# Regression Models with Lagged Dependent Variables and ARMA models

L. Magee

revised January 21, 2013

## 1 Preliminaries

### 1.1 Time Series Variables and Dynamic Models

For a time series variable $y_t$, the observations usually are indexed by a $t$ subscript instead of $i$. Unless stated otherwise, we assume that $y_t$ is observed at each period $t = 1, \ldots, n$, and these periods are evenly spaced over time, e.g. years, months, or quarters. $y_t$ can be a flow variable (e.g. GDP, trading volume), or a stock variable (e.g. capital stock) or a price or interest rate. For stock variables or prices, it can be important how they are defined, since they may vary within the period. A price at period $t$ might be defined as the price in the middle of the period, at the end of the period, the average price over the period, etc.

A model that describes how $y_t$ evolves over time is called a *time series process*, and a regression model that has terms from different time periods entering in the same equation is a *dynamic model.* An example of a dynamic model is:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + u_t$$

Models with time series variables usually are dynamic models, but not necessarily. You might have

$$y_t = \gamma_0 + \gamma_1 x_t + u_t$$

where $u_t$ is distributed independently of its past values. This is not a dynamic model, because there is nothing in it that links the different time periods.

### 1.2 Lag and First Difference Operators

The *lag operator* $L$ is defined as:

$$Ly_t = y_{t-1}$$

$L^a$ for some positive integer $a$ means lagging $y_t$ by $a$ periods. If $a = 2$ then

$$L^2 y_t = LLy_t = L(Ly_t) = Ly_{t-1} = y_{t-2}$$

*Lag polynomials* are notated as $A(L)$, $B(L)$, etc.. An example is

$$A(L) = 1 - .4L$$

Then

$$A(L)y_t = y_t - .4y_{t-1}$$

Often a lag polynomials can be inverted. Let $A(L) = 1 - \rho L$. If $|\rho|$ is less than one, then $A(L)^{-1}$ expands like a geometric series,

$$A(L)^{-1} = (1 - \rho L)^{-1} = 1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \rho^4 L^4 + \ldots$$

This expansion is used to obtain equation (2) in section 2.1.1.

The first difference, or change in $y_t$ compared to the previous period, will be denoted by the first difference operator $D$, where:

$$Dy_t = y_t - y_{t-1}$$

(The symbol $\Delta$ is used later for describing the change in $y$ due to a change in $x$.)

Note that $D^2 y_t$, the second difference of $y_t$, does <u>not</u> equal $y_t - y_{t-2}$:

$$D^2 y_t = DDy_t = D(Dy_t) = D(y_t - y_{t-1}) = Dy_t - Dy_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

### 1.3 White Noise

In dynamic models, the effects of the error terms are felt across time and across variables. To understand what the model is predicting, and to figure out how to estimate its coefficients and variances, models are commonly specified with error terms having no autocorrelation or heteroskedasticity. These are referred to as *white noise* errors or disturbances. Following Greene's textbook, we will represent a white noise variable as $u_t$. (In many other sources it is represented by $\epsilon_t$.) So the term “white noise” is a short way to say that

$$\begin{aligned} Eu_t &= 0\ ,\ \mathrm{Var}(u_t) = \sigma^2\ ,\ \text{for all } t\ ,\ \text{and} \\ Eu_t u_s &= 0\ ,\ \text{for all } t, s\ , t \neq s \end{aligned}$$

## 1.4 Stationarity

Stationarity usually means *covariance stationarity*, where the expected value, variance, and autocovariances of $y_t$ do not depend on $t$, that is, they do not change over time,

$$Ey_t, \mathrm{Var}(y_t), \mathrm{Cov}(y_t, y_{t-s}) \text{ are not functions of } t$$

In a model, the phrase “$y_t$ is stationary” can mean “$y_t$ is assumed to follow a covariance stationary time series process.” If $y_t$ is an observed time series, then “$y_t$ is stationary” is a (possibly overconfident) way of saying either “on the basis of some testing procedure, we cannot reject the null hypothesis that $y_t$ was generated by a covariance stationary process” or “we can reject the null hypothesis that $y_t$ was generated by a non-stationary process.”

White noise is stationary, but stationarity does not imply white noise. An autocorrelated process can be stationary but is not white noise.

## 1.5 Examples of nonstationary processes

(i) An example of a nonstationary process is

$$y_t = \alpha + \beta t + u_t$$

where $u_t$ is white noise. The mean of $y_t$, $\alpha + \beta t$, is a function of $t$. This process is called *trend stationary*, because apart from the trend, the rest of the process is stationary. That is, once $\beta t$ is accounted for by moving it to the left-hand side, the rest is stationary: $y_t - \beta t = \alpha + u_t$.

(ii) Another example of a nonstationary process is a *random walk*,

$$y_t = y_{t-1} + u_t$$

where $u_t$ is white noise. Taking the variance of both sides,

$$\begin{aligned}\text{Var}(y_t) &= \text{Var}(y_{t-1} + u_t) \\ \text{Var}(y_t) &= \text{Var}(y_{t-1}) + \sigma^2\end{aligned}$$

Unless $\sigma^2 = 0$, the variance of this process increases with $t$, hence must depend on $t$ and is not stationary.

