Autoregressive Moving Average Models

In: Introduction to Time Series Analysis

By: Mark Pickup

Pub. Date: 2015
Access Date: September 23, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781452282015
Online ISBN: 9781483390857
DOI: https://dx.doi.org/10.4135/9781483390857
Print pages: 113-164

Autoregressive Moving Average Models

In Chapters 3 and 4, we covered time series models that can be estimated using ordinary least squares (OLS). You were also introduced to models that are commonly estimated using maximum likelihood—the static model with AR(1) (autoregressive process of order 1) errors, the autoregressive conditional heteroskedasticity (ARCH) model, and the moving average (MA) model. In this chapter, we move on to the autoregressive moving average (ARMA) model and the Box-Jenkins approach to building such models. The chapter continues with a discussion of including exogenous regressors in our model for the purposes of estimating the magnitude of their effects and hypothesis testing. This includes a short discussion on transfer functions and intervention analysis. The chapter concludes with a discussion of an extension to ARCH models—generalized autoregressive conditional heteroskedasticity (GARCH) models.

5.1 Autoregressive Moving Average (ARMA) Models

So far, we have examined the static, finite distributed lag (FDL), lagged dependent variable (LDV), autoregressive distributed lag (ADL), ARCH, and MA models. Let us consider yet another type of time series data-generating process that we may choose to model and identify its stationarity conditions. Combining an AR(*p*) process with an MA*(q)* process (moving average process of order *q)* produces an ARMA data-generating process:

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=0}^{q} \phi_j \varepsilon_{t-j}. \qquad (5.1.1)$$

An ARMA process with a *p*-order autoregressive process and a *q*-order moving average process is denoted as ARMA(*p*,*q*). We can summarize the necessary and sufficient conditions for the stationarity of an ARMA*(p,q)* process as follows:

a. The process must have started an infinitely long time ago or must have immediately begun in equilibrium.
b. The autoregressive component of the time series process is stable (e.g., for ARMA(1,*q*), $|\alpha 1| < 1$).
c. The process cannot contain structural breaks, trending, or periodicity.
d. *q* must be finite.

These are really just a combination of the necessary and sufficient conditions for the stationarity of the AR(*p*) and MA(*q*) processes.

ARMA data models are used to model the time series dynamics in data that are suspected to be generated by processes of this sort. In doing so, we control for those dynamics that violate the assumptions of exogeneity and no serial correlation. Recall from Chapter 4 how including an autoregressive component may solve

problems of endogeneity. Exogenous regressors (and their lags) can also be included in ARMA data models. To see how this is done, first transform the ARMA model as follows. The ARMA*(p,q)* model

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t \qquad (5.1.2)$$

can also be written as

$$y_t = \beta_0 + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t, \qquad (5.1.3)$$

where the first equation is called the structural component and the second is called the disturbance component. Exogenous regressors and their lags are included in the structural component:

$$y_t = \beta_0 + \sum_{k=1} \beta_k x_k + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t. \qquad (5.1.4)$$

Sometimes ARMA models with exogenous regressors—x_ks—are called ARMAX models, but we will not be making the distinction.

We can demonstrate the equivalence of Equations 5.1.2 and 5.1.3 for the ARMA(1,1) data-generating process as follows. The representation described by Equation 5.1.2 is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \phi_1 \varepsilon_{t-1} + \varepsilon_t. \qquad (5.1.5)$$

Let $\mu_t = y_t - \dfrac{\alpha_0}{1-\alpha_1}$; therefore,

$$y_t = \mu_t + \frac{\alpha_0}{1-\alpha_1}. \qquad (5.1.6)$$

$$\text{And } y_{t-1} = \mu_{t-1} + \frac{\alpha_0}{1-\alpha_1}. \qquad (5.1.7)$$

Insert Equations 5.1.6 and 5.1.7 into Equation 5.1.5:

$$\mu_t + \frac{\alpha_0}{1-\alpha_1} = \alpha_0 + \alpha_1\left(\mu_{t-1} + \frac{\alpha_0}{1-\alpha_1}\right) + \phi_1\varepsilon_{t-1} + \varepsilon_t,$$

$$\mu_t = -\frac{\alpha_0}{1-\alpha_1} + \alpha_0 + \alpha_1\frac{\alpha_0}{1-\alpha_1} + \alpha_1\mu_{t-1} + \phi_1\varepsilon_{t-1} + \varepsilon_t.$$

