Chapter 6

Cycles I: Autoregressions and Wold's Chain Rule

We've already considered models with trend and seasonal components. In this chapter we consider a crucial third component, **cycles**. When you think of a "cycle," you probably think of the sort of rigid up-and-down pattern depicted in Figure 6.1. Such cycles can sometimes arise, but cyclical fluctuations in business, finance, economics and government are typically much less rigid. In fact, when we speak of cycles, we have in mind a much more general notion of cyclicality: any sort of stable, mean-reverting dynamics not captured by trends or seasonals.

Cycles, according to our broad interpretation, may display the sort of back-and-forth movement characterized in Figure 6.1, but they need not. All we require is that there be some stable dynamics ("covariance stationary" dynamics, in the jargon that we'll shortly introduce) that link the present to the past, and hence the future to the present. Cycles are present in most of the series that concern us, and it's crucial that we know how to model and forecast them, because their history conveys information regarding their future.

Trend and seasonal dynamics are simple, so we can capture them with simple models. Cyclical dynamics, however, are a bit more complicated, and

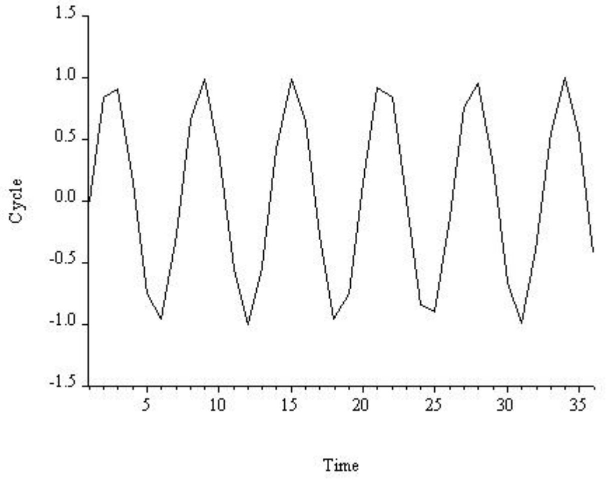


Figure 6.1: A Rigid Cyclical Pattern

consequently the cycle models we need are a bit more involved. We will emphasize autoregressive models.

Let's jump in.

6.1 Characterizing Cycles

Here we introduce methods for characterizing cyclical dynamics in model-free fashion.

6.1.1 Covariance Stationary Time Series

Basic Ideas

A **realization** of a time series is an ordered set,

$$\{..., y_{-2}, y_{-1}, y_0, y_1, y_2, ...\}.$$

Typically the observations are ordered in time – hence the name **time series** – but they don't have to be. We could, for example, examine a spatial series, such as office space rental rates as we move along a line from a point in midtown Manhattan to a point in the New York suburbs thirty miles away. But the most important case, by far, involves observations ordered in time, so that's what we'll stress.

In theory, a time series realization begins in the infinite past and continues into the infinite future. This perspective may seem abstract and of limited practical applicability, but it will be useful in deriving certain very important properties of the models we'll be using shortly. In practice, of course, the data we observe is just a finite subset of a realization, $\{y_1, ..., y_T\}$, called a **sample path**.

Shortly we'll be building models for cyclical time series. If the underlying probabilistic structure of the series were changing over time, we'd be doomed – there would be no way to relate the future to the past, because the laws governing the future would differ from those governing the past. At a minimum we'd like a series' mean and its covariance structure (that is, the covariances between current and past values) to be stable over time, in which case we say that the series is **covariance stationary**. Let's discuss covariance stationarity in greater depth. The first requirement for a series to be covariance stationary is that the mean of the series be stable over time. The mean of the series at time t is $Ey_t = \mu_t$. If the mean is stable over time, as required by covariance stationarity, then we can write $Ey_t = \mu$, for all t. Because the mean is constant over time, there's no need to put a time subscript on it.

The second requirement for a series to be covariance stationary is that its covariance structure be stable over time. Quantifying stability of the covariance structure is a bit tricky, but tremendously important, and we do it using the **autocovariance function**. The autocovariance at displacement τ is just the covariance between y_t and $y_{t-\tau}$. It will of course depend on τ, and it may also depend on t, so in general we write

$$\gamma(t, \tau) = cov(y_t, y_{t-\tau}) = E(y_t - \mu)(y_{t-\tau} - \mu).$$

If the covariance structure is stable over time, as required by covariance stationarity, then the autocovariances depend only on displacement, τ, not on time, t, and we write $\gamma(t, \tau) = \gamma(\tau)$, for all t.

The autocovariance function is important because it provides a basic summary of cyclical dynamics in a covariance stationary series. By examining the autocovariance structure of a series, we learn about its dynamic behavior. We graph and examine the autocovariances as a function of τ. Note that the autocovariance function is symmetric; that is, $\gamma(\tau) = \gamma(-\tau)$, for all τ. Typically, we'll consider only non-negative values of τ. Symmetry reflects the fact that the autocovariance of a covariance stationary series depends only on displacement; it doesn't matter whether we go forward or backward. Note also that $\gamma(0) = cov(y_t, y_t) = var(y_t)$.

There is one more technical requirement of covariance stationarity: we require that the variance of the series – the autocovariance at displacement 0, $\gamma(0)$, be finite. It can be shown that no autocovariance can be larger in absolute value than $\gamma(0)$, so if $\gamma(0) < \infty$, then so too are all the other autocovariances.

It may seem that the requirements for covariance stationarity are quite stringent, which would bode poorly for our models, almost all of which invoke covariance stationarity in one way or another. It is certainly true that many economic, business, financial and government series are not covariance stationary. An upward trend, for example, corresponds to a steadily increasing mean, and seasonality corresponds to means that vary with the season, both of which are violations of covariance stationarity.

But appearances can be deceptive. Although many series are not covariance stationary, it is frequently possible to work with models that give special treatment to nonstationary components such as trend and seasonality, so that the cyclical component that's left over is likely to be covariance stationary. We'll often adopt that strategy. Alternatively, simple transformations often appear to transform nonstationary series to covariance stationarity. For example, many series that are clearly nonstationary in levels appear covariance stationary in growth rates.

In addition, note that although covariance stationarity requires means and covariances to be stable and finite, it places no restrictions on other aspects of the distribution of the series, such as skewness and kurtosis.[1] The upshot is simple: whether we work directly in levels and include special components for the nonstationary elements of our models, or we work on transformed data such as growth rates, the covariance stationarity assumption is not as unrealistic as it may seem.

Recall that the correlation between two random variables x and y is defined by

$$corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}.$$

That is, the correlation is simply the covariance, "normalized," or "standardized," by the product of the standard deviations of x and y. Both the correlation and the covariance are measures of linear association between two random variables. The correlation is often more informative and easily interpreted, however, because the construction of the correlation coefficient guarantees that $corr(x, y) \in [-1, 1]$, whereas the covariance between the same two random variables may take any value. The correlation, moreover, does not depend on the units in which x and y are measured, whereas the covariance does. Thus, for example, if x and y have a covariance of ten million, they're not necessarily very strongly associated, whereas if they have a correlation of .95, it is unambiguously clear that they are very strongly associated.

In light of the superior interpretability of correlations as compared to covariances, we often work with the correlation, rather than the covariance, between y_t and $y_{t-\tau}$. That is, we work with the **autocorrelation function**,

[1] For that reason, covariance stationarity is sometimes called **second-order stationarity** or **weak stationarity**.

$\rho(\tau)$, rather than the autocovariance function, $\gamma(\tau)$. The autocorrelation function is obtained by dividing the autocovariance function by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \tau = 0, 1, 2, ...$$

The formula for the autocorrelation is just the usual correlation formula, specialized to the correlation between y_t and $y_{t-\tau}$. To see why, note that the variance of y_t is $\gamma(0)$, and by covariance stationarity, the variance of y at any other time $y_{t-\tau}$ is also $\gamma(0)$. Thus,

$$\rho(\tau) = \frac{cov(y_t, y_{t-\tau})}{\sqrt{var(y_t)}\sqrt{var(y_{t-\tau})}} = \frac{\gamma(\tau)}{\sqrt{\gamma(0)}\sqrt{\gamma(0)}} = \frac{\gamma(\tau)}{\gamma(0)},$$

as claimed. Note that we always have $\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$, because any series is perfectly correlated with itself. Thus the autocorrelation at displacement 0 isn't of interest; rather, only the autocorrelations *beyond* displacement 0 inform us about a series' dynamic structure.

Finally, the **partial autocorrelation function**, $p(\tau)$, is sometimes useful. $p(\tau)$ is just the coefficient of $y_{t-\tau}$ in a population linear regression of y_t on $y_{t-1}, ..., y_{t-\tau}$.[2] We call such regressions **autoregressions**, because the variable is regressed on lagged values of itself. It's easy to see that the autocorrelations and partial autocorrelations, although related, differ in an important way. The autocorrelations are just the "simple" or "regular" correlations between y_t and $y_{t-\tau}$. The partial autocorrelations, on the other hand, measure the association between y_t and $y_{t-\tau}$ after *controlling* for the effects of y_{t-1} , ..., $y_{t-\tau+1}$; that is, they measure the partial correlation between y_t and $y_{t-\tau}$.

As with the autocorrelations, we often graph the partial autocorrelations

[2] To get a feel for what we mean by "**population regression**," imagine that we have an infinite sample of data at our disposal, so that the parameter estimates in the regression are not contaminated by sampling variation – that is, they're the true population values. The thought experiment just described is a population regression.

as a function of τ and examine their qualitative shape, which we'll do soon. Like the autocorrelation function, the partial autocorrelation function provides a summary of a series' dynamics, but as we'll see, it does so in a different way.[3]

All of the covariance stationary processes that we will study subsequently have autocorrelation and partial autocorrelation functions that approach zero, one way or another, as the displacement gets large. In Figure 6.2 we show an autocorrelation function that displays gradual one-sided damping, and in Figure 6.3 we show a constant autocorrelation function; the latter could not be the autocorrelation function of a stationary process, whose autocorrelation function must eventually decay. The precise decay patterns of autocorrelations and partial autocorrelations of a covariance stationary series, however, depend on the specifics of the series. In Figure 6.4, for example, we show an autocorrelation function that displays damped oscillation – the autocorrelations are positive at first, then become negative for a while, then positive again, and so on, while continuously getting smaller in absolute value. Finally, in Figure 6.5 we show an autocorrelation function that differs in the way it approaches zero – the autocorrelations drop abruptly to zero beyond a certain displacement.

[3] Also in parallel to the autocorrelation function, the partial autocorrelation at displacement 0 is always one and is therefore uninformative and uninteresting. Thus, when we graph the autocorrelation and partial autocorrelation functions, we'll begin at displacement 1 rather than displacement 0.

