slightly different from the approximate MLE. In terms of asymptotic properties, the difference is negligible.

  – However, in finite sample there can be significant differences.

– Also, the MLE imposes that $\widehat{\phi}$ be less than one - as $\phi \rightarrow \pm\mathbf{1}$, $\ell \rightarrow -\infty$. The OLS estimate however can be either side of the unit circle.

3 Forecasting and Prediction

- Let the sample be $\{y_1, \ldots, y_T\}$. Suppose that

$$y_t = \phi y_{t-1} + \varepsilon_t \ , \quad |\phi| < 1,$$

where we first assume that ϕ is known.

- Want to forecast $y_{T+1}, y_{T+2}, \ldots, y_{T+r}$ given the sample information. We have

$$y_{T+1} = \phi y_T + \varepsilon_{T+1}.$$

Therefore, forecast y_{T+1} by

$$\widehat{y}_{T+1|T} = E[y_{T+1}|y_1, \ldots, y_T] = \phi y_T.$$

- The forecast error is ε_{T+1}, which is mean zero and has variance σ^2.

- What about forecasting r periods ahead?

$$y_{T+r} = \phi^r y_T + \phi^{r-1}\varepsilon_{T+1} + \cdots + \varepsilon_{T+r}.$$

Therefore, let

$$\widehat{y}_{T+r|T} = \phi^r y_T$$

be our forecast.

- The forecast error $\widehat{y}_{T+r|T} - y_{T+r}$ has mean zero and variance

$$\sigma^2(1 + \phi^2 + \cdots + \phi^{2r-2}).$$

- In practice, we must use an estimate of ϕ, so that

$$\widehat{y}_{T+r|T} = \widehat{\phi}^r y_T,$$

where $\widehat{\phi}$ is estimated from sample data. If γ is estimated well, then this will not make much difference.

- Forecast interval

$$\widehat{y}_{T+r|T} \pm 1.96 \cdot SD\ ,$$

$$SD = \sigma^2(1 + \phi^2 + \cdots + \phi^{2r-2}).$$

This is to be interpreted like a confidence interval. Again we must replace the unknown parameters by consistent estimates.

- How do you evaluate forecasts? After the forecast period has been realized you can compare the actual outturn with that predicted.

- For example, let $\widehat{y}_{T+r|T}$, $r = 1, \ldots, R$ be a forecast of R periods ahead and let y_{T+r}, $r = 1, \ldots, R$ be the realized data. Then define the average squared error

$$
\begin{aligned}
MSE &= \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{y}_{T+r|T} - y_{T+r}\right)^2 . \\
RMSE &= \left[\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{y}_{T+r|T} - y_{T+r}\right)^2\right]^{1/2} . \\
MAE &= \frac{1}{R}\sum_{r=1}^{R}\left|\widehat{y}_{T+r|T} - y_{T+r}\right| .
\end{aligned}
$$

4 Autocorrelation and Regression

- Regression models with correlated disturbances

$$y_t = \beta' x_t + u_t,$$

where x_t is exogenous, i.e., is determined outside the system; fixed regressors are an example. There are a number of different variations on this theme - strongly exogenous and weakly exogenous. A weakly exogenous process could include lagged dependent variables. We will for now assume strong exogeneity.

- We also suppose that

$$E(u_t u_s) \neq 0 \text{ for some } s \neq t.$$

- As an example, suppose that

$$\ln GNP = \beta_1 + \beta_2 \text{time } + u_t.$$

We expect the deviation from trend, u_t, to be positively autocorrelated reflecting the business cycle, i.e., not i.i.d. Recession quarter tends to be followed by recession quarter.

- We can write the model in matrix form

$$y = X\beta + u,$$

$$\begin{aligned} E(uu') &= \Sigma \\ &= \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-1} \\ & & & \ddots & \gamma_2 \\ & & & \ddots & \gamma_0 \end{bmatrix}. \end{aligned}$$

- The consequences for estimation and testing of β are the same as with heteroskedasticity: OLS is consistent and unbiased, but inefficient, while the SE's are wrong.

1.

