ApEc 8213: Econometric Analysis III -- Lecture #11

Time Series Econometrics, Part 4
Hansen, Chapter 14, Sections 14.30 – 14.45

This lecture will focus on estimating AR(p) time series models, testing for serial correlation, and selecting the "best" model. So it is much more applied than previous lectures. Hansen does not explain how to estimate MA(q) or ARMA(p, q) models. Hamilton (1994) explains how to estimate those models.

I. Estimating Autoregressive Models (14.30)

Suppose that Y_t is stationary, ergodic and non-deterministic. (Note that we are ***not*** assuming that it follows an AR(p) process.) Consider an AR(p) model for Y_t:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \dots + \alpha_p Y_{t-p} + e_t = X_t'\alpha + e_t \quad (14.45)$$

where $X_t = (1, Y_{t-1}, \dots Y_{t-p})'$ and $\alpha = (\alpha_0, \alpha_1, \dots \alpha_p)'$ are convenient notation. The **vector $\boldsymbol{\alpha}$** is ***defined*** as the **linear projection** of Y_t on X_t (i.e. $\alpha \equiv (E[X_tX_t'])^{-1}(E[X_tY_t])$), and **equation (14.45)** ***defines*** $\boldsymbol{e_t}$, which has variance σ^2 ($< \infty$).

The OLS estimator of this AR(p) model is:

$$\hat{\alpha} = (\textstyle\sum_{t=1}^{n} X_tX_t')^{-1}(\sum_{t=1}^{n} X_tY_t)$$

In fact, we assume that there are $n + p$ observations for Y_t because we need to use the first p as "initial conditions" for $t = 1$ to $t = p$ (since $X_1 = (1, Y_0, \ldots Y_{1-p})'$).

We also want to estimate σ^2 (variance of e_t). There are two options here:

$$\hat{\sigma}^2 = (1/n) \sum_{t=1}^{n} \hat{e}_t^2$$

$$s^2 = (1/(n - p - 1)) \sum_{t=1}^{n} \hat{e}_t^2$$

where $\hat{e}_t = Y_t - X_t'\hat{\alpha}$.

If Y_t is strictly stationary and ergodic, then by Theorem 14.5 so are X_tX_t' and X_tY_t, and both have finite means if $E[Y_t^2] < \infty$. Then, the Ergodic Theorem (Theorem 14.9) implies that:

$$(1/n)\sum_{t=1}^{n} X_t Y_t \underset{p}{\rightarrow} E[X_tY_t] \qquad (14.47)$$

$$(1/n)\sum_{t=1}^{n} X_t X_t' \underset{p}{\rightarrow} E[X_tX_t'] = \boldsymbol{Q}$$

Theorem 14.28 shows that $\boldsymbol{Q} > 0$ (is positive definite), so by the continuous mapping theorem we have:

$$\hat{\alpha} = (\sum_{t=1}^{n} X_t X_t')^{-1}(\sum_{t=1}^{n} X_t Y_t) \underset{p}{\rightarrow} \alpha$$

Thus, **$\hat{\boldsymbol{\alpha}}$ is a consistent estimate of $\boldsymbol{\alpha}$**. Hansen says that it is "straightforward" to show that $\hat{\sigma}^2$ consistently estimates σ^2.

This is summed up as:

Theorem 14.29. If Y_t is strictly stationary, ergodic, not purely deterministic (i.e. $\sigma^2 > 0$), and $E[Y_t^2] < \infty$, then for any p, $\hat{\alpha} \underset{p}{\to} \alpha$ and $\hat{\sigma}^2 \underset{p}{\to} \sigma^2$ as $n \to \infty$.

Note again: this does *not* assume that Y_t is an AR(p) process.

II. Asymptotic Distribution of OLS Estimator (14.31-14.33)

To calculate the **asymptotic distribution of $\hat{\boldsymbol{\alpha}}$**, we need to make some more assumptions. First, we **assume that e_t is an MDS**. Note that the "conditioning set" for e_t includes lagged Y_t terms, so it includes $X_t = (1, Y_{t\text{-}1}, \ldots Y_{t\text{-}p})'$. Then:

$$E[X_t e_t | \mathscr{F}_{t\text{-}1}] = X_t E[e_t | \mathscr{F}_{t\text{-}1}] = 0$$

This means that $\boldsymbol{X_t e_t}$ **is also an MDS**. Also, $\text{Var}(X_t e_t) < \infty$ if e_t has a finite fourth moment. We can assume this for e_t, that is assume that $E[e_t^4] < \infty$; Hansen shows that this implies that the fourth moment of Y_t is also $< \infty$.

Applying Theorem 14.11 (martingale difference CLT) yields:

$$(1/\sqrt{n})\sum_{t=1}^{n} X_t e_t \underset{d}{\to} N(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = E[X_t X_t' e_t^2]$.