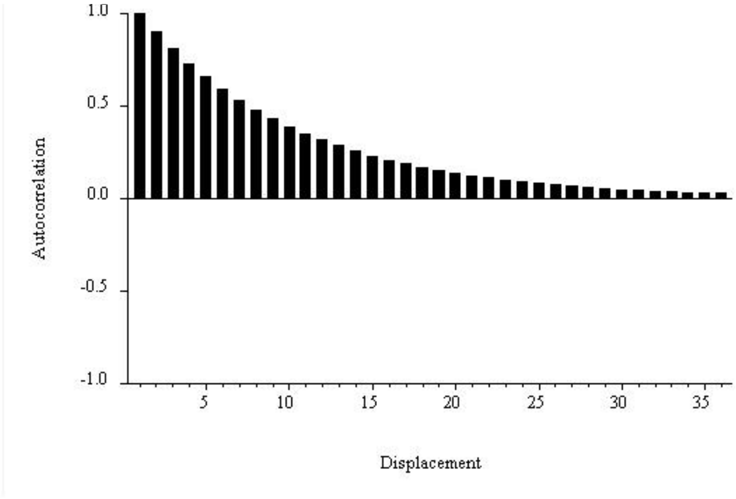


Figure 6.2: Autocorrelation Function: One-sided Gradual Damping

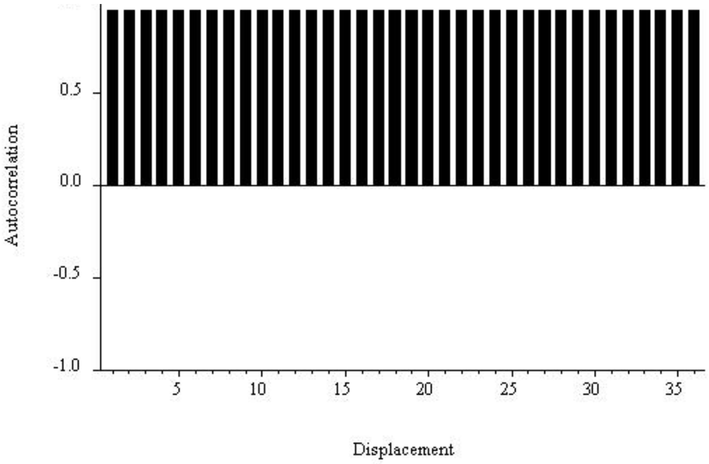


Figure 6.3: Constant Autocorrelation

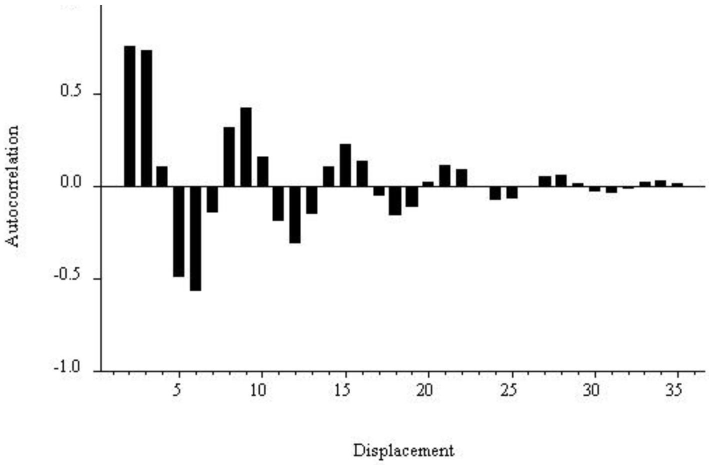


Figure 6.4: Autocorrelation Function: Gradual Damped Oscillation

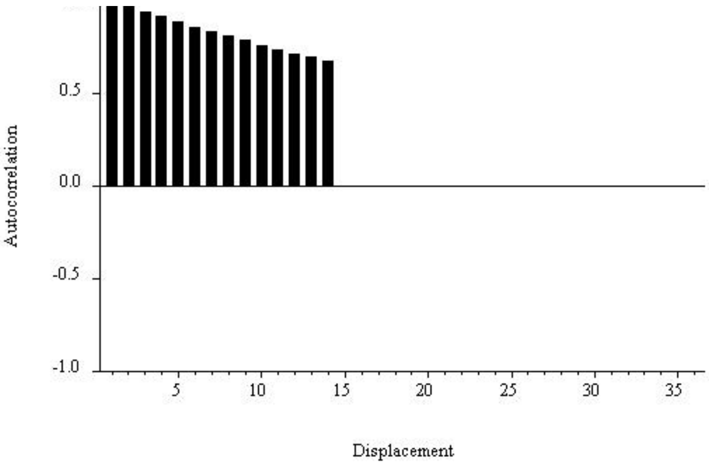


Figure 6.5: Autocorrelation Function: Sharp Cutoff

6.2 White Noise

6.2.1 Basic Ideas

Later in this chapter we'll study the population properties of certain important time series models, or **time series processes**. Before we estimate time series models, we need to understand their population properties, assuming that the postulated model is true. The simplest of all such time series processes is the fundamental building block from which all others are constructed. In fact, it's so important that we introduce it now. We use y to denote the observed series of interest. Suppose that

$$y_t = \varepsilon_t$$

$$\varepsilon_t \sim (0, \sigma^2),$$

where the "shock," ε_t , is uncorrelated over time. We say that ε_t , and hence y_t , is **serially uncorrelated**. Throughout, unless explicitly stated otherwise, we assume that $\sigma^2 < \infty$. Such a process, with zero mean, constant variance, and no serial correlation, is called **zero-mean white noise**, or simply **white noise**.[4] Sometimes for short we write

$$\varepsilon_t \sim WN(0, \sigma^2)$$

and hence

$$y_t \sim WN(0, \sigma^2).$$

Note that, although ε_t and hence y_t are serially uncorrelated, they are not necessarily serially independent, because they are not necessarily normally distributed.[5] If in addition to being serially uncorrelated, y is serially

[4] It's called white noise by analogy with white light, which is composed of all colors of the spectrum, in equal amounts. We can think of white noise as being composed of a wide variety of cycles of differing periodicities, in equal amounts.

[5] Recall that zero correlation implies independence only in the normal case.

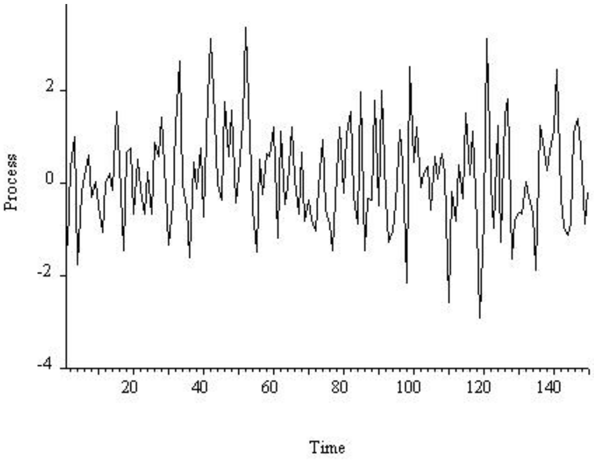


Figure 6.6: Realization of White Noise Process

independent, then we say that y is **independent white noise**.[6] We write

$$y_t \sim iid(0, \sigma^2),$$

and we say that "y is independently and identically distributed with zero mean and constant variance." If y is serially uncorrelated and normally distributed, then it follows that y is also serially independent, and we say that y is **normal white noise**, or **Gaussian white noise**.[7] We write

$$y_t \sim iidN(0, \sigma^2).$$

We read "y is independently and identically distributed as normal, with zero mean and constant variance," or simply "y is Gaussian white noise." In Figure 6.6 we show a sample path of Gaussian white noise, of length $T = 150$, simulated on a computer. There are no patterns of any kind in the series due to the independence over time.

You're already familiar with white noise, although you may not realize

[6] Another name for independent white noise is **strong white noise**, in contrast to standard serially uncorrelated **weak white noise**.

[7] Carl Friedrich Gauss, one of the greatest mathematicians of all time, discovered the normal distribution some 200 years ago; hence the adjective "Gaussian."

it. Recall that the disturbance in a regression model is typically assumed to be white noise of one sort or another. There's a subtle difference here, however. Regression disturbances are not observable, whereas we're working with an observed series. Later, however, we'll see how all of our models for observed series can be used to model unobserved variables such as regression disturbances.

Let's characterize the dynamic stochastic structure of white noise, $y_t \sim WN(0, \sigma^2)$. By construction the unconditional mean of y is $E(y_t) = 0$, and the unconditional variance of y is $var(y_t) = \sigma^2$. Note in particular that the unconditional mean and variance are constant. In fact, the unconditional mean and variance must be constant for any covariance stationary process. The reason is that constancy of the unconditional mean was our first explicit requirement of covariance stationarity, and that constancy of the unconditional variance follows implicitly from the second requirement of covariance stationarity, that the autocovariances depend only on displacement, not on time.[8]

To understand fully the linear dynamic structure of a covariance stationary time series process, we need to compute and examine its mean and its autocovariance function. For white noise, we've already computed the mean and the variance, which is the autocovariance at displacement 0. We have yet to compute the rest of the autocovariance function; fortunately, however, it's very simple. Because white noise is, by definition, uncorrelated over time, all the autocovariances, and hence all the autocorrelations, are zero beyond displacement 0.[9] Formally, then, the autocovariance function for a white noise process is

$$\gamma(\tau) = \begin{cases} \sigma^2, \tau = 0 \\ \\ 0, \tau \geq 1, \end{cases}$$

[8] Recall that $\sigma^2 = \gamma(0)$.

[9] If the autocovariances are all zero, so are the autocorrelations, because the autocorrelations are proportional to the autocovariances.

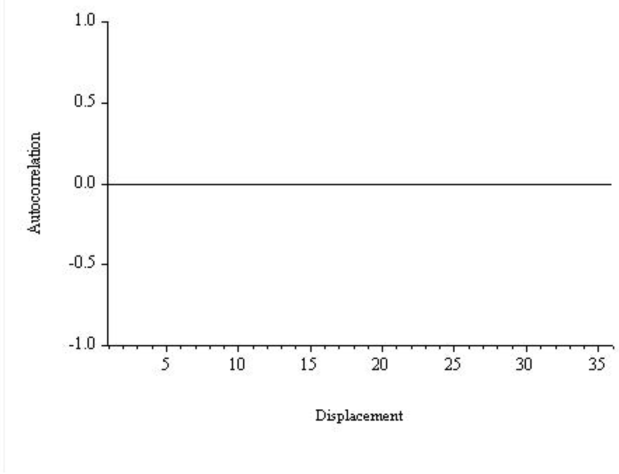


Figure 6.7: White Noise Autocorrelation Function

and the autocorrelation function for a white noise process is

$$\rho(\tau) = \begin{cases} 1, \tau = 0 \\ \\ 0, \tau \geq 1. \end{cases}$$

In Figure 6.7 we plot the white noise autocorrelation function.

Finally, consider the partial autocorrelation function for a white noise series. For the same reason that the autocorrelation at displacement 0 is always one, so too is the partial autocorrelation at displacement 0. For a white noise process, all partial autocorrelations beyond displacement 0 are zero, which again follows from the fact that white noise, by construction, is serially uncorrelated. Population regressions of y_t on y_{t-1} , or on y_{t-1} and y_{t-2} , or on any other lags, produce nothing but zero coefficients, because the process is serially uncorrelated. Formally, the partial autocorrelation function of a white noise process is

$$p(\tau) = \begin{cases} 1, \tau = 0 \\ \\ 0, \tau \geq 1. \end{cases}$$

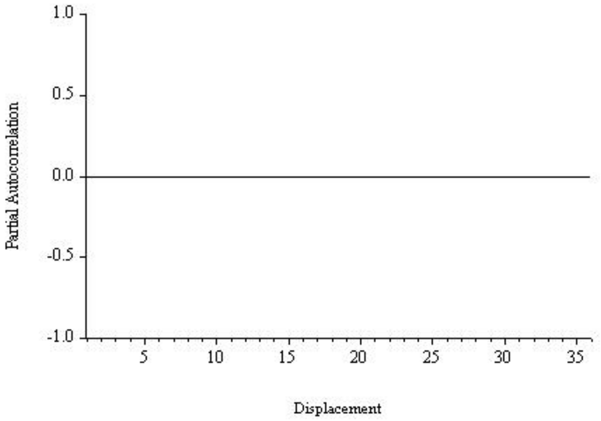


Figure 6.8: White Noise Partial Autocorrelation Function

We show the partial autocorrelation function of a white noise process in Figure 6.8. Again, it's degenerate, and exactly the same as the autocorrelation function!