Note that $-\frac{\alpha_0}{1-\alpha_1} + \alpha_0 + \alpha_1\frac{\alpha_0}{1-\alpha_1} = 0$, and so

$$\mu_t = \alpha_1\mu_{t-1} + \phi_1\varepsilon_{t-1} + \varepsilon_t. \qquad (5.1.8)$$

This is the second component of the representation described by Equation 5.1.3. Next, let $\beta_0 = \frac{\alpha_0}{1-\alpha_1}$; then, Equation 5.1.6 becomes

$$y_t = \beta_0 + \mu_t. \qquad (5.1.9)$$

This is the first component of the representation described by Equation 5.1.3.

One of the greatest difficulties of modelling a time series process with an ARMA data model is that there are commonly many alternative ARMA models that capture the same data-generating process dynamics. There is no real way of determining which model reflects the true data-generating process. Box and Jenkins (1976), therefore, proposed a method for selecting an appropriate ARMA model given the data based on a certain set of goals.

5.2 Box-Jenkins Approach to ARMA Models

Having now introduced the ARMA model, we review the Box-Jenkins approach for specifying the model. First, we need to discuss notation. When we introduced autoregressive processes, we defined an AR(2) process as

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t. \qquad (5.2.1)$$

The following data-generating process is also perfectly possible:

$$y_t = \alpha_2 y_{t-2} + \varepsilon_t. \qquad (5.2.2)$$

This process contains the second lag of the dependent variable but not the first. We need some way of distinguishing between these types of models in our notation. When it is not clear from the context, we shall distinguish these as follows:

$$y_t = \alpha_2 y_{t-2} + \varepsilon_t \quad \mathrm{AR}(p=2),$$

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t \quad \mathrm{AR}\left(p=1,2\right).$$

This will apply to any possible lag except the first. Since AR(p = 1) = AR(1), we shall just use AR(1).

Furthermore, when discussing moving average processes, we defined an MA(2) process as

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2}. \quad (5.2.3)$$

The following data-generating process is also perfectly possible:

$$y_t = \varepsilon_t + \phi_2 \varepsilon_{t-2}. \quad (5.2.4)$$

This is the process assumed in our example of an MA model earlier, in Chapter 4. When it is not clear from the context, we shall distinguish these as follows:

$$y_t = \varepsilon_t + \phi_2 \varepsilon_{t-2} \mathrm{MA}(q = 2).$$

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \mathrm{MA}(q = 1, 2).$$

Again, this will apply to any possible lag except the first. Since MA(q = 1) = MA(1), we shall just use MA(1).

We are now in a position to discuss the Box-Jenkins approach. George Box and Gwilym Jenkins (1976) proposed a three-stage approach to selecting the appropriate number of autoregressive and moving average components to include in a model:

1. Identification
2. Estimation
3. Diagnostic checking

To begin, it must be determined whether the data being modelled are stationary. A simple plot of the data can reveal nonstationarity produced by elements such as trending, periodicity, and structural breaks. If the data are not stationary, they are transformed so that they are—most commonly, the data are first differenced or seasonally differenced (discussed in Chapter 3 and to be discussed further in Chapter 6) or detrended/deseasonalized (as discussed in Chapter 3).

Box-Jenkins Approach: Identification

Once the data are determined to be stationary or transformed accordingly, the first step of the Box-Jenkins approach is to identify the appropriate autoregressive and moving average components to include in the model. The Box-Jenkins approach to doing this makes use of the autocorrelation and partial correlation functions of the observed time series.

In Chapter 2, we discussed the autocorrelations of a time series process and gave an example of an estimated autocorrelation function (ACF). Recall that under the assumption of stationarity, the autocorrelation at lag s is defined as follows:

$$\rho_s \equiv \frac{E\left(y_t - \mu_y\right)E\left(y_{t-s} - \mu_y\right)}{E\left(y_t - \mu_y\right)^2} \qquad (5.2.5)$$