To sum up:

Theorem 14.30. If Y_t is an AR(p) process, all roots of $\alpha(z)$ lie outside the unit circle (Y_t is strictly stationary and ergodic), e_t is an MDS, $\mathrm{E}[e_t | \mathscr{F}_{t\text{-}1}] = 0$, $\mathrm{E}[e_t^4] < \infty$, and $\mathrm{E}[e_t^2] > 0$, then as $n \to \infty$,

$$\sqrt{n}(\hat{\alpha} - \alpha) \underset{d}{\to} \mathrm{N}(0, \mathbf{V})$$

where $\boldsymbol{V} = \boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}$.

Note: ***Unlike the previous section***, we are **assuming that $\boldsymbol{Y_t}$ follows an AR(p) process**. Also, the $\boldsymbol{\Sigma}$ matrix allows for heteroscedasticity in the sense the variance of e_t could be related to X_t.

Asymptotic Distribution under Homoscedasticity

Theorem 14.30 **can be simplified if we assume that $\boldsymbol{e_t}$ is a homoscedastic MDS**. That is, assume that $\mathrm{E}[e_t^2 | \mathscr{F}_{t\text{-}1}] = \sigma^2$. Then $\boldsymbol{\Sigma} = \mathrm{E}[X_t X_t' e_t^2] = \mathrm{E}[X_t X_t' \mathrm{E}[e_t^2 | \mathscr{F}_{t\text{-}1}]] = \boldsymbol{Q}\sigma^2$, and we have:

Theorem 14.31. Under the assumptions of Theorem 14.30, and the assumption that $\mathrm{E}[e_t^2 | \mathscr{F}_{t\text{-}1}] = \sigma^2$, as $n \to \infty$:

$$\sqrt{n}(\hat{\alpha} - \alpha) \underset{d}{\to} \mathrm{N}(0, \boldsymbol{V}^0)$$

where $\boldsymbol{V}^0 = \sigma^2\boldsymbol{Q}^{-1}$.

In the special case of a simple AR(1) model with $\alpha_0 = 0$ (no intercept term), $\boldsymbol{Q} = \mathrm{E}[Y_t^2] = \sigma^2/(1 - \alpha_1^2)$, so that:

$$\sqrt{n}(\hat{\alpha} - \alpha) \underset{d}{\rightarrow} \mathrm{N}(0, 1 - \alpha_1^2)$$

This shows that **as $\boldsymbol{\alpha_1}$ is closer to 1** (in absolute value) **the more precisely $\hat{\boldsymbol{\alpha}}$ is estimated**. Intuitively, the role played by α_1 via $\alpha_1 Y_{t-1}$ in determining Y_t, relative to the role played by e_t, is increasing. This increases the "signal to noise" ratio in determining Y_t.

In fact, there is **no reason to expect $\boldsymbol{e_t}$ to be homoscedastic**, so you should **use the results of Theorem 14.30**, and make sure your software does that (it may not!).

Asymptotic Distribution under General Dependence

Recall that MDS errors are a special case of more general white noise errors. **If we do not assume MDS errors** but instead assume white noise errors, **or if we relax the assumption that $\boldsymbol{Y_t}$ is an AR(p) process**, so that we are using the AR(p) model only as an approximation for Y_t, then **we cannot use the MDS CLT** (Theorem 14.11). **However**, if we **assume $\boldsymbol{Y_t}$ is a strong mixing process**, we can use the CLT for mixing processes (Theorem 14.15).

Theorem 14.32. Assume that Y_t is a strictly stationary, ergodic and, for some $r > 4$, $E[|Y_t|^r] < \infty$ and the mixing coefficients satisfy $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-4/r} < \infty$. Define α as the best linear projection coefficients, $\alpha = (E[X_t X_t'])^{-1} E[X_t Y_t]$ for an AR(p) model with projection errors e_t. The OLS estimator of α is $\hat{\alpha}$. Then:

$$\boldsymbol{\Omega} = \sum_{\ell=-\infty}^{\infty} E[X_{t-\ell} X_t' e_t e_{t-\ell}]$$

is convergent and

$$\sqrt{n}(\hat{\alpha} - \alpha) \underset{d}{\rightarrow} N(0, \boldsymbol{V})$$

where $\boldsymbol{V} = \boldsymbol{Q}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}^{-1}$.

The difference here, **compared to Theorem 14.30**, is the difference between $\boldsymbol{\Sigma}$ in that theorem and $\boldsymbol{\Omega}$ here. Here, $\boldsymbol{\Omega}$ includes correlation over time in $X_t e_t$.

III. Covariance Matrix Estimation

Under the "correct specification" assumptions of Theorem 14.30, covariance matrix estimation is the same as used for cross-sectional data. More specifically, if e_t is assumed to be **homoscedastic**, then:

$$\widehat{\boldsymbol{V}}^0 = \hat{\sigma}^2 \widehat{\boldsymbol{Q}}^{-1}, \quad \text{where } \widehat{\boldsymbol{Q}} = \frac{1}{n} \sum_{t=1}^{n} X_t X_t'$$

You can also use s^2 (see page 2) instead of $\hat{\sigma}^2$.