White noise is very special, indeed degenerate in a sense, as what happens to a white noise series at any time is uncorrelated with anything in the past, and similarly, what happens in the future is uncorrelated with anything in the present or past. But understanding white noise is tremendously important for at least two reasons. First, as already mentioned, processes with much richer dynamics are built up by taking simple transformations of white noise.

Second, the goal of all time series modeling (and 1-step-ahead forecasting) is to reduce the data (or 1-step-ahead forecast errors) to white noise. After all, if such forecast errors aren't white noise, then they're serially correlated, which means that they're forecastable, and if forecast errors are forecastable then the forecast can't be very good. Thus it's important that we understand and be able to recognize white noise.

Thus far we've characterized white noise in terms of its mean, variance, autocorrelation function and partial autocorrelation function. Another characterization of dynamics involves the mean and variance of a process, *conditional* upon its past. In particular, we often gain insight into the dynamics in a process by examining its conditional mean.[10] In fact, throughout our study of time series, we'll be interested in computing and contrasting the **unconditional mean and variance** and the **conditional mean and variance** of various processes of interest. Means and variances, which convey information about location and scale of random variables, are examples of what statisticians call **moments**. For the most part, our comparisons of the conditional and unconditional moment structure of time series processes will focus on means and variances (they're the most important moments), but sometimes we'll be interested in higher-order moments, which are related to properties such as skewness and kurtosis.

For comparing conditional and unconditional means and variances, it will simplify our story to consider independent white noise, $y_t \sim iid(0, \sigma^2)$. By the same arguments as before, the unconditional mean of y is 0 and the unconditional variance is σ^2. Now consider the conditional mean and variance, where the information set Ω_{t-1} upon which we condition contains either the past history of the observed series, $\Omega_{t-1} = y_{t-1}, y_{t-2}, ...$, or the past history of the shocks, $\Omega_{t-1} = \varepsilon_{t-1}, \varepsilon_{t-2} ....$ (They're the same in the white noise case.) In contrast to the unconditional mean and variance, which must be constant by covariance stationarity, the conditional mean and variance need not be constant, and in general we'd expect them *not* to be constant. The unconditionally expected growth of laptop computer sales next quarter may be ten percent, but expected sales growth may be much higher, *conditional* upon knowledge that sales grew this quarter by twenty percent. For the independent white noise process, the conditional mean is

$$E(y_t|\Omega_{t-1}) = 0,$$

[10] If you need to refresh your memory on conditional means, consult any good introductory statistics book, such as Wonnacott and Wonnacott (1990).

and the conditional variance is

$$var(y_t|\Omega_{t-1}) = E[(y_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}] = \sigma^2.$$

Conditional and unconditional means and variances are identical for an independent white noise series; there are no dynamics in the process, and hence no dynamics in the conditional moments.

6.3 Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions

Now suppose we have a sample of data on a time series, and we don't know the true model that generated the data, or the mean, autocorrelation function or partial autocorrelation function associated with that true model. Instead, we want to use the data to estimate the mean, autocorrelation function, and partial autocorrelation function, which we might then use to help us learn about the underlying dynamics, and to decide upon a suitable model or set of models to fit to the data.

6.3.1 Sample Mean

The mean of a covariance stationary series is

$$\mu = Ey_t.$$

A fundamental principle of estimation, called the **analog principle**, suggests that we develop estimators by replacing expectations with sample averages. Thus our estimator for the population mean, given a sample of size T, is the **sample mean**,

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

Typically we're not directly interested in the estimate of the mean, but it's needed for estimation of the autocorrelation function.

6.3.2 Sample Autocorrelations

The autocorrelation at displacement τ for the covariance stationary series y is

$$\rho(\tau) = \frac{E\left[(y_t - \mu)(y_{t-\tau} - \mu)\right]}{E[(y_t - \mu)^2]}.$$

Application of the analog principle yields a natural estimator,

$$\hat{\rho}(\tau) = \frac{\frac{1}{T}\sum_{t=\tau+1}^{T}\left[(y_t - \bar{y})(y_{t-\tau} - \bar{y})\right]}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\sum_{t=\tau+1}^{T}\left[(y_t - \bar{y})(y_{t-\tau} - \bar{y})\right]}{\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

This estimator, viewed as a function of τ, is called the **sample autocorrelation function**, or **correlogram**. Note that some of the summations begin at $t = \tau + 1$, not at $t = 1$; this is necessary because of the appearance of $y_{t-\tau}$ in the sum. Note that we divide those same sums by T, even though only $T - \tau$ terms appear in the sum. When T is large relative to τ (which is the relevant case), division by T or by $T - \tau$ will yield approximately the same result, so it won't make much difference for practical purposes, and moreover there are good mathematical reasons for preferring division by T.

It's often of interest to assess whether a series is reasonably approximated as white noise, which is to say whether all its autocorrelations are zero in population. A key result, which we simply assert, is that if a series is white noise, then the distribution of the sample autocorrelations in large samples is

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right).$$

Note how simple the result is. The sample autocorrelations of a white noise series are approximately normally distributed, and the normal is always a convenient distribution to work with. Their mean is zero, which is to say the sample autocorrelations are unbiased estimators of the true autocorrelations, which are in fact zero. Finally, the variance of the sample autocorrelations is approximately $1/T$ (equivalently, the standard deviation is $1/\sqrt{T}$), which is easy to construct and remember. Under normality, taking plus or minus two standard errors yields an approximate 95% confidence interval. Thus, if the series is white noise, approximately 95% of the sample autocorrelations should fall in the interval $0 \pm 2/\sqrt{T}$. In practice, when we plot the sample autocorrelations for a sample of data, we typically include the "two standard error bands," which are useful for making informal graphical assessments of whether and how the series deviates from white noise.

The two-standard-error bands, although very useful, only provide 95% bounds for the sample autocorrelations taken one at a time. Ultimately, we're often interested in whether a series is white noise, that is, whether *all* its autocorrelations are *jointly* zero. A simple extension lets us test that hypothesis. Rewrite the expression

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right)$$

as

$$\sqrt{T}\hat{\rho}(\tau) \sim N(0, 1).$$

Squaring both sides yields[11]

$$T\hat{\rho}^2(\tau) \sim \chi^2_1.$$

It can be shown that, in addition to being approximately normally distributed, the sample autocorrelations at various displacements are approximately independent of one another. Recalling that the sum of independent χ^2 variables is also χ^2 with degrees of freedom equal to the sum of the degrees

[11] Recall that the square of a standard normal random variable is a χ^2 random variable with one degree of freedom. We square the sample autocorrelations $\hat{\rho}(\tau)$ so that positive and negative values don't cancel when we sum across various values of τ, as we will soon do.

of freedom of the variables summed, we have shown that the **Box-Pierce Q-statistic**,

$$Q_{BP} = T \sum_{\tau=1}^{m} \hat{\rho}^2(\tau),$$

is approximately distributed as a χ^2_m random variable under the null hypothesis that y is white noise.[12] A slight modification of this, designed to follow more closely the χ^2 distribution in small samples, is

$$Q_{LB} = T(T+2) \sum_{\tau=1}^{m} \left(\frac{1}{T-\tau} \right) \hat{\rho}^2(\tau).$$

Under the null hypothesis that y is white noise, Q_{LB} is approximately distributed as a χ^2_m random variable. Note that the **Ljung-Box Q-statistic** is the same as the Box-Pierce Q statistic, except that the sum of squared autocorrelations is replaced by a weighted sum of squared autocorrelations, where the weights are $(T+2)/(T-\tau)$. For moderate and large T, the weights are approximately 1, so that the Ljung-Box statistic differs little from the Box-Pierce statistic.

Selection of m is done to balance competing criteria. On one hand, we don't want m too small, because after all, we're trying to do a joint test on a large part of the autocorrelation function. On the other hand, as m grows relative to T, the quality of the distributional approximations we've invoked deteriorates. In practice, focusing on m in the neighborhood of $\sqrt{T}$ is often reasonable.

6.3.3 Sample Partial Autocorrelations

Recall that the partial autocorrelations are obtained from population linear regressions, which correspond to a thought experiment involving linear regression using an infinite sample of data. The sample partial autocorrelations

[12] m is a maximum displacement selected by the user. Shortly we'll discuss how to choose it.

correspond to the same thought experiment, except that the linear regression is now done on the (feasible) sample of size T. If the fitted regression is

$$\hat{y}_t = \hat{c} + \hat{\beta}_1 y_{t-1} + ... + \hat{\beta}_\tau y_{t-\tau},$$

then the **sample partial autocorrelation** at displacement τ is

$$\hat{p}(\tau) \equiv \hat{\beta}_\tau.$$

Distributional results identical to those we discussed for the sample autocorrelations hold as well for the sample *partial* autocorrelations. That is, if the series is white noise, approximately 95% of the sample partial autocorrelations should fall in the interval $\pm 2/\sqrt{T}$. As with the sample autocorrelations, we typically plot the sample partial autocorrelations along with their two-standard-error bands.

A "**correlogram analysis**" simply means examination of the sample autocorrelation and partial autocorrelation functions (with two standard error bands), together with related diagnostics, such as Q statistics.

We don't show the sample autocorrelation or partial autocorrelation at displacement 0, because as we mentioned earlier, they equal 1.0, by construction, and therefore convey no useful information. We'll adopt this convention throughout.

Note that the sample autocorrelation and partial autocorrelation are identical at displacement 1. That's because at displacement 1, there are no earlier lags to control for when computing the sample partial autocorrelation, so it equals the sample autocorrelation. At higher displacements, of course, the two diverge.

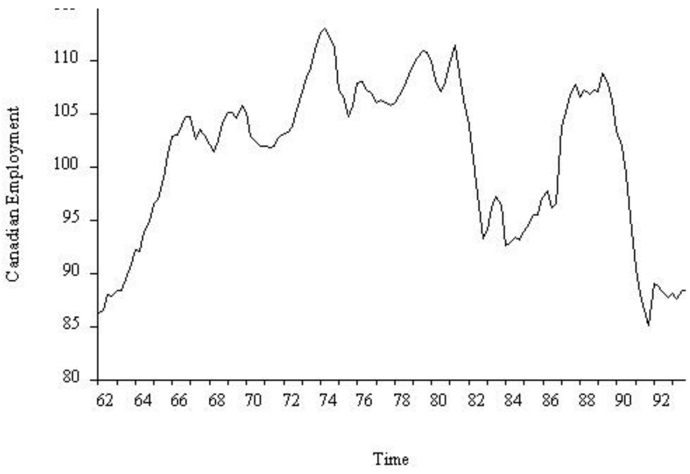


Figure 6.9: Canadian Employment Index

6.4 Canadian Employment I: Characterizing Cycles

To illustrate the ideas we've introduced, we examine a quarterly, seasonally-adjusted index of Canadian employment, 1962.1 - 1993.4, which we plot in Figure 6.9. The series displays no trend, and of course it displays no seasonality because it's seasonally adjusted. It does, however, appear highly serially correlated. It evolves in a slow, persistent fashion – high in business cycle booms and low in recessions.