# 2 Time Series Models

## 2.1 ARMA models

An important time series model in statistics is the *autoregressive moving average* (ARMA$(p, q)$) model

$$y_t = \alpha + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + u_t + \gamma_1 u_{t-1} + \ldots + \gamma_q u_{t-q}$$

where $u_t$ is white noise. Letting $A(L) = 1 - \rho_1 L - \ldots - \rho_p L^p$ and $B(L) = 1 + \gamma_1 L + \ldots + \gamma_q L^q$, then the ARMA$(p, q)$ model can be written compactly as

$$A(L)y_t = \alpha + B(L)u_t$$

The $A(L)y_t$ term is the *autoregressive* part and the $B(L)u_t$ term is the *moving average* part.

An ARMA$(p, 0)$ process is more simply called an AR$(p)$ process,

$$A(L)y_t = \alpha + u_t$$

This is a regression model where $y_t$ is regressed on its own lags.

Similarly, an ARMA$(0, q)$ process is called an MA$(q)$ process,

$$y_t = \alpha + B(L)u_t$$

which expresses $y_t$ as a weighted average of the current and past $q$ disturbances.

Some generalizations of ARMA models include

(i) the *autoregressive integrated moving average* (ARIMA($p, d, q$)) model

$$D^d y_t = \alpha + \rho_1 D^d y_{t-1} + \ldots + \rho_p D^d y_{t-p} + u_t + \gamma_1 u_{t-1} + \ldots + \gamma_q u_{t-q}$$

which is an ARMA model applied to the $d^{th}$ difference of $y_t$, and

(ii) the ARMAX model

$$y_t = \alpha + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + x_t'\beta + u_t + \gamma_1 u_{t-1} + \ldots + \gamma_q u_{t-q}$$

which augments the ARMA model with $k$ other regressor variables through a $k \times 1$ vector $x_t$. The inclusion of $x_t$ makes this model look more like a typical econometric model with lagged $y_t$ regressors, although MA errors are not used very often in econometrics.

If $x_t$ includes lags, and the MA aspect of the errors is removed, then we have a *dynamically complete* regression model, discussed later in this section.

### 2.1.1 AR and MA representations of the same process

A dynamic model can be expressed in different ways. A simple example is the AR(1) model, which can be expressed as an MA($\infty$) as follows, if $|\rho| < 1$.

$$\begin{aligned}
y_t &= \gamma + \rho y_{t-1} + u_t && (1)\\
y_t - \rho y_{t-1} &= \gamma + u_t \\
A(L) y_t &= \gamma + u_t, \quad \text{where} \quad A(L) = 1 - \rho L \\
y_t &= A(L)^{-1}\gamma + A(L)^{-1} u_t, \quad \text{if } A(L) \text{ is invertible} \\
y_t &= (1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \rho^4 L^4 + \ldots)(\gamma + u_t) \\
y_t &= (1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \rho^4 L^4 + \ldots)(\gamma + u_t) \\
y_t &= \left(\frac{\gamma}{1-\rho}\right) + (u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \rho^4 u_{t-4} + \ldots) && (2)
\end{aligned}$$

Suppose that $u_t$ is distributed independently of its past values. The dynamics in model (1) are captured by the $y_{t-1}$ regressor. But it is the same process as model (2), where the dynamics are captured by an series of lagged disturbance terms multiplied by coefficients that are gradually decreasing in magnitude. This is an example of a model that has both an *autoregressive representation* (1), convenient for estimating the coefficients and for prediction, and a *moving average representation* (2), which often is more useful for interpreting the effects of shocks on future values. For example, $u_{t-s}$ can be thought of as a "shock" to the process that occurred $s$ periods before time $t$. Its coefficient, $\rho_s$, gives the per-unit effect of that shock on the $y$ value $s$ periods later, at time $t$.

Note that $A(L)^{-1}\gamma = \frac{\gamma}{1-\rho}$. This follows from the fact that the lag of a constant (over time) is that same constant: $L^a\gamma = \gamma$ for all $a$. Then

$$\begin{aligned} A(L)^{-1}\gamma &= (1 + \rho L + \rho^2 L^2 + \rho^3 L^3 + \rho^4 L^4 + \dots)\gamma \\ &= \gamma + \rho\gamma + \rho^2\gamma + \rho^3\gamma + \rho^4\gamma + \dots \\ &= \gamma(1 + \rho + \rho^2 + \rho^3 + \rho^4 + \dots) \\ &= \frac{\gamma}{1-\rho} \end{aligned}$$

### 2.1.2 Explosive series and unit roots

If $|\rho| > 1$, the magnitudes of the coefficients on the lag operators get larger as the lags go "further back". Roughly speaking, that would mean that what happened 100 periods ago would have a bigger effect on the present than what happened last period. This is an *explosive* time series process, which sounds kind of exciting, but is not considered very useful in time series econometrics. There is a great deal of interest in the borderline case where $|\rho| = 1$, specifically when $\rho = 1$ (not so much $\rho = -1$). This value is associated with *unit root* time series, which will be dealt with later.