In the more general case where e_t can be **heteroscedastic**:

$$\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{Q}}^{-1}, \quad \text{where } \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n} X_t X_t' \hat{e}_t^{2} \qquad (14.48)$$

As usual, standard errors for the individual coefficient estimates ($\hat{\alpha}_j$), which can be denoted by $s(\hat{\alpha}_j)$, are the square root of the diagonal elements of $\widehat{\boldsymbol{V}}$. Thus we have:

Theorem 14.33. Under the assumptions of Theorem 14.30, as $n \to \infty$, $\widehat{\boldsymbol{V}} \underset{p}{\to} \boldsymbol{V}$ and:

$$\frac{\hat{\alpha}_j - \alpha_j}{s(\hat{\alpha}_j)} \underset{d}{\to} \mathrm{N}(0,\, 1) \quad \text{for all } j = 0, 1, \ldots p$$

Covariance Matrix Estimation under General Dependence

Under the **less restrictive assumptions** of Theorem 14.32, we **need an estimate of $\boldsymbol{\Omega}$**. Such estimates are called **Heteroscedasticity and Autocorrelation Consistent (HAC)** or **Heteroscedasticity and Autocorrelation Robust (HAR)** covariance matrix estimators.
To see how this works, define the vector series $u_t = X_t e_t$ and autocovariance matrices $\boldsymbol{\Gamma}(\ell) = \mathrm{E}[u_{t-\ell} u_t']$, so that:

$$\boldsymbol{\Omega} = \sum_{\ell=-\infty}^{\infty} \boldsymbol{\Gamma}(\ell)$$

This sum is convergent (see Theorem 14.32), so the $\boldsymbol{\Gamma}(\ell)$ matrices converge to zero as $\ell \to \infty$. Thus **$\boldsymbol{\Omega}$ can be approximated by a finite sum of these matrices**, such as:

$$\mathbf{\Omega}_M = \sum_{\ell=-M}^{M} \mathbf{\Gamma}(\ell)$$

where M is sometimes called the lag truncation number (others call it the bandwith). **An estimator of $\mathbf{\Gamma}(\ell)$ is**:

$$\hat{\mathbf{\Gamma}}(\ell) = \frac{1}{n}\sum_{t=1+\ell}^{n+\ell} \hat{u}_{t-\ell}\hat{u}_t{}'$$

where $\hat{u}_t = X_t\hat{e}_t$. By the ergodic theorem we can show that for any ℓ, $\hat{\mathbf{\Gamma}}(\ell) \underset{p}{\rightarrow} \mathbf{\Gamma}(\ell)$. Thus for any fixed M, the estimator

$$\hat{\mathbf{\Omega}}_M = \sum_{\ell=-M}^{M} \hat{\mathbf{\Gamma}}(\ell) \qquad (14.49)$$

is a consistent estimate for $\mathbf{\Omega}_M$ (but not necessarily for $\mathbf{\Omega}$).

How should *M* be selected? We can specify that it should increase with n. Hansen says that if the rate at which M increases is "sufficiently slow", then $\hat{\mathbf{\Omega}}_M$ converges to $\mathbf{\Omega}$.

In practice, there are **two problems regarding the choice of *M***. First, $\hat{\mathbf{\Omega}}_M$ can change "non-smoothly" with M, making this estimate sensitive to the choice of M. Second, $\hat{\mathbf{\Omega}}_M$ may not be positive semi-definite (and thus may not be a valid covariance matrix estimator), which could happen if Y_t is strongly negatively correlated.

Both of these problems **can be solved** by using a ***weighted* sum** of $\hat{\mathbf{\Gamma}}(\ell)$. Newey and West (1987) suggested:

$$\widehat{\boldsymbol{\Omega}}_{\text{nw}} = \sum_{\ell=-M}^{M} \left(1 - \frac{|\ell|}{M+1}\right) \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\ell}) \qquad (14.50)$$

Newey and West showed that this estimator is postitive semi-definite and a smooth function of M.

For $\widehat{\boldsymbol{\Omega}}_{\text{nw}}$ to be a consistent estimate of $\boldsymbol{\Omega}$, M must increase to infinity with n. Sufficient conditions are:

Theorem 14.34. Under the assumptions of Theorem 14.32, and assuming $\sum_{\ell=1}^{\infty} \alpha(\boldsymbol{\ell})^{1/2 - 4/r} < \infty$, if $M \rightarrow \infty$ but also $M^3/n = O(1)$, then as $n \rightarrow \infty$, $\widehat{\boldsymbol{\Omega}}_{\text{nw}} \underset{p}{\rightarrow} \boldsymbol{\Omega}$.

Note: $M^3/n = O(1)$ means that M grows no faster than $n^{1/3}$.