To get a feel for the dynamics operating in the employment series we perform a correlogram analysis.[13] The results appear in Table 1. Consider first the Q statistic.[14] We compute the Q statistic and its p-value under the null hypothesis of white noise for values of m (the number of terms in the sum that underlies the Q statistic) ranging from one through twelve. The p-value is consistently zero to four decimal places, so the null hypothesis of white noise is decisively rejected.

Now we examine the sample autocorrelations and partial autocorrelations. The sample autocorrelations are very large relative to their standard errors

[13] A "correlogram analysis" simply means examination of the sample autocorrelation and partial autocorrelation functions (with two standard error bands), together with related diagnostics, such as Q statistics.

[14] We show the Ljung-Box version of the Q statistic.

and display slow one-sided decay.[15] The sample partial autocorrelations, in contrast, are large relative to their standard errors at first (particularly for the 1-quarter displacement) but are statistically negligible beyond displacement 2.[16] In Figure 6.10 we plot the sample autocorrelations and partial autocorrelations along with their two standard error bands.

It's clear that employment has a strong cyclical component; all diagnostics reject the white noise hypothesis immediately. Moreover, the sample autocorrelation and partial autocorrelation functions have particular shapes – the autocorrelation function displays slow one-sided damping, while the partial autocorrelation function cuts off at displacement 2. Such patterns, which summarize the dynamics in the series, can be useful for suggesting candidate forecasting models. Such is indeed the case.

[15] We don't show the sample autocorrelation or partial autocorrelation at displacement 0, because as we mentioned earlier, they equal 1.0, by construction, and therefore convey no useful information. We'll adopt this convention throughout.

[16] Note that the sample autocorrelation and partial autocorrelation are identical at displacement 1. That's because at displacement 1, there are no earlier lags to control for when computing the sample partial autocorrelation, so it equals the sample autocorrelation. At higher displacements, of course, the two diverge.

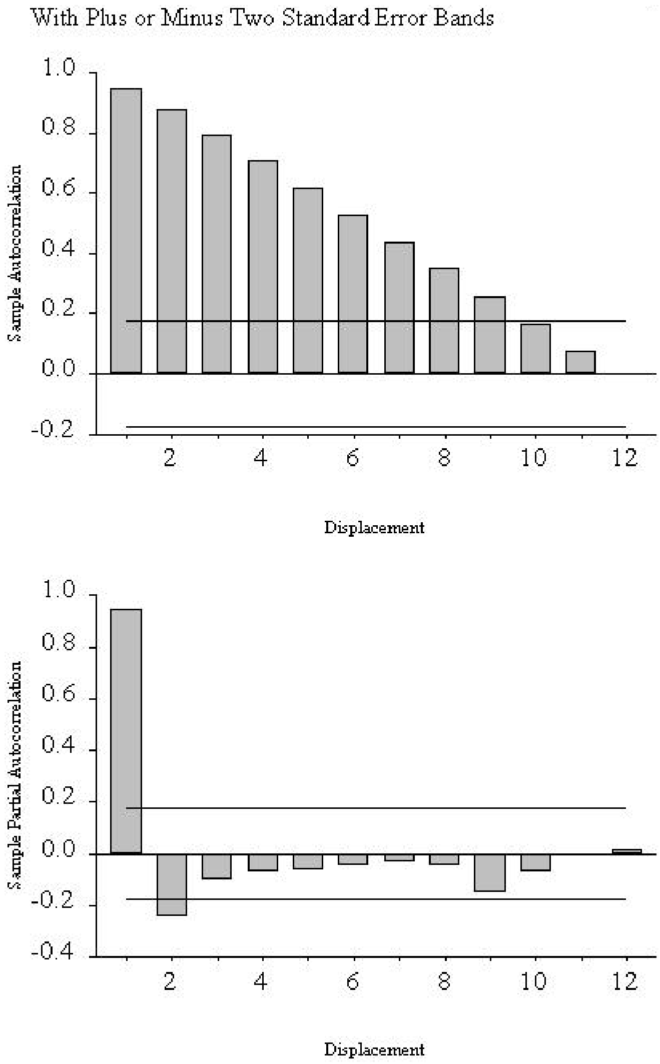


Figure 6.10: Sample Autocorrelation and Sample Partial Autocorrelation

6.5 Modeling Cycles With Autoregressions

6.5.1 Some Preliminary Notation: The Lag Operator

The **lag operator** and related constructs are the natural language in which time series models are expressed. If you want to understand and manipulate time series models – indeed, even if you simply want to be able to read the software manuals – you have to be comfortable with the lag operator. The lag operator, L, is very simple: it "operates" on a series by lagging it. Hence $Ly_t = y_{t-1}$. Similarly, $L^2 y_t = L(L(y_t)) = L(y_{t-1}) = y_{t-2}$, and so on. Typically we'll operate on a series not with the lag operator but with a **polynomial in the lag operator**. A lag operator polynomial of degree m is just a linear function of powers of L, up through the m-th power,

$$B(L) = b_0 + b_1 L + b_2 L^2 + ...b_m L^m.$$

To take a very simple example of a lag operator polynomial operating on a series, consider the m-th order lag operator polynomial L^m, for which

$$L^m y_t = y_{t-m}.$$

A well-known operator, the first-difference operator Δ, is actually a first-order polynomial in the lag operator; you can readily verify that

$$\Delta y_t = (1 - L) y_t = y_t - y_{t-1}.$$

As a final example, consider the second-order lag operator polynomial $1 + .9L + .6L^2$ operating on y_t. We have

$$(1 + .9L + .6L^2) y_t = y_t + .9 y_{t-1} + .6 y_{t-2},$$

which is a weighted sum, or **distributed lag**, of current and past values. All time-series models, one way or another, must contain such distributed lags, because they've got to quantify how the past evolves into the present and future; hence lag operator notation is a useful shorthand for stating and manipulating time-series models.

Thus far we've considered only finite-order polynomials in the lag operator; it turns out that infinite-order polynomials are also of great interest. We write the infinite-order lag operator polynomial as

$$B(L) = b_0 + b_1 L + b_2 L^2 + ... = \sum_{i=0}^{\infty} b_i L^i.$$

Thus, for example, to denote an infinite distributed lag of current and past shocks we might write

$$B(L)\varepsilon_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ... = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}.$$

At first sight, infinite distributed lags may seem esoteric and of limited practical interest, because models with infinite distributed lags have infinitely many parameters $(b_0, b_1, b_2, ...)$ and therefore can't be estimated with a finite sample of data. On the contrary, and surprisingly, it turns out that models involving infinite distributed lags are central to time series modeling, as we shall soon see in detail.

6.5.2 Autoregressive Processes

Here we emphasize a very important model of cycles, the **autoregressive** **(AR) model**.

We begin by characterizing the autocorrelation function and related quantities under the assumption that the AR model is the DGP.[17] These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which

[17] Sometimes we call time series models of cycles "time series processes," which is short for **stochastic processes**.

is necessary to perform intelligent modeling. They enable us to make statements such as "If the data were really generated by an autoregressive process, then we'd expect its autocorrelation function to have property x." Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the AIC and the SIC, to suggest candidate models, which we then estimate.

The autoregressive process is a natural time-series model of cycles. It's simply a *stochastic difference equation*, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

6.5.3 Autoregressive Disturbances and Lagged Dependent Variables

You already know the first-order autoregressive ($AR(1)$) model as a model of cyclical dynamics in regression disturbances. Recall, in particular the Durbin-Watson environment that we introduced earlier in Chapter 3:

$$y_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t$$

$$v_t \overset{iid}{\sim} N(0, \sigma^2).$$

To strip things to their essentials, suppose that the only regressor is an intercept.[18] Then we have:

$$y_t = \mu + \varepsilon_t \tag{6.1}$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t$$

[18] In later chapters we'll bring in trends, seasonals, and other standard "x variables."

$$v_t \sim iid(0, \sigma^2).$$

Now let us manipulate this "regression with serially-correlated disturbances" as follows. Because

$$y_t = \mu + \varepsilon_t,$$

we have

$$y_{t-1} = \mu + \varepsilon_{t-1},$$

so

$$\phi y_{t-1} = \phi\mu + \phi\varepsilon_{t-1}. \tag{6.2}$$

Subtracting 6.2 from 6.1 produces

$$y_t - \phi y_{t-1} = \mu(1 - \phi) + (\varepsilon_t - \phi\varepsilon_{t-1}),$$

or

$$y_t = \mu(1 - \phi) + \phi y_{t-1} + v_t.$$

Hence we have arrived at a model of "regression a lagged dependent variable with iid disturbances." The two models are mathematically identical. LDV with classical disturbances does the same thing as no LDV with serially-correlated disturbances. Each approach "mops up" serial correlation not explained by other regressors. (And in this extreme case, there are no other regressors.)

In this chapter we'll focus on univariate models with LDV's, and again, to isolate the relevant issues we'll focus on models with *only* LDV's. Later, in Chapter 16, we'll add x's as well.

The $AR(1)$ Process for Observed Series

The first-order autoregressive process, $AR(1)$ for short, is

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

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L) y_t = \varepsilon_{t.}$$

In Figure 6.11 we show simulated realizations of length 150 of two $AR(1)$ processes; the first is

$$y_t = .4 y_{t-1} + \varepsilon_t,$$

and the second is

$$y_t = .95 y_{t-1} + \varepsilon_t,$$

where in each case

$$\varepsilon_t \sim iidN(0, 1),$$

and the same innovation sequence underlies each realization. The fluctuations in the $AR(1)$ with parameter $\phi = .95$ appear much more persistent that those of the $AR(1)$ with parameter $\phi = .4$. Thus the $AR(1)$ model is capable of capturing highly persistent dynamics.

A certain condition involving the autoregressive lag operator polynomial must be satisfied for an autoregressive process to be covariance stationary. The condition is that all roots of the autoregressive lag operator polynomial must be outside the unit circle. In the $AR(1)$ case we have

$$(1 - \phi L) y_t = \varepsilon_t,$$

so the autoregressive lag operator polynomial is $1 - \phi L$, with root $1/\phi$. Hence the $AR(1)$ process is covariance stationary if $|\phi| < 1$.

Let's investigate the moment structure of the $AR(1)$ process. If we begin with the $AR(1)$ process,

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

and substitute backward for lagged y's on the right side, we obtain the so-

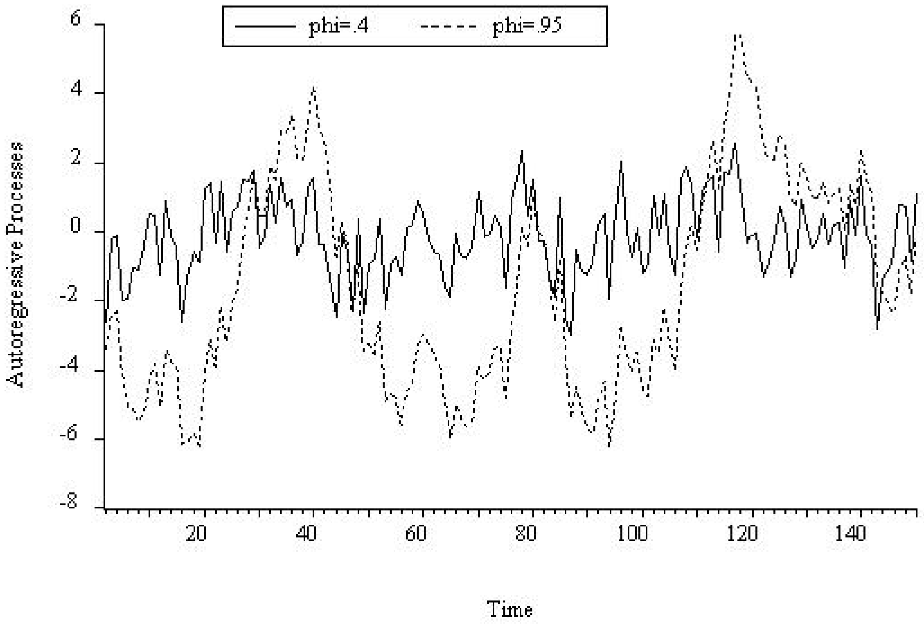


Figure 6.11: Realizations of Two AR(1) Processes

called **"moving-average representation"**

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...$$

The existence of a moving-average representation is very intuitive. Ultimately the ε's are the only things that move y, so it is natural that we should be able to express y in terms of the history of ε. We will have much more to say about that in Chapter 7. The existence of a moving-average representation is also very useful, because it facilitates some important calculations, to which we now turn.