## 2.2 Dynamically complete models

*Dynamically complete* models are time series regression models with no autocorrelation in the errors. Let $x_t$ and $z_t$ be two time series explanatory variables, then such a model is

$$y_t = \alpha + \rho_1 y_{t-1} + \dots + \rho_p y_{t-p} + \beta_0 x_t + \dots + \beta_a x_{t-a} + \gamma_0 z_t + \dots + \gamma_b z_{t-b} + u_t \tag{3}$$

which can be written more compactly as

$$A(L)y_t = \alpha + B(L)x_t + C(L)z_t + u_t$$

where $u_t$ is white noise. There can be more right-hand side variables.

## 2.3 Models with autocorrelated errors vs. dynamically complete models

Some of the above models have an MA component in the error terms, and therefore have autocorrelated errors, while others have white noise errors. Suppose we have specified a model with white noise errors, but find evidence of autocorrelation in the residuals of the fitted model. (Tests for autocorrelation are discussed in section 4.2.2.) There are two main ways to adjust the model to deal with this. One is to model the autocorrelation in the errors, and the other is to include more lagged regressors until there no longer is evidence of such autocorrelation. This second approach (making the model *dynamically complete* (Wooldridge (2009, pp.396-9))) has become the more popular one. In this section it is shown how these two approaches are related, and why the second approach has become more popular.

### 2.3.1 Autoregressive errors

Suppose the model is

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{4}$$

and autocorrelation is suspected in the error term $\epsilon_t$. Of the two ways of modeling autocorrelation that we have seen, AR and MA processes, AR is by far the more common. Next, we will argue why AR errors are more commonly used than MA errors. As usual, let $u_t$ represent white noise. If the model has mean-zero AR($p$) errors, then

$$\begin{aligned}
\epsilon_t &= \rho_1\epsilon_{t-1} + \ldots + \rho_p\epsilon_{t-p} + u_t \text{ , or equivalently} \\
A(L)\epsilon_t &= u_t \text{ , where } A(L) = 1 - \rho_1 L - \ldots - \rho_p L^p \text{ , or} \\
\epsilon_t &= A(L)^{-1}u_t
\end{aligned}$$

Let $p = 1$, giving AR(1) or first-order autoregressive errors. It has been the most commonly-employed version, and will simplify notation for our purposes. Then (4) becomes

$$y_t = \alpha + \beta x_t + (1 - \rho_1 L)^{-1}u_t \tag{5}$$

This AR error process implies a gradual decrease in the error autocorrelations as the lag increases. $E\epsilon_t\epsilon_{t-1} = E(\rho_1\epsilon_{t-1} + u_t)\epsilon_{t-1} = \rho_1\text{Var}(\epsilon_t)$, assuming $|\rho_1| < 1$ and therefore that the process is stationary. Continuing this approach would show that $E\epsilon_t\epsilon_{t-s} = \rho_1^s \text{Var}(\epsilon_t)$, and therefore

$$\text{Corr}(\epsilon_t\epsilon_{t-s}) = \frac{\text{Cov}(\epsilon_t\epsilon_{t-s})}{\sqrt{\text{Var}(\epsilon_t)\text{Var}(\epsilon_{t-s})}} = \frac{\rho_1^s \text{Var}(\epsilon_t)}{\sqrt{\text{Var}(\epsilon_t)\text{Var}(\epsilon_t)}} = \rho_1^s$$

If the residual autocorrelations die out gradually rather than suddenly, as would be implied by an MA($q$) error process, then AR errors likely provide a better description of the autocorrelation pattern.

***Estimation with AR errors***

Both AR and MA errors can be estimated efficiently by Generalized Least Squares when there are no lagged $y$ regressors. The GLS estimator of the regression coefficients when there are AR(1) errors (e.g. GLS estimation of $\alpha$ and $\beta$ in (5)) is very similar to the following "quasi-differencing" procedure. First, estimate $\rho_1$. Many estimators have been proposed, a simple one being the sample correlation between the OLS residual $e_t$ and its first lag $e_{t-1}$. Then take the approximate model that results from substituting $\hat{\rho}_1$ for $\rho_1$ in (5) and multiply through by $(1 - \hat{\rho}_1 L)$,

$$\begin{aligned} y_t &= \alpha + \beta x_t + (1 - \hat{\rho}_1 L)^{-1} u_t \\ (1 - \hat{\rho}_1 L) y_t &= (1 - \hat{\rho}_1 L)\alpha + \beta(1 - \hat{\rho}_1 L) x_t + u_t \\ \tilde{y}_t &= \alpha^* + \beta \tilde{x}_t + u_t \end{aligned}$$

where $\tilde{y}_t = y_t - \hat{\rho}_1 y_{t-1}$ and $\tilde{x}_t = x_t - \hat{\rho}_1 x_{t-1}$ are *quasi-differenced* variables. OLS estimation of $\alpha^*$ and $\beta$ in the last equation is very similar to the GLS estimator. It is known as the *Cochrane-Orcutt estimator* and has been used since the 1940s.