But **how does one select *M*?**. Several papers in the 1990s examined this. For $\widehat{\boldsymbol{\Omega}}_{\text{nw}}$, Andrews (1991) suggested:

$$M = \left(6\frac{\rho^2}{(1-\rho^2)^2}\right)^{1/3} n^{1/3}$$

where ρ is the serial correlation of u_t ($= X_t e_t$). Andrews suggested a way to estimate ρ (if it is scalar. Hansen also suggests trying $\rho = 0.50$, which implies $M \approx 1.4 \times n^{1/3}$.

IV. Serial Correlation Tests, Model Selection (14.36-14.38)

Perhaps the first thing that one should test for is whether Y_t is serially correlated over time. Consider an AR(p) model:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + e_t$$

where e_t is an MDS error term. The **null hypothesis** is that the **coefficients on all the lagged Y_t terms are zero**:

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$$

$$H_1: \alpha_j \neq 0 \text{ for at least one } j \geq 1$$

This can be tested using a standard Wald test. Estimate the equation by OLS, and use the covariance matrix given in equation (14.48). **Question**: Why not use the Newey-West correlation matrix in equation (14.50)?

Of course, this test requires that you have an idea for what p should be. If there is some autocorrelation it will probably be strongest in the first one or two terms, so you could choose $p = 2$. **Question**: Why not "cast a wide net" at set p to a much larger number, such as 5, 10 or 20? Hansen also suggests setting p to match seasonal periodicity such as 4 for quarterly data and 12 for monthly data.

Testing for Omitted Serial Correlation

Suppose that you have an AR(p) model and you want to test whether there is additional correlation in Y_t that is not accounted for in your model. Your model is:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + u_t \qquad (14.52)$$

The **null hypothesis** is that ***u_t* is serially uncorrelated**. The **alternative hypothesis** is that there is **serial correlation in *u_t***, which we can model as:

$$u_t = \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} + e_t \qquad (14.53)$$

The **null and alternative hypotheses** are:

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_q = 0$$

$$H_1: \theta_j \neq 0 \text{ for at least one } j \geq 1$$

One (too) simple approach would be to estimate equation (14.52), generate $\hat{u}_t$ based on those estimates, and estimate (14.53) using the generated $\hat{u}_t$ terms. Yet using OLS for the second estimate would not give correct standard errors. One could get the **correct standard errors using GMM** methods, **but there is an easier method**.

Consider the simple case where $q = 1$, so that equation (14.53) is $u_t = \theta_1 u_{t-1} + e_t$. Lag equation (14.52) by 1 period:

$$Y_{t-1} = \alpha_0 + \alpha_1 Y_{t-2} + \alpha_2 Y_{t-3} + \ldots + \alpha_p Y_{t-p-1} + u_{t-1}$$

Multiply this by θ and subtract from equation (14.52):

$$Y_t - \theta_1 Y_{t-1} = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + u_t$$

$$- \theta_1 \alpha_0 - \theta_1 \alpha_1 Y_{t-2} - \theta_1 \alpha_2 Y_{t-3} - \ldots - \theta_1 \alpha_p Y_{t-p-1} - \theta_1 u_{t-1}$$

$$Y_t = \alpha_0(1 - \theta_1) + (\alpha_1 + \theta_1)Y_{t\text{-}1} + (\alpha_2 - \theta_1\alpha_1)Y_{t\text{-}2} + \ldots - \theta_1\alpha_p Y_{t\text{-}p\text{-}1} + e_t$$

This AR(p+1) process **simplifies to an AR(p) process if $\boldsymbol{\theta_1 = 0}$**. Thus the null hypothesis is that the coefficient on $Y_{t\text{-}p\text{-}1}$ is zero, which can be tested using a standard t-test.

More generally, one can test whether the error u_t is an AR(q) process by testing whether Y_t is an AR(p+q) process. This can be done using a **Wald test** for the **null hypothesis** that the **coefficients on $\boldsymbol{Y_{t\text{-}p\text{-}1}}$ to $\boldsymbol{Y_{t\text{-}p\text{-}q}}$ all equal zero**.

In almost all cases, testing for this serial correlation is not testing something derived from an economic theory. **What is really happening here is that one is engaged in "model selection"**, that is trying to find the specification that best fits the data. Next, consider **"model selection tools"**.

Model Selection

A **general method** to select a model is to **minimize the Akaike information criterion (AIC)**:

$$\text{AIC}(p) = n \times \log(\hat{\sigma}^2(p)) + 2p$$

where $\hat{\sigma}^2(p)$ is the estimated residual variance from an AR(p) regression. In general, adding more lagged Y_t terms (increasing p) will improve the fit of the model and thus reduce $\hat{\sigma}^2(p)$, but notice that the **AIC adds a "penalty"** ($2p$) **for adding new parameters**.

The AIC can be interpreted as estimating the divergence between the fitted model and the "true" conditional density. It is also a monotonic transformation of an estimator of the squared error of a "one-time-period-ahead" forecast based on the model. Whatever interpretation you choose, the way to use the AIC is to **choose the model** (specification) **with the lowest AIC**.