From the moving average representation of the covariance stationary $AR(1)$

process, we can compute the unconditional mean and variance,

$$E(y_t) = E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...)$$

$$= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + ...$$

$$= 0$$

and

$$var(y_t) = var(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...)$$

$$= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + ...$$

$$= \sigma^2 \textstyle\sum_{i=0}^{\infty} \phi^{2i}$$

$$= \tfrac{\sigma^2}{1-\phi^2}.$$

The conditional moments, in contrast, are

$$E(y_t|y_{t-1}) = E(\phi y_{t-1} + \varepsilon_t|y_{t-1})$$

$$= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1})$$

$$= \phi y_{t-1} + 0$$

$$= \phi y_{t-1}$$

and

$$var(y_t|y_{t-1}) = var((\phi y_{t-1} + \varepsilon_t)|y_{t-1})$$

$$= \phi^2 var(y_{t-1}|y_{t-1}) + var(\varepsilon_t|y_{t-1})$$

$$= 0 + \sigma^2$$

$$= \sigma^2.$$

Note in particular that the simple way that the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

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

so that multiplying both sides of the equation by $y_{t-\tau}$ we obtain

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

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \phi\gamma(\tau - 1).$$

This is called the **Yule-Walker equation**. It is a recursive equation; that is, given $\gamma(\tau)$, for any τ, the Yule-Walker equation immediately tells us how to get $\gamma(\tau + 1)$. If we knew $\gamma(0)$ to start things off (an "initial condition"), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know $\gamma(0)$; it's just the variance of the process, which we already showed to be

$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}.$$

Thus we have

$$\gamma(0) = \frac{\sigma^2}{1-\phi^2}$$

$$\gamma(1) = \phi\frac{\sigma^2}{1-\phi^2}$$

$$\gamma(2) = \phi^2\frac{\sigma^2}{1-\phi^2},$$

and so on. In general, then,

$$\gamma(\tau) = \phi^\tau\frac{\sigma^2}{1-\phi^2}, \tau = 0, 1, 2, ...$$

Dividing through by $\gamma(0)$ gives the autocorrelations,

$$\rho(\tau) = \phi^\tau, \tau = 0, 1, 2, ...$$

Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach zero in the limit as the displacement approaches infinity. If ϕ is positive, the autocorrelation decay is one-sided. If ϕ is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is $\phi > 0$, but either way, the autocorrelations damp gradually. In Figure 6.12 and 6.13 we show the autocorrelation functions for $AR(1)$ processes with parameters $\phi = .4$ and $\phi = .95$. The persistence is much stronger when $\phi = .95$.

Finally, the partial autocorrelation function for the $AR(1)$ process cuts off abruptly; specifically,

$$p(\tau) = \begin{cases} \phi, \tau = 1 \\ \\ 0, \tau > 1. \end{cases} .$$

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an $AR(1)$, the first partial autocorrelation is just the autoregressive coefficient, and coefficients on all longer lags are zero.

In Figures 6.14 and 6.15 we show the partial autocorrelation functions for our two $AR(1)$ processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (.4 and .95, respectively), and at longer displacements, the partial autocorrelations are zero.

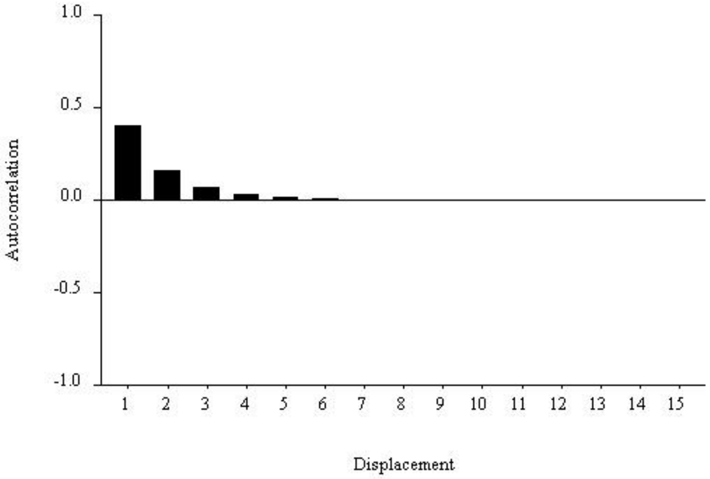


Figure 6.12: Population Autocorrelation Function: $\rho = .4$

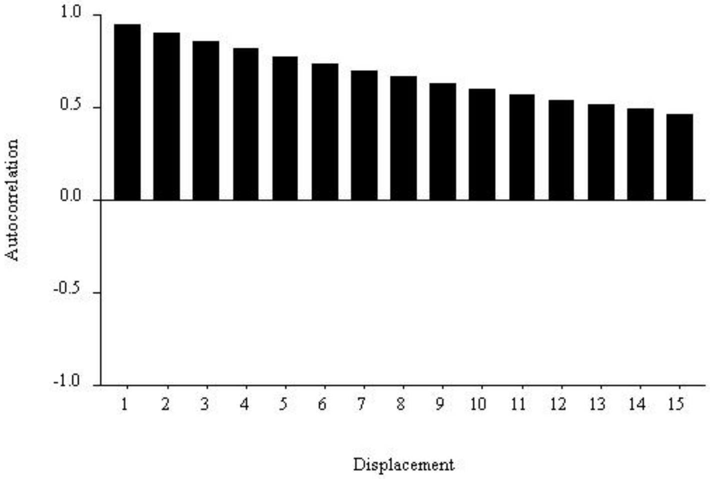


Figure 6.13: Population Autocorrelation Function: $\rho = .95$

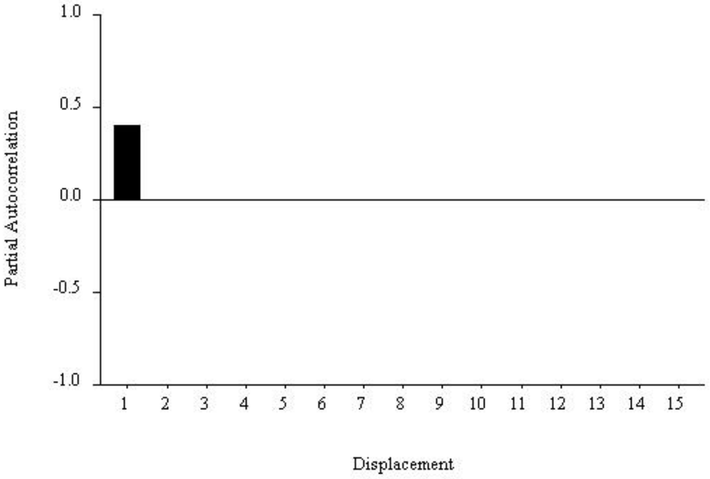


Figure 6.14: Partial Autocorrelation Function: $\rho = .4$

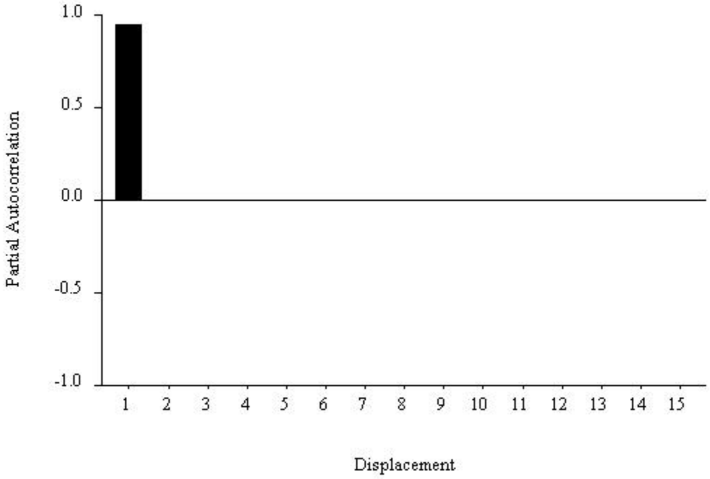


Figure 6.15: Partial Autocorrelation Function: $\rho = .95$

6.5.4 The $AR(p)$ Process

The general p-th order autoregressive process, or $AR(p)$ for short, is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

In lag operator form we write

$$\Phi(L) y_t = (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) y_t = \varepsilon_{t.}$$

In our discussion of the $AR(p)$ process we dispense with mathematical derivations and instead rely on parallels with the $AR(1)$ case to establish intuition for its key properties.

An $AR(p)$ process is covariance stationary if and only if all roots of the autoregressive lag operator polynomial $\Phi(L)$ are outside the unit circle.[19]

The autocorrelation function for the general $AR(p)$ process, as with that of the $AR(1)$ process, decays gradually with displacement. Finally, the $AR(p)$ partial autocorrelation function has a sharp cutoff at displacement p, for the same reason that the $AR(1)$ partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the $AR(p)$ autocorrelation function in a bit greater depth. The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the $AR(1)$ autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the $AR(1)$ case with a positive coefficient, but it can also have damped oscillation in ways that $AR(1)$ can't have. In the $AR(1)$ case, the only possible oscillation occurs when the coefficient is negative, in which case

[19] A necessary condition for covariance stationarity, which is often useful as a quick check, is $\sum_{i=1}^{p} \phi_i < 1$. If the condition is satisfied, the process may or may not be stationary, but if the condition is violated, the process can't be stationary.