***Disadvantages of AR errors***

In current practice, models with AR errors largely have been replaced by models with more lagged explanatory variables and white noise errors. One disadvantage of AR errors is that the dynamic effects are difficult to interpret. When the dynamics are only in observed variables, a more direct description of "the effect of $x$ on $y$" over time is possible.

Another disadvantage of AR errors arises when there are lagged $y$ regressors. Consider the model

$$y_t = \alpha + \gamma y_{t-1} + \beta x_t + \epsilon_t \ , \text{ where } \epsilon_t = \rho\epsilon_{t-1} + u_t \tag{6}$$

$y_{t-1}$ is an endogenous regressor. The model itself implies a correlation between $y_t$ and $\epsilon_t$, and therefore between $y_{t-1}$ and $\epsilon_{t-1}$. Unless $\rho = 0$, there must then be a correlation between the regressor $y_{t-1}$ and the error term $\epsilon_t$, since $y_{t-1} \Rightarrow \epsilon_{t-1} \Rightarrow \epsilon_t$. Instrumental variable estimators have been proposed. Unlike many other endogenenous-regressor problems, however, there is another way out, which is described next.

***From AR errors to dynamically complete models***

The once-popular time series model with AR errors has been largely replaced by dynamically complete models, which have more lags and no autocorrelation in the errors. The reasons for this will be illustrated here by converting from one to the other and pointing out the advantages of the latter.

First, rewrite model (6) with the autocorrelation removed by multiplying through by $(1-\rho L)$,

$$\begin{aligned} y_t &= \alpha + \gamma y_{t-1} + \beta x_t + (1-\rho L)^{-1} u_t \\ (1-\rho L) y_t &= (1-\rho L)\alpha + \gamma (1-\rho L) y_{t-1} + \beta (1-\rho L) x_t + u_t \\ y_t &= \alpha^* + \rho y_{t-1} + \gamma y_{t-1} - \gamma \rho y_{t-2} + \beta x_t - \beta \rho x_{t-1} + u_t \end{aligned} \tag{7}$$

Model (7) has new lagged regressors which replace the dynamics that were in the error term of (6). Simplifying the notation for the regression coefficients, (7) can be written as

$$y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 x_t + \theta_4 x_{t-1} + u_t \tag{8}$$

Eliminating the autocorrelation in the errors has solved the endogenous regressor problem, although comparing (7) to (8) reveals a restriction which can be written as $\theta_1\theta_3\theta_4 = -\theta_4^2 + \theta_2\theta_3^2$. This awkward-looking nonlinear restriction arises from having multiplied through the original model by the "common factor" $(1-\rho L)$, and is referred to as a *common factor restriction.* In current practice, it is typically not imposed. Researchers' starting point is a model more like (8) than (6). Having estimated (8), it is not natural to ask whether $\theta_1\theta_3\theta_4 = -\theta_4^2 + \theta_2\theta_3^2$. It is more natural to ask if $\theta_2 = 0$, or $\theta_4 = 0$, or whether $y_{t-3}$ and/or $x_{t-3}$ should be included as regressors, or whether the error term really is white noise. The goal is to specify a dynamically complete model without including lags unnecessarily.

## 2.4 Means and Variances

### 2.4.1 Conditional and unconditional means

If in Section 2.1.1, we assume that $Eu_t = 0$ for all $t$ in model (1), then (2) makes it easy to see that $Ey_t = \frac{\gamma}{1-\rho}$. This may seem to contradict (1), which says that $Ey_t = \gamma + \rho y_{t-1}$. This apparent contradiction arises because the regression model (1) on the one hand shows the expected value of $y_t$ *conditional on* $y_{t-1}$. On the other hand, (2) does not have any regressor variables and has a zero-mean error term, so its constant term represents the *unconditional mean* of $y_t$. Summarizing with more careful expectation notation,

$$\begin{aligned} Ey_t &= \frac{\gamma}{1-\rho} \quad \text{(unconditional mean)} \\ E(y_t|y_{t-1}) &= \gamma + \rho y_{t-1} \quad \text{(conditional mean)} \end{aligned}$$

The unconditional mean does not depend on $t$. This is one of the necessary conditions for a process to be stationary. The unconditional mean of $y_t$ can be derived by using the assumed stationarity of $y_t$ to justify the restriction $Ey_t = Ey_{t-1} = y^*$, say. Take the expectation of both sides of the above conditional mean expression with respect to both $y_t$ and $y_{t-1}$ and then solve for $y*$.

$$\begin{aligned} E_{y_{t-1}}E(y_t|y_{t-1}) &= \gamma + \rho E_{y_{t-1}} y_{t-1} \\ E(y_t) &= \gamma + \rho y* \\ y* &= \gamma + \rho y* \\ y* &= \frac{\gamma}{1-\rho} \end{aligned}$$