One thing to be careful about is that you should **always compare estimates using the same sample size**, and **changing *p* will change the sample size**. Hansen recommends using the highest value of p for the models you want to compare to "fix" the sample size that you will use for all the models. For example, if the highest p is 5, and you have a total sample of 100, set n to 95 and use only observations 6 to 100; observations 1 to 5 will be used only to provide lagged Y_t for observations 6-10.

Hansen provides some examples in Section 14.39 for U.S. macro data of how to use AIC to select the "best" model.

V. Time Series Models (14.40)

Time series models introduce other (stationary) variables. Hansen starts by not worrying much about whether estimates are causal, but simply focuses on linear projections of Y_t on some variables X_t. **To start**, let Y_t be the "dependent" variable and let $\boldsymbol{X}_t$ be a "**vector of**

regressors" that includes an intercept (constant term) and also **could include lagged values of Y_t**. The model is:

$$Y_t = X_t'\beta + e_t \qquad (14.54)$$

The **vector β is not defined as a causal relationship** but is simply **defined as a linear projection** of Y_t on X_t:

$$\beta = (\mathrm{E}[X_tX_t'])^{-1}\mathrm{E}[X_tY_t] \qquad (14.55)$$

Once β is defined this way, then **e_t is defined by equation (14.54)**. That is, it is simply a "**projection error**", and so it has the property $\mathrm{E}[X_te_t] = 0$. Any other properties of e_t will be determined by the properties of X_t and Y_t.

Note finally that there is an implicit assumption here that $\mathrm{E}[X_tX_t'] > 0$ (is positive definite), since this is needed to be able to invert this matrix as in equation (14.55).

The OLS estimate of β as defined in equation (14.55) is:

$$\hat{\beta} = (\textstyle\sum_{t=1}^{n} X_tX_t')^{-1}(\textstyle\sum_{t=1}^{n} X_tY_t).$$

The **assumption** that the joint distribution **(Y_t, X_t)** is **strictly stationary and ergodic** implies that **$\hat{\beta}$** is a **consistent** estimate of β. **Under the assumptions of Theorem (14.32), $\hat{\beta}$ is asymptotically normal** with the (general dependence) asymptotic covariance matrix given by that theorem.

If we make a somewhat stronger assumption that e_t is an MDS, the asymptotic covariance matrix simplifies to the one given for Theorem (14.30).

The assumption that e_t is MDS requires that $\mathrm{E}[e_t | \mathscr{F}_{t-1}] = 0$ where $\mathscr{F}_{t-1}$ is the information set on which (e_{t-1}, X_t) is adapted (which means $\mathrm{E}[e_{t-1}, X_t | \mathscr{F}_{t-1}] = (e_{t-1}, X_t)$). It may seem strange that past information can predict X_t, but in some contexts it is OK, e.g. for autoregression models $X_t = (1, Y_{t-1}, \ldots Y_{t-p})$. This assumption gives us the following:

$$\mathrm{E}[X_t e_t | \mathscr{F}_{t-1}] = X_t \mathrm{E}[e_t | \mathscr{F}_{t-1}] = 0$$

Thus $(X_t e_t, \mathscr{F}_t)$ is an MDS, so we can apply the MDS CLT.

We can summarize this as follows:

Theorem 14.35. If (Y_t, X_t) is strictly stationary, ergodic, with finite second moments, and $\boldsymbol{Q} = \mathrm{E}[X_t X_t'] > 0$, then β in equation (14.55) is uniquely defined and the OLS estimator $\hat{\beta}$ is consistent: $\hat{\beta} \underset{p}{\rightarrow} \beta$.

Moreover, if $\mathrm{E}[e_t | \mathscr{F}_{t-1}] = 0$ where $\mathscr{F}_{t-1}$ is an information set to which (e_{t-1}, X_t) is adapted, $\mathrm{E}[|Y_t|^4] < \infty$, and $\mathrm{E}[\|X_t\|^4] < \infty$, then:

$$\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\rightarrow} \mathrm{N}(0, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}) \qquad (14.56)$$

as $n \rightarrow \infty$, where $\boldsymbol{\Omega} = \mathrm{E}[X_t X_t' e_t^2]$.

Alternatively, if for some $r > 4$, $\mathrm{E}[|Y_t|^r] < \infty$, $\mathrm{E}[\|X_t\|^r] < \infty$, and the mixing coefficients for (Y_t, X_t) satisfy $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1 - 4/r} < \infty$, then (14.56) holds with:

$$\mathbf{\Omega} = \sum_{\ell=-\infty}^{\infty} \mathrm{E}[X_{t-\ell} X_t' e_t e_{t-\ell}].$$

Subsection (14.41) considers a specific time series model where the "other" variables could include lagged values of those variables. These are ARDL models, where DL is for "distributed lag". This is optional material.

VI. Time Trends (14.42)

Many time series have means that change over time. One simple way to specify this is:

$$Y_t = T_t + u_t$$

where T_t represents a "trend" component and u_t is the stochastic component.