- Specifically,

$$var(\widehat{\beta}) = (X'X)^{-1}X'\mathbf{\Sigma}X(X'X)^{-1},$$

where

$$\psi_T = X'\mathbf{\Sigma}X = \sum_{t=1}^{T}\sum_{s=1}^{T} x_t x_t' \gamma_{|t-s|}.$$

- A naive implementation of the White strategy is going to fail here, i.e.,

$$\begin{aligned}\widehat{\psi}_T &= X'\begin{pmatrix} \widehat{u}_1^2 & \widehat{u}_1\widehat{u}_2 & \cdots & \widehat{u}_1\widehat{u}_T \\ & \widehat{u}_1^2 & & \\ & & \ddots & \\ & & & \widehat{u}_T^2 \end{pmatrix} X \\ &= \sum_{t=1}^{T}\sum_{s=1}^{T} x_t x_t' \widehat{u}_t \widehat{u}_s\end{aligned}$$

is inconsistent. This is basically because there are too many random variables in the sample matrix, in fact order T^2, whereas in the independent but heterogeneous case there were only order T terms.

1.

- The correct approach is to use some downweighting that concentrates weight on a smaller fraction of their terms. Bartlett/White/Newey/West SE's: Replace by sample equivalents and use weights

$$w(j) = 1 - \frac{j}{n+1},$$

so that

$$\widehat{\psi}_T = \sum\sum_{t,s:|t-s|\leq n(T)} X_t X_s' w\left(|t-s|\right) \widehat{u}_t \widehat{u}_s.$$

This also ensures a positive definite covariance matrix estimate. Provides consistent standard errors.

- An alternative strategy is to parameterize u_t by, say, an ARMA process and do maximum likelihood

$$\ell = -\frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}(y - X\beta)' \Sigma(\theta)^{-1}(y - X\beta).$$

- Efficient estimate of β (under Gaussianity) is a sort of GLS

$$\widehat{\beta}_{ML} = \left(X'\boldsymbol{\Sigma}(\widehat{\theta})^{-1}X\right)^{-1} X'\boldsymbol{\Sigma}(\widehat{\theta})^{-1}y,$$

where $\widehat{\theta}$ is the MLE of θ. This will be asymptotically efficient when the chosen parametric model is correct.

5 Testing for Autocorrelation

- Suppose that we observe u_t, which is generated from an $AR(1)$ process
$$u_t = \rho u_{t-1} + \varepsilon_t,$$
where ε_t are i.i.d.

- The null hypothesis is that u_t is i.i.d., i.e.,
$$H_0 : \rho = 0 \quad \text{vs.} \quad H_A : \rho \neq 0.$$
This is used as (a) general diagnostic, and (b) efficient markets.

- General strategy: use LR, Wald or LM tests to detect departures.

- The LM test is easiest, this is based on
$$\begin{aligned} LM &= T \left(\frac{\sum_t \widehat{u}_t \widehat{u}_{t-1}}{\sum_t \widehat{u}_{t-1}^2} \right)^2 \\ &= T r_1^2 \xrightarrow{D} \chi_1^2, \end{aligned}$$

where $\widehat{u}_t$ are the OLS residuals. Therefore, we reject the null hypothesis when LM is large relative to the critical value from χ^2_1.

- This approach is limited to two-sided alternatives. We can however also use the signed version, $T^{\frac{1}{2}} r_1$, which satisfies

$$T^{\frac{1}{2}} r_1 \xrightarrow{D} N(0,1)$$

under the null hypothesis.

- The Durbin-Watson d is

$$d = \frac{\sum_{t=2}^{T} (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T} \widehat{u}_t^2}.$$

This is always printed out by many regression packages.

- Using the approximation

$$d \approx 2(1 - r_1),$$

we have [under the null hypothesis]

$$T^{\frac{1}{2}} \left(1 - \frac{d}{2}\right) \rightarrow N(0, 1).$$

- Generalization (test against $AR(p)$). Suppose that

$$u_t = \rho_1 u_{t-1} + \ldots + \rho_p u_{t-p} + \varepsilon_t,$$

where ε_t are i.i.d. The null hypothesis is that u_t is i.i.d., i.e.,

$$H_0 : \rho_1 = \ldots = \rho_p = 0$$

$$\text{vs.} \quad H_A \text{some } \rho_j \neq 0.$$

- Box-Pierce Q

$$Q = T \sum_{j=1}^{P} r_j^2 \xrightarrow{D} \chi_P^2.$$