### 2.4.2 Deriving unconditional moments of a stationary AR(1)

Consider $y_t$ where

$$y_t = 6 + .4y_{t-1} + u_t$$

and $u_t$ is white noise. Since it is not specified otherwise, and usually it is not, we assume there are no *starting conditions*. (An example of a starting condition is $y_0 = 0$.) When there are no starting conditions, we are assuming that we observe a segment of a process that started long enough ago to have settled into its steady-state behaviour, in which a variable fluctuates around a constant long-run unconditional mean. The above process can be written as

$$\begin{aligned} (1-.4L)y_t &= 6 + u_t \\ y_t &= (1-.4L)^{-1}6 + (1-.4L)^{-1}u_t \\ &= \frac{6}{1-.4} + u_t + .4u_{t-1} + .4^2u_{t-2} + \ldots \\ &= 10 + \sum_{j=0}^{\infty} .4^j u_{t-j} \end{aligned}$$

From the last expression we see that $Ey_t = 10$. The white-noise assumption leads to variance, covariance and correlation results

$$\begin{aligned}
\text{Var}(y_t) &= E(\sum_{j=0}^{\infty} .4^j u_{t-j})^2 = \sum_{j=0}^{\infty} (.4^j)^2 E u_{t-j}^2 = (\sum_{j=0}^{\infty} (.4^{2j}))\sigma^2 \\
&= (\sum_{j=0}^{\infty} .16^j)\sigma^2 = \frac{\sigma^2}{1-.16} = \frac{\sigma^2}{.84}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(y_t, y_{t-s}) &= E(\sum_{j=0}^{\infty} .4^j u_{t-j})(\sum_{\ell=0}^{\infty} .4^\ell u_{t-\ell-s}) = E(\sum_{j=0}^{\infty} .4^{j+s} u_{t-j-s})(\sum_{\ell=0}^{\infty} .4^\ell u_{t-\ell-s}) \\
&= E\sum_{j=0}^{\infty} (.4^{j+s} u_{t-j-s})(.4^j u_{t-s-j}) = E\sum_{j=0}^{\infty} (.4^{j+s} u_{t-j-s})(.4^j u_{t-s-j}) \\
&= \sum_{j=0}^{\infty} .4^{2j+s} E u_{t-j-s}^2 = .4^s \sigma^2 \sum_{j=0}^{\infty} .4^{2j} = \frac{.4^s \sigma^2}{.84}
\end{aligned}$$

$$\text{Corr}(y_t, y_{t-s}) = \frac{\text{Cov}(y_t, y_{t-s})}{\sqrt{\text{Var}(y_t)\text{Var}(y_{t-s})}} = \frac{\text{Cov}(y_t, y_{t-s})}{\text{Var}(y_t)} = (\frac{.4^s\sigma^2}{.84})/(\frac{\sigma^2}{.84}) = .4^s$$

The covariances are not zero, so $y_t$ is not white noise. But because the mean, variance, and covariances do not change over time, $y_t$ is stationary.

### 2.4.3 Long-run expected values in dynamic regression models

The method of section 2.4.2 can be generalized to obtain the unconditional (on past $y$ values) mean of $y$, conditional on other explanatory variables, when the latter are assigned values that are fixed over time. Consider

$$\begin{aligned}
y_t &= \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \gamma_0 x_t + \gamma_1 x_{t-1} + \ldots + \gamma_r x_{t-r} + u_t \text{ or} \\
A(L) y_t &= \alpha + B(L) x_t + u_t
\end{aligned}$$

where $A(L) = 1 - \rho_1 L - \ldots - \rho_p L^p$ and $B(L) = \gamma_0 + \gamma_1 L + \ldots + \gamma_r L^r$.

The long-run or steady-state value of $y$ when $x = x^*$ at all periods is found by letting the solution be $y^*$, say, then substituting these long-run values in the regression model, setting the disturbance term to zero, and solving for $y^*$, as follows:

$$
\begin{aligned}
&y^* = \alpha + \rho_1 y^* + \rho_2 y^* + \ldots + \rho_p y^* + \gamma_0 x^* + \gamma_1 x^* + \ldots + \gamma_r x^* \\
&y^*(1 - \rho_1 - \rho_2 - \ldots - \rho_p) = \alpha + (\gamma_0 + \gamma_1 + \ldots + \gamma_r) x^* \\
&y^* = \frac{\alpha}{1 - \rho_1 - \rho_2 - \ldots - \rho_p} + \left( \frac{\gamma_0 + \gamma_1 + \ldots + \gamma_r}{1 - \rho_1 - \rho_2 - \ldots - \rho_p} \right) x^*
\end{aligned}
$$

A notational trick with the lag polynomials makes this manipulation easier to write. When a lag polynomial operates on something that is constant over time, the $L$ operator does not change its value, so this operation is equivalent to multiplying by one. In that case we can replace $A(L)$, for example, by $A(1) = 1 - \rho_1 - \ldots - \rho_p$, since $A(L)y^* = (1 - \rho_1 - \rho_2 - \ldots - \rho_p)y^* = A(1)y^*$. With this notation, the above derivation can be written as

$$
\begin{aligned}
A(1)y^* &= \alpha + B(1)x^* \\
y^* &= \frac{\alpha}{A(1)} + \left( \frac{B(1)}{A(1)} \right) x^*
\end{aligned}
$$