The trend is often modeled as linear or quadratic:

$$T_t = \beta_0 + \beta_1 t \quad \text{(linear)}$$

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 \quad \text{(quadratic)}$$

Clearly Y_t is not stationary (**why?**). A common way to make Y_t stationary is to "detrend" it: estimate T_t and subtract it from Y_t. For example, if T_t is linear, use OLS to estimate:

$$Y_t = \beta_0 + \beta_1 t + u_t \qquad (14.57)$$

Subtracting $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$ from Y_t yields $\hat{u}_t$, which is (asymptotically) stationary.

Hansen says that "most of our conventional procedures work just fine with time trends … as regressors." That is, we do not need to detrend (14.57) but instead we can estimate it directly. He shows (pages 503-04) that applying OLS to (14.57) yields consistent estimates of the β terms. In fact, for an AR-DL model that adds a linear or quadratic time trend, OLS estimation with the linear or quadratic trend will produce consistent estimates of all parameters.

In Section 14.43, Hansen provides an example of estimating DL and AR-DL models of the "Phillips curve" that depicts the relationship between inflation and unemployment.

Chapter 14 ends with a (rather simple) discussion of Granger Causality in (Section 14.44), testing for omitted serial correlation in u_t in ARDL models (Section 14.45) and using the bootstrap for time series model estimation (Section 14.44). All of this is optional material.

VII. Exogeneity and Causality (Davidson and MacKinnon, Chapter 18, Section 2)

Economists and econometricians often refer to **exogenous** or **endogenous** variables without being precise about the meaning. In fact, there several types of exogeneity. For time series models (models where the observations have a specific order over time) the concepts are rather subtle. For "non time series models" exogeneity is much simpler and quite intuitive. Finally, there is the concept of **causality**, which is related to exogeneity but is a distinct concept.

In econometrics, a simple definition of exogenous is an X variable that is **uncorrelated with the error term. In contrast**, many **economic models use "exogenous" in a different way**. They specify that exogenous variables "come from outside the model" or "are not controlled by the economic decisionmaker".

This is **not the same** as the "econometric" definition. First, it is **possible that an exogenous variable** (i.e. one determined outside the system) **is correlated with the error** in a regression. An example is estimation of a farm production function. Suppose that two variables that determine productivity are rainfall and sunshine. These are clearly exogenous in the economic sense. They are also correlated (it only rains on cloudy days). Suppose we have data on rainfall but not on sunshine, so sunshine is an omitted variable that goes into the error term. In this case rainfall is correlated with the error term, since rainfall and sunshine are correlated.

Second, there can be a model with **variables that are endogenous in the economic sense** on the right-hand side of an econometric model, **yet the error term is not correlated with them**. Consider another farm production function. Suppose output is determined by fertilizer, labor and rainfall. We do not have data on rainfall, so it ends up in the error term. Fertilizer and labor inputs are clearly endogenous in the economic sense; the farmer chooses them. Yet the error term (rainfall) is pretty random and thus may be uncorrelated with fertilizer and labor (assuming they are chosen before rainfall occurs).

We now examine econometric/statistical definitions of exogeneity that are "deeper" than lack of correlation between the error and a right-hand side variable (but are still not the same as the "economic" definition of exogeneity).

Precise Statistical Definitions of Exogeneity

With this introduction, we now turn to the general econometric/statistical approach to exogeneity. Start with a system of **M structural equations** (new notation):

$$\mathbf{y}_t'\Gamma + \mathbf{x}_t'\mathbf{B} = \varepsilon_t'$$

where Γ is an M×M matrix (M is the number of **y** variables) and **B** is a K×M matrix (K is the number of **x** variables). What do we mean when we say that $\mathbf{x}_t'$ is exogenous? The **first concept** is strict exogeneity: $\mathbf{x}_t'$ is **strictly exogenous** if all elements of $\mathbf{x}_t'$ are **independent** of all the error terms in the vector ε_s', for all t from 1 to T and all s from 1 to T. This is denoted:

$$\mathbf{x}_t' \perp\!\!\!\perp \varepsilon_s' \quad \text{for all } t, s = 1, 2, \ldots T$$

where "$\perp\!\!\!\perp$" denotes statistical independence.

For time series and panel data models this may be too strong. In particular, we may want to **use past values of y as exogenous variables**. In this notation such variables are part of the **x** variables, not the **y** variables. A **weaker condition** that lets us use past values of y as x variables is **predetermined**: $\mathbf{x}_t'$ is independent only of current and future values of ε_s':

$$\mathbf{x}_t' \perp\!\!\!\perp \varepsilon_{t+s}' \quad \text{for all } t = 1, 2, \ldots T, \text{ for all } s \geq 0$$

This allows one to use past values of **y**'s as **x** variables.