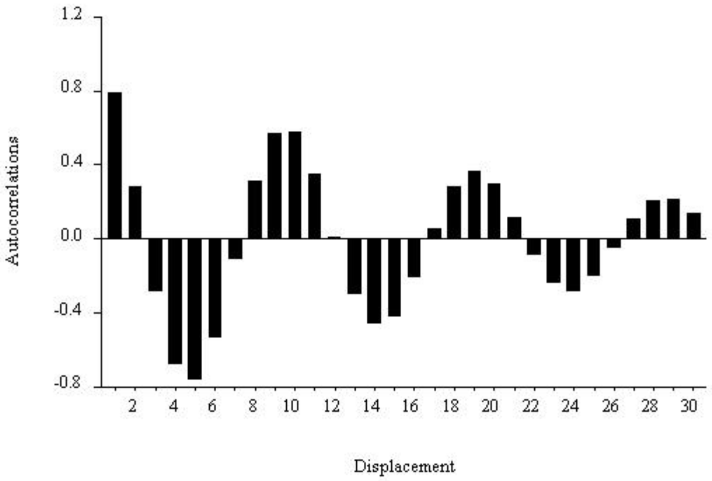


Figure 6.16: Autocorrelation Function of AR(2) with Complex Roots

the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.[20] Consider, for example, the $AR(2)$ process,

$$y_t = 1.5y_{t-1} - .9y_{t-2} + \varepsilon_{t.}$$

The corresponding lag operator polynomial is $1 - 1.5L + .9L^2$, with two complex conjugate roots, $.83 \pm .65i$. The inverse roots are $.75 \pm .58i$, both of which are close to, but inside, the unit circle; thus the process is covariance stationary. It can be shown that the autocorrelation function for an $AR(2)$ process is

$$\rho(0) = 1$$

$$\rho(\tau) = \phi_1\rho(\tau - 1) + \phi_2\rho(\tau - 2), \tau = 2, 3, ...$$

$$\rho(1) = \frac{\phi_1}{1 - \phi_2}$$

Using this formula, we can evaluate the autocorrelation function for the

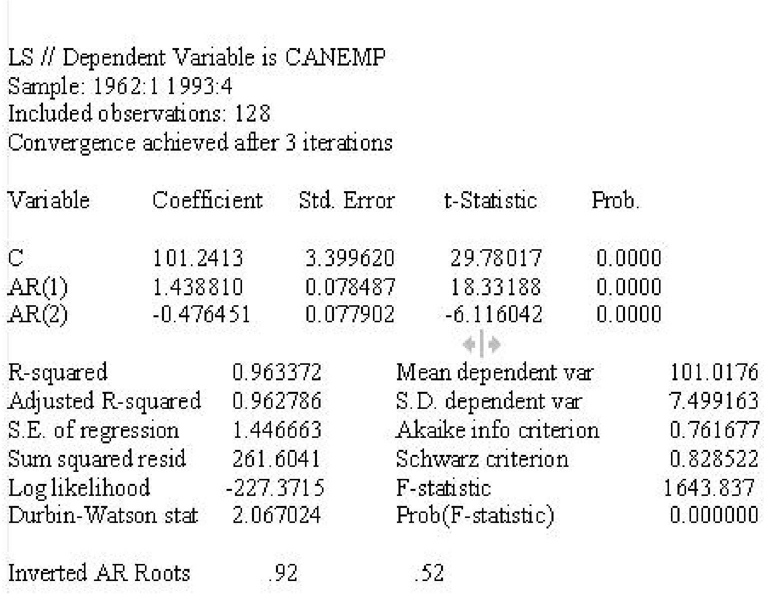
[20] Note that complex roots can't occur in the $AR(1)$ case.

process at hand; we plot it in Figure 6.16. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

6.6 Canadian Employment II: Modeling Cycles

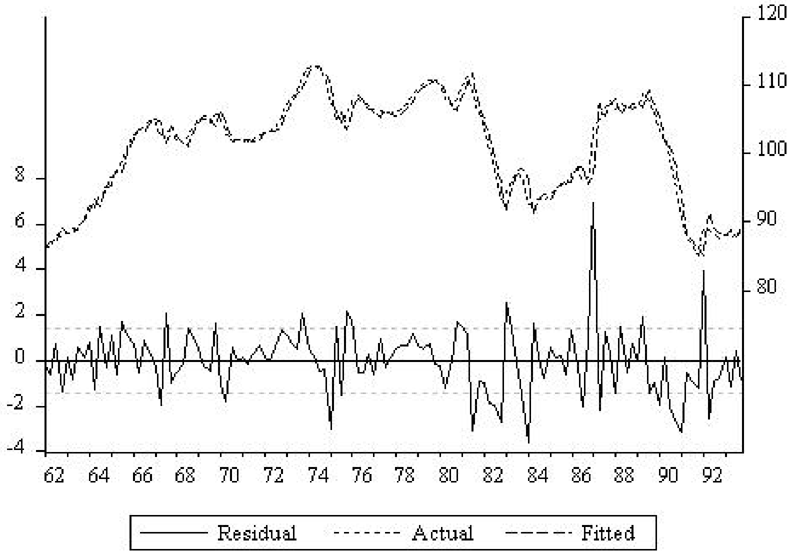
The sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy – just standard OLS regressions. In the $AR(1)$ case, we simply run an ordinary least squares regression of y on one lag of y; in the $AR(p)$ case, we regress y on p lags of y.

We estimate $AR(p)$ models, $p = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the $AR(2)$ is best. To save space, we report only the results of $AR(2)$ estimation in Table 6.17a. The estimation results look good, and the residuals (Figure 6.17b) look like white noise. The residual correlogram (Table 6.18, Figure 6.19) supports that conclusion.



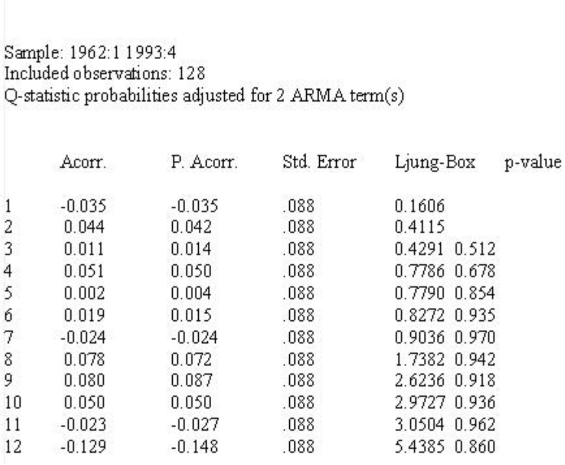
```
LS // Dependent Variable is CANEMP
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 3 iterations

Variable        Coefficient    Std. Error     t-Statistic     Prob.

C               101.2413       3.399620       29.78017        0.0000
AR(1)           1.438810       0.078487       18.33188        0.0000
AR(2)           -0.476451      0.077902       -6.116042       0.0000

R-squared             0.963372      Mean dependent var       101.0176
Adjusted R-squared    0.962786      S.D. dependent var       7.499163
S.E. of regression    1.446663      Akaike info criterion    0.761677
Sum squared resid     261.6041      Schwarz criterion        0.828522
Log likelihood        -227.3715     F-statistic              1643.837
Durbin-Watson stat    2.067024      Prob(F-statistic)        0.000000

Inverted AR Roots          .92           .52
```

(a) Employment: AR(2) Model



(b) Employment: AR(2) Model, Residual Plot

Figure 6.17: Employment: AR(2) Model



```
Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 2 ARMA term(s)

      Acorr.      P. Acorr.     Std. Error    Ljung-Box    p-value

1     -0.035      -0.035        .088          0.1606
2      0.044      0.042         .088          0.4115
3      0.011      0.014         .088          0.4291  0.512
4      0.051      0.050         .088          0.7786  0.678
5      0.002      0.004         .088          0.7790  0.854
6      0.019      0.015         .088          0.8272  0.935
7     -0.024      -0.024        .088          0.9036  0.970
8      0.078      0.072         .088          1.7382  0.942
9      0.080      0.087         .088          2.6236  0.918
10     0.050      0.050         .088          2.9727  0.936
11    -0.023      -0.027        .088          3.0504  0.962
12    -0.129      -0.148        .088          5.4385  0.860
```

Figure 6.18: Employment: AR(2) Model, Residual Correlogram

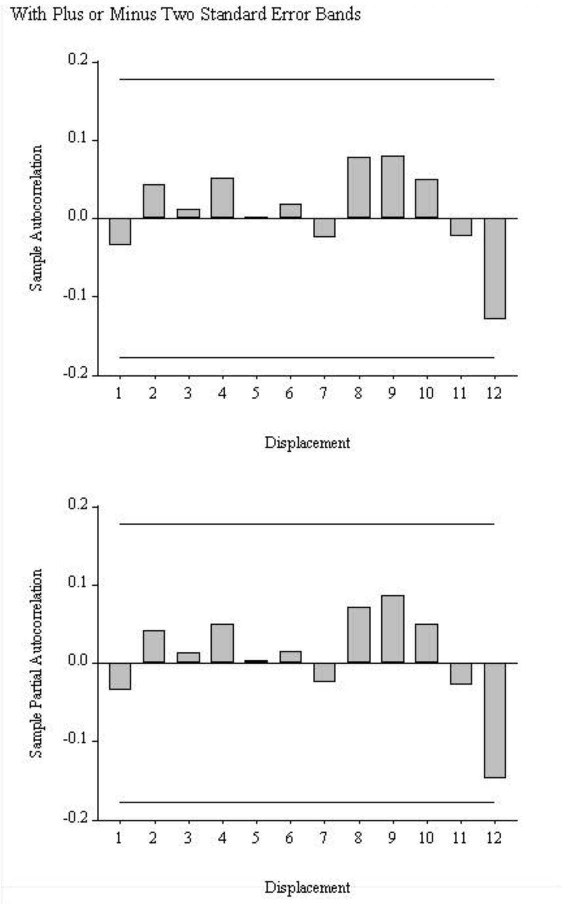


Figure 6.19: Employment: AR(2) Model, Residual Sample Autocorrelation and Partial Autocorrelation

6.7 Forecasting Cycles with Autoregressions

6.7.1 On the FRV Problem

We have seen that the FRV problem arises in general, but not in cross sections, and not in deterministic-trend time-series environments, and not in deterministic-seasonal time-series environments. The same is true in certain other time-series environments.

In particular, forget about trends and seasonals for the moment. Still the FRV problem does not arise if the RHS variables are *lagged* sufficiently relative the the forecast horizon of interest. Suppose, for example, that an acceptable model is

$$y_t = \beta_1 + \beta_2 x_{t-1} + \varepsilon_t. \tag{6.3}$$

The RHS variable is lagged by one period, so model 6.3 is immediately usable for 1-step-ahead forecasting without the FRV problem. More lags of x can of course be included; the key for 1-step-ahead forecasting is that all variables on the right be lagged by at least one period.

Forecasting more than one step ahead in model 6.3, however, would appear to lead again to the FRV problem: If we want to forecast h steps ahead, then all variables on the right must be lagged by at least h periods, not just by 1 period. Perhaps surprisingly, it actually remains *easy* to circumvent the FRV problem in autoregressive models. For example, models with $y_t \rightarrow y_{t-1}$ or $y_t \rightarrow y_{t-1}, x_{t-1}$ can effectively be transformed to models with $y_t \rightarrow y_{t-h}$ or $y_t \rightarrow y_{t-h}, x_{t-h}$, as we will see in this section.

6.7.2 Information Sets, Conditional Expectations, and Linear Projections

By now you've gotten comfortable with the idea of an **information set**. Here we'll use that idea extensively. We denote the time-T information set by Ω_T. Think of the information set as containing the available past history of the series,

$$\Omega_T = \{y_T,\ y_{T-1},\ y_{T-2},\ ...\},$$

where for theoretical purposes we imagine history as having begun in the infinite past.

Based upon that information set, we want to find the **optimal forecast** of y at some future time $T+h$. The optimal forecast is the one with the smallest loss on average, that is, the forecast that minimizes **expected loss**. It turns out that under reasonably weak conditions the optimal forecast is the **conditional mean**,

$$E(y_{T+h}|\Omega_T),$$

the expected value of the future value of the series being forecast, conditional upon available information.