The autocorrelations are denoted as ρ_s, where s indicates the "order" of the autocorrelation; for example, $\rho_1 = \text{Corr}(y_t, y_{t-1})$, $\rho_2 = \text{Corr}(y_t, y_{t-2})$, …, $\rho_s = \text{Corr}(y_t, y_{t-s})$. Also, recall that the ACF is a plot of the autocorrelations for a range of lags, beginning with a lag of 1. Plotting ρ_s against s is called the ACF or correlogram. Calculating the autocorrelations requires knowledge of the series mean and variance. Recall from Chapter 2 that these are not known for the true data-generating process but can be estimated from the sample data, assuming that the series is stationary or has been transformed accordingly:

$$\bar{\rho}_s = \frac{\sum_{t=s+1}^{T}(y_t - \bar{y})(y_{t-s} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \qquad (5.2.6)$$

These estimated autocorrelations are useful in identifying the order of the autoregressive and moving average components of an ARMA process. The pattern of the estimated autocorrelations in the ACF can tell us something about the time series process. For example, the AR(1) data-generating process is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t. \qquad (5.2.7)$$

For this time series process, $\rho_1 = \alpha_1$ and ρ_2, the correlation between y_t and y_{t-2}, is equal to the correlation between y_t and y_{t-1} multiplied by the correlation between y_{t-1} *and* y_{t-2}. From Equation 5.2.7,

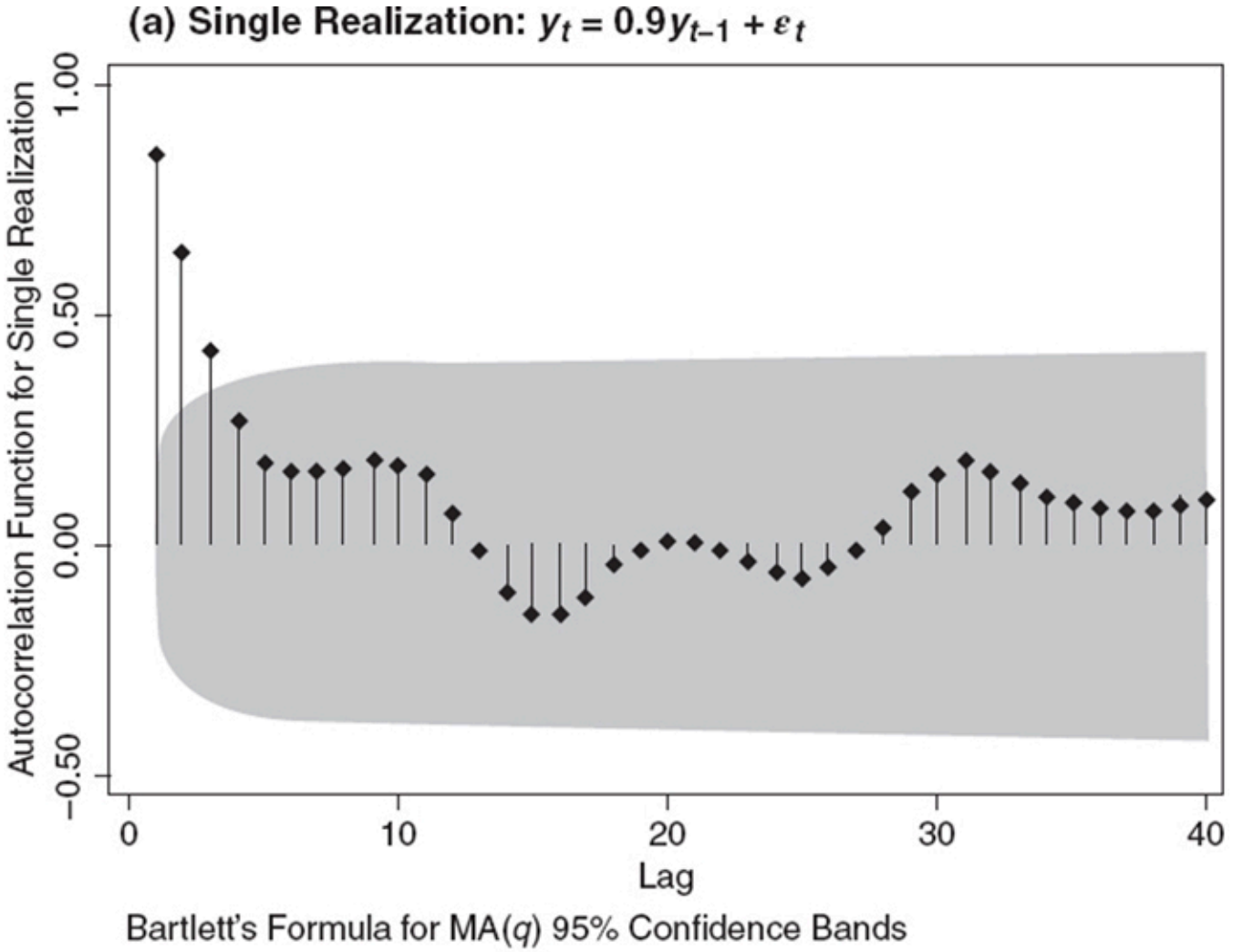
$$y_{t-1} = \alpha_0 + \alpha_1 y_{t-2} + \varepsilon_{t-1}.$$