This approach generalizes easily to several "$x$" variables. A regression model with $k$ regressor variables and their lags can be written compactly as

$$
A(L)y_t = \alpha + B_1(L)x_{1t} + B_2(L)x_{2t} + \ldots + B_k(L)x_{kt} + u_t
$$

When each regressor variable is fixed at a constant value over time, $x_{jt} = x_j^*, j = 1, \ldots, k$, then the long-run or steady-state mean of $y$ is

$$
E(y) = \frac{\alpha}{A(1)} + \sum_{j=1}^{k} \left( \frac{B_j(1)}{A(1)} \right) x_j^*
$$

We require $A(1) > 0$, that is, $\sum_{\ell=1}^{p} \rho_\ell < 1$, which is similar to the requirement $\rho < 1$ seen earlier in the single-lag model.

# 3 Interpreting the Coefficients of Dynamic Models

Here again is the dynamic model (3) from Section 2.2.

$$
y_t \;=\; \alpha + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + \beta_0 x_t + \ldots + \beta_a x_{t-a} + \gamma_0 z_t + \ldots + \gamma_b z_{t-b} + u_t
$$

The coefficients are not very informative in isolation. For example, the coefficient $\beta_2$ on the regressor $x_{t-2}$ indicates the change in $y_t$ resulting from a one-unit change in $x_{t-2}$, holding the other regressors constant. But the model itself implies that a one-unit change in $x_{t-2}$ would have changed $y_{t-1}$ and $y_{t-2}$ (unless $\beta_1 = \beta_2 = 0$). The holding-all-else constant thought experiment is not very appealing for dynamic models.

A better way to interpret these coefficients is to use them to trace out the effect that a change in $x_t$ would have on $y_t$, $y_{t+1}$, $y_{t+2}$, etc. This can be done recursively as follows. Let $\Delta y$ represent the change in $y$ resulting from a change in $x$ at time $t$ only, denoted as $\Delta x_t$ (as distinct from the first difference operator $D$ introduced earlier). The effects of $\Delta x_t$ on $y_{t+1}$, $y_{t+2}$, etc. are derived recursively from the original model by adjusting the $t$ subscript in to $t+1$, $t+2$, etc. as required:

$$\begin{aligned}
\Delta y_t &= \beta_0 \Delta x_t \\
\Delta y_{t+1} &= \rho_1 \Delta y_t + \beta_1 \Delta x_t \\
&= \rho_1(\beta_0 \Delta x_t) + \beta_1 \Delta x_t \\
&= (\rho_1 \beta_0 + \beta_1)\Delta x_t \\
\Delta y_{t+2} &= \rho_1 \Delta y_{t+1} + \rho_2 \Delta y_t + \beta_2 \Delta x_t \\
&= \rho_1((\rho_1\beta_0 + \beta_1)\Delta x_t) + \rho_2(\beta_0 \Delta x_t) + \beta_2 \Delta x_t \\
&= (\rho_1^2 \beta_0 + \rho_1 \beta_1)\Delta x_t + (\rho_2 \beta_0)\Delta x_t + \beta_2 \Delta x_t \\
&= (\rho_1^2 \beta_0 + \rho_1 \beta_1 + \rho_2 \beta_0 + \beta_2)\Delta x_t
\end{aligned}$$

and so on.

When the process is stationary, then the long run effect of this one-time-only change in $x$ settles to zero, that is, $\lim_{s\to\infty} \frac{\Delta y_{t+s}}{\Delta x_t} = 0$.

Another way to think about the effect of $x$ on $y$ in this model is to trace out the effect of a "permanent" change in $x$ beginning at time $t$ that lasts into the future. That is, let $\Delta x_t = \Delta x_{t+1} = \Delta x_{t+2} = \ldots = \Delta_x$, say. In general, this has a non-zero long run effect on $y$. Applied to the above model, these effects are

$$\begin{aligned}
\Delta y_t &= \beta_0 \Delta_x \\
\Delta y_{t+1} &= \rho_1 \Delta y_t + \beta_1 \Delta_x + \beta_0 \Delta_x \\
&= \rho_1(\beta_0 \Delta_x) + (\beta_1 + \beta_0)\Delta_x \\
&= (\rho_1 \beta_0 + \beta_1 + \beta_0)\Delta_x \\
\Delta y_{t+2} &= \rho_1 \Delta y_{t+1} + \rho_2 \Delta y_t + \beta_2 \Delta_x + \beta_1 \Delta_x + \beta_0 \Delta_x \\
&= \rho_1((\rho_1 \beta_0 + \beta_1 + \beta_0)\Delta_x) + \rho_2(\beta_0 \Delta_x) + \beta_2 \Delta_x + \beta_1 \Delta_x + \beta_0 \Delta_x \\
&= (\rho_1^2 \beta_0 + \rho_1 \beta_0 + \rho_1 \beta_1 + \rho_2 \beta_0 + \beta_2 + \beta_1 + \beta_0)\Delta_x \\
&\vdots \\
\Delta y_{t+\infty} &= \frac{B(1)}{A(1)} = \frac{\beta_0 + \ldots + \beta_a}{1 - \rho_1 - \ldots - \rho_p}
\end{aligned}$$