It turns out that **predeterminedness is not very useful** without specifying more about the equation that the **x** variables (including lagged values of **y**) are in. **For example**, consider a very simple model:

$$y_t = \beta x_t + \varepsilon_{1t}$$

$$x_t = \delta_1 x_{t-1} + \delta_2 y_{t-1} + \varepsilon_{2t}$$

$$\text{where } E[\varepsilon_t \varepsilon_t'] = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \; E[\varepsilon_t \varepsilon_s'] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ if } s \neq t.$$

In this model **x_t is not predetermined** in the first equation if $\sigma_{12} \neq 0$ (why?), so OLS estimation of the first equation will be biased (and inconsistent).

Next, consider the expectation of **y_t** conditional on **x_t** and past values of **x_t** and **y_t**.

$$E[y_t \mid x_t, y_{t-1}, x_{t-1}, \ldots] = \beta x_t + E[\varepsilon_{1t} \mid x_t, y_{t-1}, x_{t-1}, \ldots]$$

$$= \beta x_t + E[\varepsilon_{1t} | \varepsilon_{2t}] \quad (\varepsilon\text{'s independent over time})$$

$$= \beta x_t + (\sigma_{12}/\sigma_{22})\varepsilon_{2t} \quad (\text{check any statistics textbook})$$

$$= \beta x_t + (\sigma_{12}/\sigma_{22})(x_t - \delta_1 x_{t-1} - \delta_2 y_{t-1}).$$

This implies that we can write:

$$y_t = (\beta + \sigma_{12}/\sigma_{22})x_t - \delta_1\sigma_{12}/\sigma_{22}x_{t-1} - \delta_2\sigma_{12}/\sigma_{22}y_{t-1} + v_t$$

where v_t is uncorrelated with x_t.

So **rewriting the equation for y_t** so that it is a function not only of x_t but also of x_{t-1} and y_{t-1}, gives an equation where **x_t is predetermined** since it is no longer correlated with the error term. Note that in that equation the parameter on x_t is not β, but $(\beta + \sigma_{12}/\sigma_{22})$, and the error is also different. The **intuition** is: **correlation between the error term and some variable x depends on the other variables in the model**.

For times series analysis economists want something like predeterminedness that clarifies exactly what models or equations are under consideration. This leads to **weak exogeneity**, a more abstract concept.

To define weak exogeneity, first note that the **joint distribution** of the variables in $\mathbf{y}_t'$ and $\mathbf{x}_t'$ ***conditional on past values of x′ and y′*** can be expressed as the product of the joint distribution of $\mathbf{y}_t'$ conditional on $\mathbf{x}_t'$ and past values of $\mathbf{x}_t'$ and $\mathbf{y}_t'$ and the joint distribution of $\mathbf{x}_t'$ conditional on past values of $\mathbf{x}_t'$ and $\mathbf{y}_t'$. That is, **regardless of the exogeneity of** $\mathbf{x}_t'$ **we can write**:

$$f_t(\mathbf{x}_t', \mathbf{y}_t' \mid \Omega_{t-1};\theta) = f_t^y(\mathbf{y}_t' \mid \mathbf{x}_t', \Omega_{t-1};\theta)\times f_t^x(\mathbf{x}_t' \mid \Omega_{t-1};\theta)$$

where Ω_{t-1} represents past values of x and y from time 1 to t-1, and θ is the vector of parameters that determine the joint distribution of **x** and **y** conditional on Ω_{t-1}.

We are now ready to define weak exogeneity. Suppose we can divide θ into two components: $\theta = (\theta_1, \theta_2)$. The variables $\mathbf{x}_t'$ are **weakly exogenous for estimation of** θ_1 if we can rewrite the above conditional density function as:

$$f_t(\mathbf{x}_t', \mathbf{y}_t' \mid \Omega_{t-1};\theta) = f_t^y(\mathbf{y}_t' \mid \mathbf{x}_t', \Omega_{t-1};\theta_1) \times f_t^x(\mathbf{x}_t' \mid \Omega_{t-1};\theta_2)$$

This is pretty abstract, but the **intuition** is that **$\mathbf{y}_t'$ has no "feedback" into the current** (marginal) **distribution of $\mathbf{x}_t'$** over and above what past values of y′ and x′ have.

The key things to remember are:

1. Weak exogeneity is less restrictive compared to strict exogeneity; you can use past values of y as regressors.

2. Weak exogeneity is relevant only for models that have some kind of time series element to them.

3. Predeterminedness is **not** weak exogeneity. **Predeterminedness refers to a specific equation in a model**, while **weak exogeneity refers to the joint distribution of all variables in all equations of the model** and all the parameters and equations underlying that joint distribution.

Granger (Non)Causality

Granger causality is less abstract. In fact, the real concept of interest is called Granger noncausality. Recall that $\mathbf{x}_t'$ is **weakly exogenous** if $\mathbf{y}_t'$ has no explanatory power for $\mathbf{x}_t'$, but it is **still possible for past values of y′** (i.e. $\mathbf{y}_{t-1}'$, $\mathbf{y}_{t-2}'$, etc.) to play a role in determining $\mathbf{x}_t'$ even after conditioning on past values of **x′**. **Granger noncausality rules out**

any role for past values of $\mathbf{y}'$ after accounting for the predictive power of past values of the $\mathbf{x}$ variables.