In general, the conditional mean need not be a linear function of the elements of the information set. Because linear functions are particularly tractable, we prefer to work with **linear forecasts** – forecasts that are linear in the elements of the information set – by finding the best linear approximation to the conditional mean, called the **linear projection**, denoted

$$P(y_{T+h}|\Omega_T).$$

This explains the common term "**linear least squares forecast**." The linear projection is often very useful and accurate, because the conditional mean is often close to linear. In fact, in the Gaussian case the conditional expectation is exactly linear, so that

$$E(y_{T+h}|\Omega_T) = P(y_{T+h}|\Omega_T).$$

6.7.3 Point Forecasts for Autoregressions: Wold's Chain Rule

A very simple recursive method for computing optimal h-step-ahead point forecasts, for any desired h, is available for autoregressions.

The recursive method, called the **chain rule of forecasting**, is best learned by example. Consider the $AR(1)$ process,

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

$$\varepsilon_t \sim WN(0, \sigma^2).$$

First we construct the optimal 1-step-ahead forecast, and then we construct the optimal 2-step-ahead forecast, which depends on the optimal 1-step-ahead forecast, which we've already constructed. Then we construct the optimal 3-step-ahead forecast, which depends on the already-computed 2-step-ahead forecast, which we've already constructed, and so on.

To construct the 1-step-ahead forecast, we write out the process for time $T+1$,

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

Then, projecting the right-hand side on the time-T information set, we obtain

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

Now let's construct the 2-step-ahead forecast. Write out the process for time $T+2$,

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

Then project directly on the time-T information set to get

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Note that the future innovation is replaced by 0, as always, and that we have directly replaced the time $T+1$ value of y with its earlier-constructed optimal forecast. Now let's construct the 3-step-ahead forecast. Write out the process for time $T+3$,

$$y_{T+3} = \phi y_{T+2} + \varepsilon_{T+3}.$$

Then project directly on the time-T information set,

$$y_{T+3,T} = \phi y_{T+2,T}.$$

The required 2-step-ahead forecast was already constructed.

Continuing in this way, we can recursively build up forecasts for any and all future periods. Hence the name "chain rule of forecasting." Note that, for the $AR(1)$ process, only the most recent value of y is needed to construct optimal forecasts, for any horizon, and for the general $AR(p)$ process only the p most recent values of y are needed. In particular, for our $AR(1)$ case,

$$y_{T+h,T} = \phi^h y_T.$$

As usual, in truth the parameters are unknown and so must be estimated, so we turn infeasible forecasts into feasible ("operational") forecasts by inserting the usual estimates where unknown parameters appear.

It is worth noting that thanks to Wold's chain rule we have now solved the FRV problem for autoregressions, as we did earlier for cross sections, trends, and seasonals! We have of course worked through the calculations in detail only for the $AR(1)$ case, but the approach is identical for the general $AR(p)$ case.

6.7.4 Density Forecasts

The chain rule is a device for simplifying the computation of point forecasts. Density forecasts require a bit more work. Let us again work through the $AR(1)$ case in detail, assuming normality and ignoring parameter estimation uncertainty.

We know that

$$y_{T+h} \ \sim \ N(y_{T+h,T}, \sigma_h^2),$$

where $\sigma_h^2 = var(y_{T+h}|\Omega_T)$ and $\Omega_T = \{y_T, y_{T-1}, ...\}$. Using Wold's chain rule we already derived the formula for $y_{T+h,T}$, so all we need is the h-step-ahead forecast error variance, σ_h^2.

First let us simply assert the general result. It is

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

Now let us derive the general result. First recall that the optimal forecasts are

$$y_{T+1,T} = \phi y_T$$

$$y_{T+2,T} = \phi^2 \ y_T$$

$$y_{T+h,T} = \phi^h \ y_T.$$

Second, note that the corresponding forecast errors are

$$e_{T+1,T} = (y_{T+1} - y_{T+1,T}) = \varepsilon_{T+1}$$

$$e_{T+2,T} = (y_{T+2} - y_{T+2,T}) = \phi\varepsilon_{T+1} + \varepsilon_{T+2}$$

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \phi\varepsilon_{T+h-1} + ... + \phi^{h-1}\varepsilon_{T+1}.$$

Third, note that the corresponding forecast error variances are

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1+\phi^2)$$

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

QED

Note that the limiting h-step-ahead forecast error variance is

$$\lim_{h \to \infty} \sigma_h^2 = \frac{\sigma^2}{1-\phi^2},$$

the unconditional variance of the $AR(1)$ process. (The conditioning information becomes progressively less valuable as $h \to \infty$ in covariance stationary environments, so the conditional variance converges to the unconditional variance.)

As usual, in truth the parameters are unknown and so must be estimated, so we turn infeasible forecasts into feasible ("operational") forecasts by inserting the usual estimates where unknown parameters appear. In addition, and also as usual, we can account for non-normality and parameter-estimation uncertainty using simulation methods. (Of course simulation could be used even under normality).

Density forecasts for higher-ordered autoregressions proceed in similar fashion. Point forecasts at any horizon come from Wold's chain rule. Under normality we still need the corresponding h-step forecast-error variances, we we infer from the moving-average representation. Dropping normality and using simulation methods does not even require the variance calculation.

6.8 Canadian Employment III: Forecasting

Now we put our forecasting technology to work to produce autoregressive point and interval forecasts for Canadian employment. Recall that the best autoregressive model was an $AR(2)$. In Figure 6.20 we show the 4-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much less quickly, as seems natural given the high persistence of employment. The 4-quarter-ahead point forecast, in fact, is still well below the mean. Similarly, the 95% error bands grow gradually and haven't approached their long-horizon values by four quarters out.

Figures 6.20 and 6.21 make clear the nature of the autoregressive forecasts. In Figure 6.21 we show the employment history, 4-quarter-ahead $AR(2)$ extrapolation forecast, and the realization. The $AR(2)$ forecast appears quite accurate; the mean squared forecast error is 1.3.

Figure 6.22 presents the 12-step-ahead extrapolation forecast, and Figure 6.23 presents a much longer-horizon extrapolation forecast. Eventually the unconditional mean *is* approached, and eventually the error bands do go flat, but only for very long-horizon forecasts, due to the high persistence in employment, which the $AR(2)$ model captures.

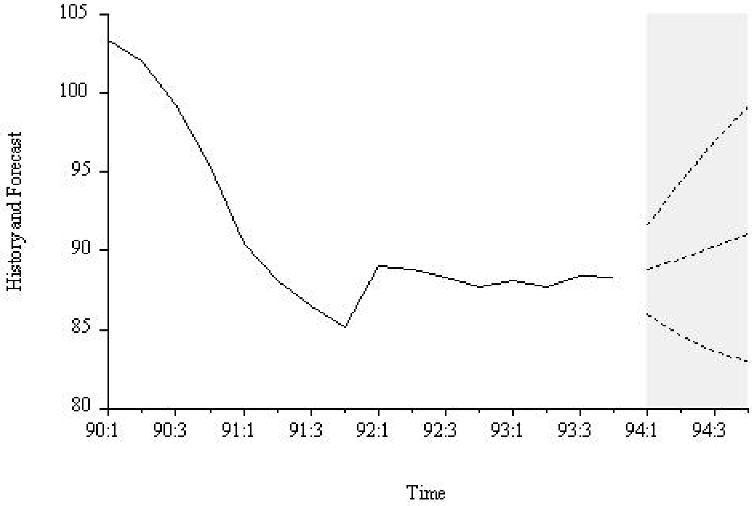


Figure 6.20: Employment History and Forecast: AR(2)

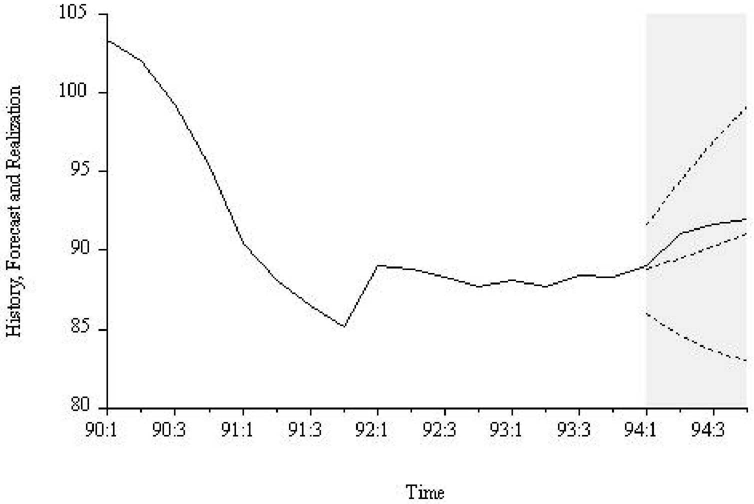


Figure 6.21: Employment History, Forecast, and Realization: AR(2)

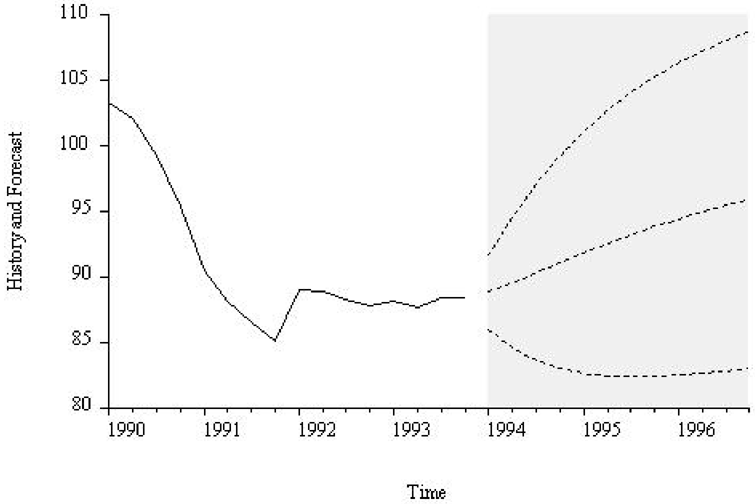


Figure 6.22: Employment History and Long-Horizon Forecast: AR(2)

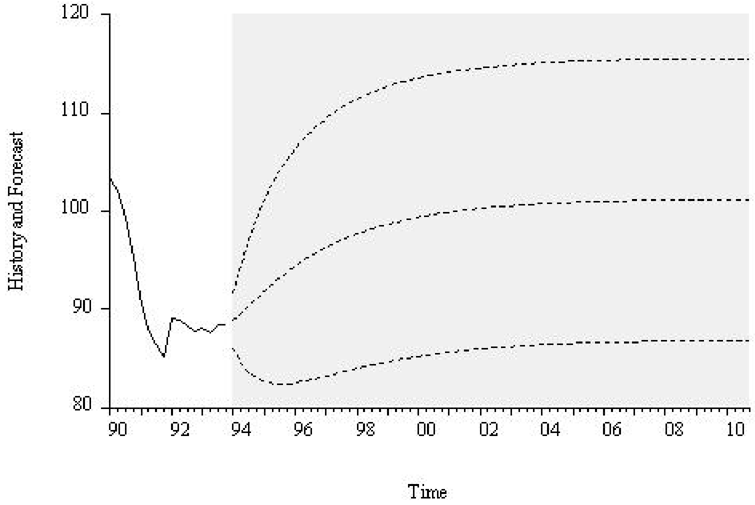


Figure 6.23: Employment History and Very Long-Horizon Forecast: AR(2)

6.9 Exercises, Problems and Complements

1. From FRED get Industrial Production, Total Index, 2012=100, Quarterly, Not Seasonally Adjusted, 1919:Q1-latest. First, hold out 2014.1-latest. Select and estimate your preferred model (deterministic trend + deterministic seasonal + autoregressive cyclical dynamics) using 1919:Q1-2013:Q4, and use your estimated model to generate a path forecast 2014:Q1-latest. Second, hold out nothing. Re-select and re-estimate using 1919:Q1-latest, and use your estimated model to generate a path forecast for the next eight quarters.