Numerical applications are not as notationally cumbersome. Suppose a fitted dynamic model is

$$y_t = 5 + .3y_{t-1} + .1y_{t-2} + 4.2x_t + 1.7x_{t-1} - .7z_t + .3z_{t-1} - .2z_{t-2} + e_t$$

The effects on $y$ of a one-time-only change in $x_t$ are

$$\begin{aligned}
\Delta y_t &= 4.2\Delta x_t \\
\Delta y_{t+1} &= .3\Delta y_t + 1.7\Delta x_t \\
&= .3(4.2\Delta x_t) + 1.7\Delta x_t = 2.96\Delta x_t \\
\Delta y_{t+2} &= .3\Delta y_{t+1} + .1\Delta y_t + 0\Delta x_t \\
&= .3(2.96\Delta x_t) + .1(4.2\Delta x_t) + 0 = 1.308\Delta x_t \\
\Delta y_{t+3} &= .3\Delta y_{t+2} + .1\Delta y_{t+1} \\
&= .3(1.308\Delta x_t) + .1(2.96\Delta x_t) = .6884\Delta x_t \\
\Delta y_{t+4} &= (.3 \times .6884 + .1 \times 1.308)\Delta x_t = .33732\Delta x_t \\
&\vdots \\
\Delta y_{t+\infty} &= 0 \times \Delta x_t
\end{aligned}$$

and the effects on $y$ of a permanent change in $x$ are

$$\begin{aligned}
\Delta y_t &= 4.2\Delta_x \\
\Delta y_{t+1} &= .3\Delta y_t + 4.2\Delta x_{t+1} + 1.7\Delta x_t \\
&= .3(4.2\Delta_x) + 4.2\Delta_x + 1.7\Delta_x = 7.16\Delta_x \\
\Delta y_{t+2} &= .3\Delta y_{t+1} + .1\Delta y_t + 4.2\Delta x_{t+2} + 1.7\Delta x_{t+1} + 0\Delta x_t \\
&= .3(7.16\Delta_x) + .1(4.2\Delta_x) + 4.2\Delta_x + 1.7\Delta_x + 0\Delta_x = 8.468\Delta_x \\
&\vdots \\
\Delta y_{t+\infty} &= = \frac{4.2+1.7}{1-.3-.1} = \frac{5.9}{.6} = 9.833
\end{aligned}$$

These permanent-change effects are cumulative sums of the previously-derived one-time change effects. This is reflected in the Stata terminology *simple* and *cumulative IRF*s. (IRFs are *impulse response functions*, which express these effects as a function of the time elapsed after the change in $x$. They usually are associated with VARs and VECMs, which are discussed in the next set of notes.)

# 4 Estimation and Testing

Models like (3) and (8) are estimated by OLS. There is no autocorrelation in $u_t$, so the lagged $y$ regressors are not correlated with $u_t$. As long as $x$ is exogenous and there is no autocorrelation in the errors, then the regressors values at time $t$ and earlier are not correlated with $u_t$.

## 4.1 OLS: biased, but consistent

There is one other detail to consider, though, that does not arise in cross-section data. Let $z_t = [1\ y_{t-1}\ y_{t-2}\ x_t\ x_{t-1}]'$ be the $5 \times 1$ vector of regressor values at time $t$, and $\theta$ is the coefficient vector. Then (8) can be written as

$$y_t = z_t'\theta + u_t$$

and the OLS estimator is

$$\hat{\theta} = (\sum_t z_t z_t')^{-1} \sum_t z_t y_t$$

The usual decomposition of $\hat{\theta}$ is

$$\begin{aligned}
\hat{\theta} &= (\sum_t z_t z_t')^{-1} \sum_t z_t y_t \\
&= (\sum_t z_t z_t')^{-1} \sum_t z_t (z_t'\theta + u_t) \\
&= (\sum_t z_t z_t')^{-1} \sum_t z_t z_t'\theta + (\sum_t z_t z_t')^{-1} \sum_t z_t u_t \\
&= \theta + (\sum_t z_t z_t')^{-1} \sum_t z_t u_t
\end{aligned}$$

Even though we assume that the elements of $z_t$ are not correlated with $u_t$, there will be non-zero correlations between $u_t$ and some elements of $z_{t+s}$, $s > 0$, because the lagged $y$ regressors appearing in $z_{t+s}$ are correlated with $u_t$. This causes a correlation between $z_t u_t$ and some elements in $(\sum_t z_t z_t')^{-1}$, so that

$$E(\sum_t z_t z_t')^{-1} \sum_t z_t u_t \neq 0$$

which implies that $\hat{\theta}$ is biased. Despite this bias, the following argument shows that plim $\hat{\theta} = \theta$, implying that this bias shrinks to zero as $n \to \infty$.

Since $E z_t u_t = 0$, then $E z_t (y_t - z_t'\theta) = 0$. Solving for $\theta$ gives $\theta = (E z_t z_t')^{-1} E z_t y_t$.