The definition is: Past values of $\mathbf{y}$ (denoted by $\mathbf{y}_{t-1}$) do not Granger cause $\mathbf{x}_t'$ (current values of $\mathbf{x}$) if and only if:

$$f_t^x(\mathbf{x}_t' \mid \Omega_{t-1}) = f_t^x(\mathbf{x}_t' \mid \mathbf{x}_{t-1})$$

where $\mathbf{x}_{t-1}$ represents all past values of x variables. Thus past values of $\mathbf{y}$ have no predictive power for current values of $\mathbf{x}$ once past values of $\mathbf{x}$ have been used.

Granger noncausality is concerned with the ability to forecast future values of $\mathbf{x}$ and $\mathbf{y}$ variables. If it holds past values of $\mathbf{y}$ do not help forecast current and future $\mathbf{x}$.

Three final comments:

1. Granger noncausality and weak exogeneity are not the same thing. You can have one without the other and neither implies the other. Going back to the 2-equation model on p.21, y does not Granger cause x if and only if $\delta_2 = 0$. In contrast, x is weakly exogenous with respect to estimating β if and only if $\sigma_{12} = 0$ (Davidson and MacKinnon, pp.628-9).

2. Granger noncausality is used for forecasting, and weak exogeneity is used to estimate structural models.

3. For specific parameters, if $\mathbf{x}_t'$ is weakly exogenous ***and*** past values of $\mathbf{y}$ do not Granger cause $\mathbf{x}_t'$, then we say that $\mathbf{x}_t'$ is strongly exogenous.

ARDL Material

Static, Distributed Lag (DL) & Autoregressive DL Models

Consider a model for Y_t that includes additional variables, the ***vector*** Z_t, that does **not** includes lagged values of Y_t:

$$Y_t = \alpha + Z_t'\beta + e_t$$

This **static model** is equation (14.54) with $X_t = (1, Z_t')'$. This shows how Y_t and Z_t "co-move", but it is **not clear what the causal relationship is**.

Causality is clearer if Y_t is determined by **past values of $\boldsymbol{Z_t}$**:

$$Y_t = \alpha + Z_{t-1}'\beta_1 + Z_{t-2}'\beta_2 + \ldots + Z_{t-q}'\beta_q + e_t$$

The **cumulative impact** of Z_t on Y_t is the sum of the β coefficients. This is the **distributed lag (DL)** model.

A more flexible model adds **lagged values of $\boldsymbol{Y_t}$**:

$$Y_t = \alpha_0 + \alpha_1 Y_{t\text{-}1} + \ldots + \alpha_p Y_{t\text{-p}} + Z_{t-1}'\beta_1 + Z_{t-2}'\beta_2 + \ldots + Z_{t-q}'\beta_q + e_t$$

This is the **autoregressive distributed lag (AR-DL)** model. Both the DL and the AR-DL models are examples of equation (14.54). Hansen says that an AR-DL model with p and q "sufficiently large" will have e_t that is "approximately white noise" and so the model "can be interpreted as dynamically well specified" and "conventional standard error methods can be used." The "long-run multiplier" is:

$$\frac{\beta_1 + \ldots + \beta_q}{1 - \alpha_1 - \ldots - \alpha_p}$$

VII. Tests for Serial Correlation in Regression Models (14.45)

Consider an AR-DL model w/ possibly omitted serial correlation:

$$Y_t = \alpha_0 + \alpha_1 Y_{t\text{-}1} + \ldots + \alpha_p Y_{t\text{-p}} + Z_{t\text{-}1}'\beta_1 + \ldots + Z_{t\text{-}q}'\beta_q + u_t \quad (14.61)$$

If u_t is serially correlated, it can be modeled as:

$$u_t = \theta_1 u_{t\text{-}1} + \ldots + \theta_r u_{t\text{-}r} + e_t \qquad (14.62)$$

The hypotheses are:

$$H_0\text{: } \theta_1 = \ldots = \theta_r = 0$$

$$H_1\text{: } \theta_j \neq 0 \text{ for some } j \geq 1$$

There are **two ways to test H₀**. First, estimate (14.61) and (14.62) sequentially by OLS. This is messy and so "is not recommended". Second, combine the equations into a single model. Using lag operators, rewrite them as:

$$\alpha(L)Y_t = \alpha_0 + \beta(L)Z_{t-1} + u_t \quad \text{and} \quad \theta(L)u_t = e_t$$

Then apply $\theta(L)$ to the first equation:

$$\theta(L)\alpha(L)Y_t = \theta(L)\alpha_0 + \theta(L)\beta(L)Z_{t-1} + \theta(L)u_t$$

where $\theta(L)\alpha(L)$ and $\theta(L)\beta(L)$ yield $p + r$ and $q + r$ order polynomials. If H_0 holds, they are p and q order polynomials. To check this, estimate the AR-DL model with $p + r$ and $q + r$ lags, and test that the final r lags of Y_t and Z_t equal 0.