2. More on the stability condition for $AR(1)$ processes.

   The key stability condition is $|\phi| < 1$. Recall $y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$. This implies that $var(y_t) = \sum_{j=0}^{\infty} \phi^{2j} \sigma^2$, which is the sum of a geometric series. Hence:

   $$var(y_t) = \frac{\sigma^2}{1-\phi^2} \text{ if } |\phi| < 1$$

   $var(y_t) = \infty$ otherwise

3. A more complete picture of $AR(1)$ stability.

   The following are all aspects in which covariance stationarity corresponds to a nice, stable environment.

   (a) Series y_t is persistent but eventually reverts to a fixed mean.

   (b) Shocks ε_t have persistent effects but eventually die out. Hint: Consider $y_t = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}, \ |\phi| < 1$.

   (c) Autocorrelations $\rho(\tau)$ nonzero but decay to zero.

   (d) Autocorrelations $\rho(\tau)$ depend on τ (of course) but not on time. Hint: Use back substitution to relate y_t and y_{t-2}. How does it compare to the relation between y_t and y_{t-1} when $|\phi| < 1$?

(e) Series y_t varies but not too extremely. Hint: Consider $var(y_t) = \frac{\sigma^2}{1-\phi^2}$, $|\phi| < 1$.

4. Autocorrelation functions of covariance stationary series.

   While interviewing at a top investment bank, your interviewer is impressed by the fact that you have taken a course on forecasting. She decides to test your knowledge of the autocovariance structure of covariance stationary series and lists five autocovariance functions:

   a. $\gamma(t, \tau) = \alpha$

   b. $\gamma(t, \tau) = e^{-\alpha\tau}$

   c. $\gamma(t, \tau) = \alpha\tau$

   d. $\gamma(t, \tau) = \frac{\alpha}{\tau}$, where α is a positive constant. Which autocovariance function(s) are consistent with covariance stationarity, and which are not? Why?

5. Autocorrelation vs. partial autocorrelation.

   Describe the difference between autocorrelations and partial autocorrelations. How can autocorrelations at certain displacements be positive while the partial autocorrelations at those same displacements are negative?

6. Sample autocorrelation functions of trending series.

   A tell-tale sign of the slowly-evolving nonstationarity associated with trend is a sample autocorrelation function that damps extremely slowly.

   a. Find three trending series, compute their sample autocorrelation functions, and report your results. Discuss.

   b. Fit appropriate trend models, obtain the model residuals, compute their sample autocorrelation functions, and report your results. Discuss.

7. Sample autocorrelation functions of seasonal series.

   A tell-tale sign of seasonality is a sample autocorrelation function with sharp peaks at the seasonal displacements (4, 8, 12, etc. for quarterly data, 12, 24, 36, etc. for monthly data, and so on).

   a. Find a series with both trend and seasonal variation. Compute its sample autocorrelation function. Discuss.

   b. Detrend the series. Discuss.

   c. Compute the sample autocorrelation function of the detrended series. Discuss.

   d. Seasonally adjust the detrended series. Discuss.

   e. Compute the sample autocorrelation function of the detrended, seasonally-adjusted series. Discuss.

8. Lag operator expressions, I.

   Rewrite the following expressions without using the lag operator.

   a. $(L^{\tau})y_t = \varepsilon_t$

   b. $y_t = \left(\frac{2+5L+.8L^2}{L-.6L^3}\right)\varepsilon_t$

   c. $y_t = 2\left(1+\frac{L^3}{L}\right)\varepsilon_t.$

9. Lag operator expressions, II.

   Rewrite the following expressions in lag operator form.

   a. $y_t + y_{t-1} + ... + y_{t-N} = \alpha + \varepsilon_t + \varepsilon_{t-1} + ... + \varepsilon_{t-N}$, where α is a constant

   b. $y_t = \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t.$

10. Simulating time series processes.

Many cutting-edge estimation and forecasting techniques involve simulation. Moreover, simulation is often a good way to get a feel for a model and its behavior. White noise can be simulated on a computer using **random number generators**, which are available in most statistics, econometrics and forecasting packages.

(a) Simulate a Gaussian white noise realization of length 200. Call the white noise ε_t. Compute the correlogram. Discuss.

(b) Form the distributed lag $y_t = \varepsilon_t \ + .9\varepsilon_{t-1}$, t = 2, 3, ..., 200. Compute the sample autocorrelations and partial autocorrelations. Discuss.

(c) Let $y_1 = 1$ and $y_t = .9y_{t-1} + \varepsilon_t$, t = 2, 3, ..., 200. Compute the sample autocorrelations and partial autocorrelations. Discuss.

11. Diagnostic checking of model residuals.

If a forecasting model has extracted all the systematic information from the data, then what's left – the residual – should be white noise. More precisely, the true innovations are white noise, and if a model is a good approximation to the DGP then its 1-step-ahead forecast errors should be approximately white noise. The model residuals are the in-sample analog of out-of-sample 1-step-ahead forecast errors. Hence the usefulness of various tests of the hypothesis that residuals are white noise.

The Durbin-Watson test is the most popular. Recall the Durbin-Watson test statistic, discussed in Chapter 2,

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

Note that

$$\sum_{t=2}^{T}(e_t - e_{t-1})^2 \approx 2\sum_{t=2}^{T} e_t^2 - 2\sum_{t=2}^{T} e_t e_{t-1}.$$

Thus

$$DW \approx 2(1 - \hat{\rho}(1)),$$

so that the Durbin-Watson test is effectively based only on the first sample autocorrelation and really only tests whether the first autocorrelation is zero. We say therefore that the Durbin-Watson is a test for **first-order serial correlation**. In addition, the Durbin-Watson test is not valid in the presence of lagged dependent variables.[21] On both counts, we'd like a more general and flexible framework for diagnosing serial correlation. The residual correlogram, comprised of the residual sample autocorrelations, the sample partial autocorrelations, and the associated Q statistics, delivers the goods.

(a) When we discussed the correlogram in the text, we focused on the case of an observed time series, in which case we showed that the Q statistics are distributed as χ^2_m. Now, however, we want to assess whether unobserved model disturbances are white noise. To do so, we use the model residuals, which are estimates of the unobserved disturbances. Because we fit a model to get the residuals, we need to account for the degrees of freedom used. The upshot is that the distribution of the Q statistics under the white noise hypothesis is better approximated by a χ^2_{m-K} random variable, where K is the number of parameters estimated. That's why, for example, we don't report (and in fact the software doesn't compute) the p-values for the Q statistics associated with the residual correlogram of our employment forecasting model until $m > K$.

(b) **Durbin's h test** is an alternative to the Durbin-Watson test. As

[21] Following standard, if not strictly appropriate, practice, in this book we often report and examine the Durbin-Watson statistic even when lagged dependent variables are included. We always supplement the Durbin-Watson statistic, however, with other diagnostics such as the residual correlogram, which remain valid in the presence of lagged dependent variables, and which almost always produce the same inference as the Durbin-Watson statistic.

with the Durbin-Watson test, it's designed to detect first-order serial correlation, but it's valid in the presence of lagged dependent variables. Do some background reading as well on Durbin's h test and report what you learned.

(c) The **Breusch-Godfrey test** is another alternative to the Durbin-Watson test. It's designed to detect p^{th}-order serial correlation, where p is selected by the user, and is also valid in the presence of lagged dependent variables. Do some background reading on the Breusch-Godfrey procedure and report what you learned.

(d) Which do you think is likely to be most useful to you in assessing the properties of residuals from forecasting models: the residual correlogram, Durbin's h test, or the Breusch-Godfrey test? Why?

12. Forecast accuracy across horizons.

You are a consultant to MedTrax, a large pharmaceutical company, which released a new ulcer drug three months ago and is concerned about recovering research and development costs. Accordingly, MedTrax has approached you for drug sales projections at 1- through 12-month-ahead horizons, which it will use to guide potential sales force realignments. In briefing you, MedTrax indicated that it expects your long-horizon forecasts (e.g., 12-month-ahead) to be just as accurate as your short-horizon forecasts (e.g., 1-month-ahead). Explain to MedTrax why that is unlikely, even if you do the best forecasting job possible.

13. Forecasting an $AR(1)$ process with known and unknown parameters.

Use the chain rule to forecast the $AR(1)$ process,

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

For now, assume that all parameters are known.

a. Show that the optimal forecasts are

$$y_{T+1,T} = \phi y_T$$

$$y_{T+2,T} = \phi^2 \; y_T$$

$$y_{T+h,T} = \phi^h \; y_T.$$

b. Show that the corresponding forecast errors are

$$e_{T+1,T} = (y_{T+1} - y_{T+1,T}) = \varepsilon_{T+1}$$

$$e_{T+2,T} = (y_{T+2} - y_{T+2,T}) = \phi\varepsilon_{T+1} + \varepsilon_{T+2}$$

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \phi\varepsilon_{T+h-1} + ... + \phi^{h-1}\varepsilon_{T+1}.$$

c. Show that the forecast error variances are

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \phi^2)$$

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}.$$

d. Show that the limiting forecast error variance is

$$\begin{matrix} \lim \\ h \to \infty \end{matrix} \quad \sigma_h^2 = \frac{\sigma^2}{1 - \phi^2},$$

the unconditional variance of the $AR(1)$ process.

e. Now assume that the parameters are unknown and so must be estimated. Make your expressions for both the forecasts and the forecast error variances operational, by inserting least squares estimates where unknown parameters appear, and use them to produce an operational point forecast and an operational 90% interval forecast for $y_{T+2,T}$.

14. Forecast error variances in models with estimated parameters.

    As we've seen, computing forecast error variances that acknowledge parameter estimation uncertainty is very difficult; that's one reason why we've ignored it. We've learned a number of lessons about optimal forecasts while ignoring parameter estimation uncertainty, such as:

    a. Forecast error variance grows as the forecast horizon lengthens.

    b. In covariance stationary environments, the forecast error variance approaches the (finite) unconditional variance as the horizon grows.

    Such lessons provide valuable insight and intuition regarding the workings of forecasting models and provide a useful benchmark for assessing actual forecasts. They sometimes need modification, however, when parameter estimation uncertainty is acknowledged. For example, in models with estimated parameters:

    a. Forecast error variance needn't grow monotonically with horizon. Typically we *expect* forecast error variance to increase monotonically with horizon, but it doesn't *have* to.

    b. Even in covariance stationary environments, the forecast error variance needn't converge to the unconditional variance as the forecast horizon lengthens; instead, it may grow without bound. Consider, for example, forecasting a series that's just a stationary $AR(1)$ process around a linear trend. With known parameters, the point forecast will converge to the trend as the horizon grows, and the forecast error variance will converge to the unconditional variance of the $AR(1)$ process. With estimated parameters, however, if the estimated trend parameters are even the slightest bit different from the true values (as they almost surely will be, due to sampling variation), that error will be magnified as the horizon grows, so the forecast error variance will grow.

Thus, results derived under the assumption of known parameters should be viewed as a benchmark to guide our intuition, rather than as precise rules.

15. Direct vs. indirect autoregressive forecasts.

6.10 Notes