Now consider the plim of $\hat{\theta}$.

$$\begin{aligned}
\text{plim } \hat{\theta} &= \text{plim}(\sum_t z_t z_t')^{-1} \sum_t z_t y_t \\
&= (\text{plim}(n^{-1} \sum_t z_t z_t')^{-1})(\text{plim}(n^{-1} \sum_t z_t y_t)) \\
&= (E z_t z_t')^{-1}(E z_t y_t) = \theta
\end{aligned}$$

where the last line follows from applying the LLN separately to each of the two sample means. Therefore $\hat{\theta}$ is a consistent estimator of $\theta$. The fact that $\hat{\theta}$ is biased in finite samples tends to be ignored in practice.

## 4.2 Specification Tests

The two main specification issues are: how many lags to apply to each of the regressor variables, and whether or not the errors are autocorrelated.

### 4.2.1 Choosing the number of lags

In practice, there may be many more explanatory variables than the one '$x$' that appears in (8). Which lags should be included for each of these variables?

One way to streamline this decision-making, given some maximum lag $p$, is to require that all of the lower-order lags be included. That is, if $p = 3$, include all of the variables $x_t$, $x_{t-1}$, $x_{t-2}$ and $x_{t-3}$. For example, suppose the coefficient on $x_{t-1}$ is not significant and you remove this regressor. Even though its removal is 'supported' by the test, the restriction itself usually has little intuitive appeal. It restricts the pattern of the dynamic effects of $x$ on $y$ in a hard-to-describe and possibly nonintuitive way. One exception is when the data are quarterly or monthly and there is some seasonal pattern (e.g. a spike in consumption in December). Then you might specify one or two lags, followed by a gap, and another longer lag (e.g. 12 lags for quarterly data) to capture the stochastic part of this seasonal pattern.

A way to further simplify this decision is to require that the same number of lags be used for each variable. The following procedures can easily be adapted to this restriction.

***Testing up and testing down***

One way to select the maximum lag is to test whether the coefficient of the maximum lag of a given model equals zero. If the test accepts, reduce the lag by one and repeat. If it rejects, stop. Starting with a lot of lags and working down in this way is called *testing-down*.

Another approach is to start with a small number of lags. Add one lag, and if the new lag is not significant, then stop. If it is significant, then keep it, add another one, test it, etc. This is called *testing-up*.

Testing-down and testing-up both are prone to stopping at a bad number of lags due to unavoidable type I and type II errors. They are sensitive to two arbitrary choices: the number of lags you start at and the significance level of the test. Testing-down seems to be the more popular of the two.

***Model selection criteria***

Another approach to lag length selection is to estimate the model at many different lag lengths and compute a *model selection criterion* for each. Then pick the lag length that gave the best value of the criterion. These criteria are functions of the sum of squared residuals and number of regressors, handling the tradeoff between model parsimony and goodness-of-fit. Of the many criteria that have been suggested, the most commonly-used one for lag selection is Akaike's information criterion (AIC):

$$AIC = n \ln(RSS) + 2k$$

where n is sample size, $RSS$ is the sum of squared residuals, and $k$ is the number of unrestricted regression coefficients. A smaller $AIC$ is better.

### 4.2.2 Testing for autocorrelation

The traditional test of the null hypothesis of no autocorrelation in the errors is the Durbin-Watson test. Their original test is not applicable when there is a lagged $y$, but they developed the Durbin-Watson $h$ statistic for that situation. These tests still are used, but have been largely supplanted by the Box-Pierce, Ljung-Box, and Breusch-Godfrey tests.

The *Box-Pierce test statistic* is

$$Q_{BP} = n \sum_{j=1}^{p} \hat{r}_j^2$$

where $\hat{r}_j$ is the estimated $j^{th}$ order autocorrelation in the OLS residuals. The null hypothesis is $H_0 : \rho_1 = \ldots = \rho_p = 0$, where $\rho_j$ is the $j^{th}$ order autocorrelation of the disturbance process.

The *Ljung-Box test statistic* adjusts $Q_{BP}$ to more closely follow its asymptotic null distribution in finite samples.

$$Q_{LB} = n(n+2) \sum_{j=1}^{p} \frac{\hat{r}_j^2}{n-j}$$

The *Breusch-Godfrey test statistic* is $nR^2$, where $R^2$ is from the regression

$$e_t = z_t'\theta + \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_p e_{t-p} + \text{error}$$

where the errors in question are $u_t$ from $y_t = z_t'\theta + u_t$, $e_t$ are the OLS residuals.

For each of these three tests, $n$ is the number of observations used in the regression (which depends on both the number of observations in the original data set and the number of lags). The null hypothesis is $H_0 : \rho_1 = \ldots = \rho_p = 0$, where $\rho_j$ is the $j^{th}$ order autocorrelation of the disturbance process. The asymptotic null distribution of all three test statistics is chi-square with $p$ d.f., and $p$ is chosen by the researcher. The rejection region is in the right tail.

These tests' accept/reject rules are simpler that the Durbin-Watson tests, they allow for more general lagged dependent variable specifications, and they enable testing for autocorrelations beyond the first order.