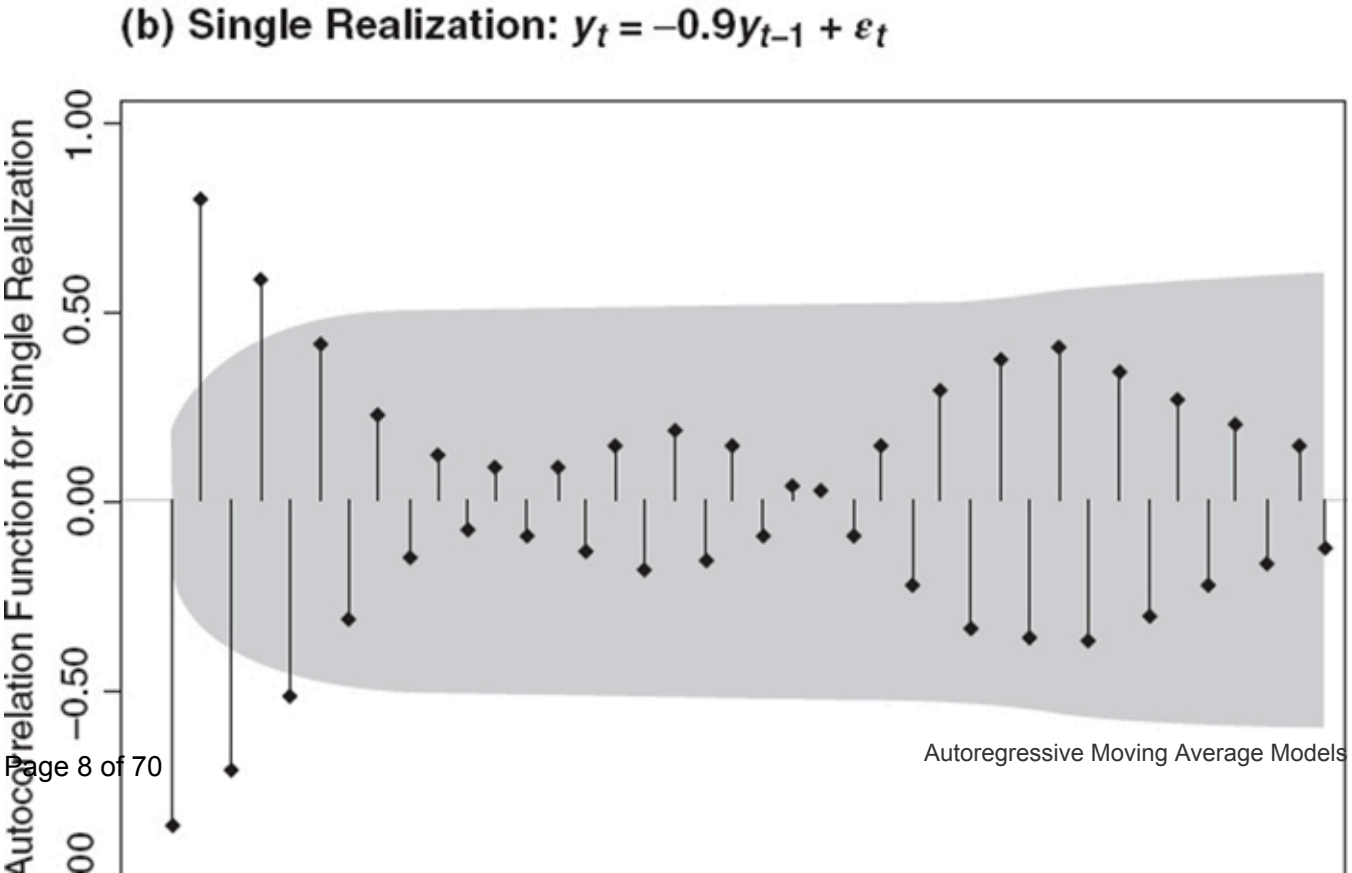
Therefore, $\rho_2 = \alpha_1 \times \alpha_1 = \alpha_1^2$. Generally, the AR(1) time series process has the following autocorrelations for $s \geq 1$:

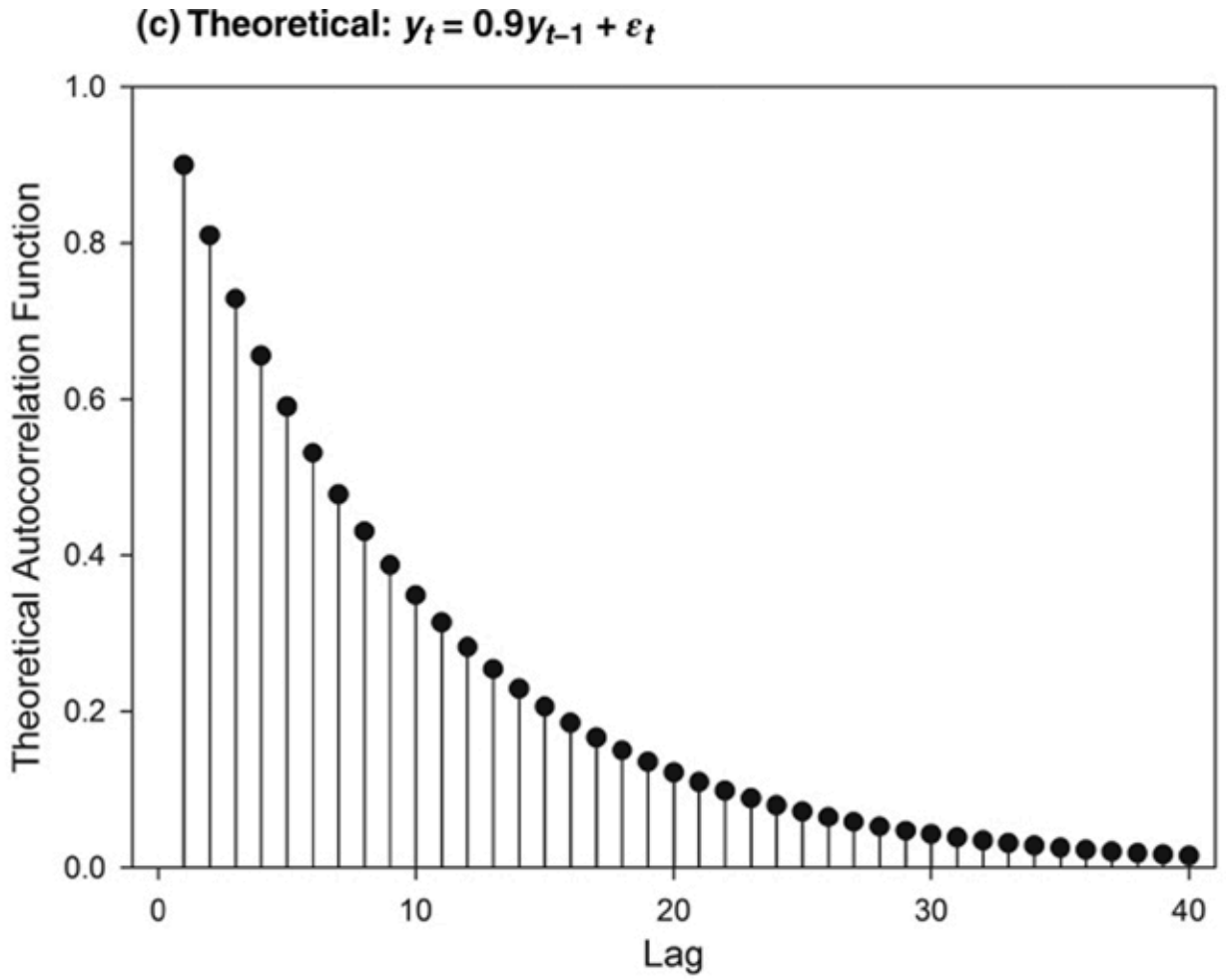
$$\rho_s = \alpha_1^s. \qquad (5.2.8)$$

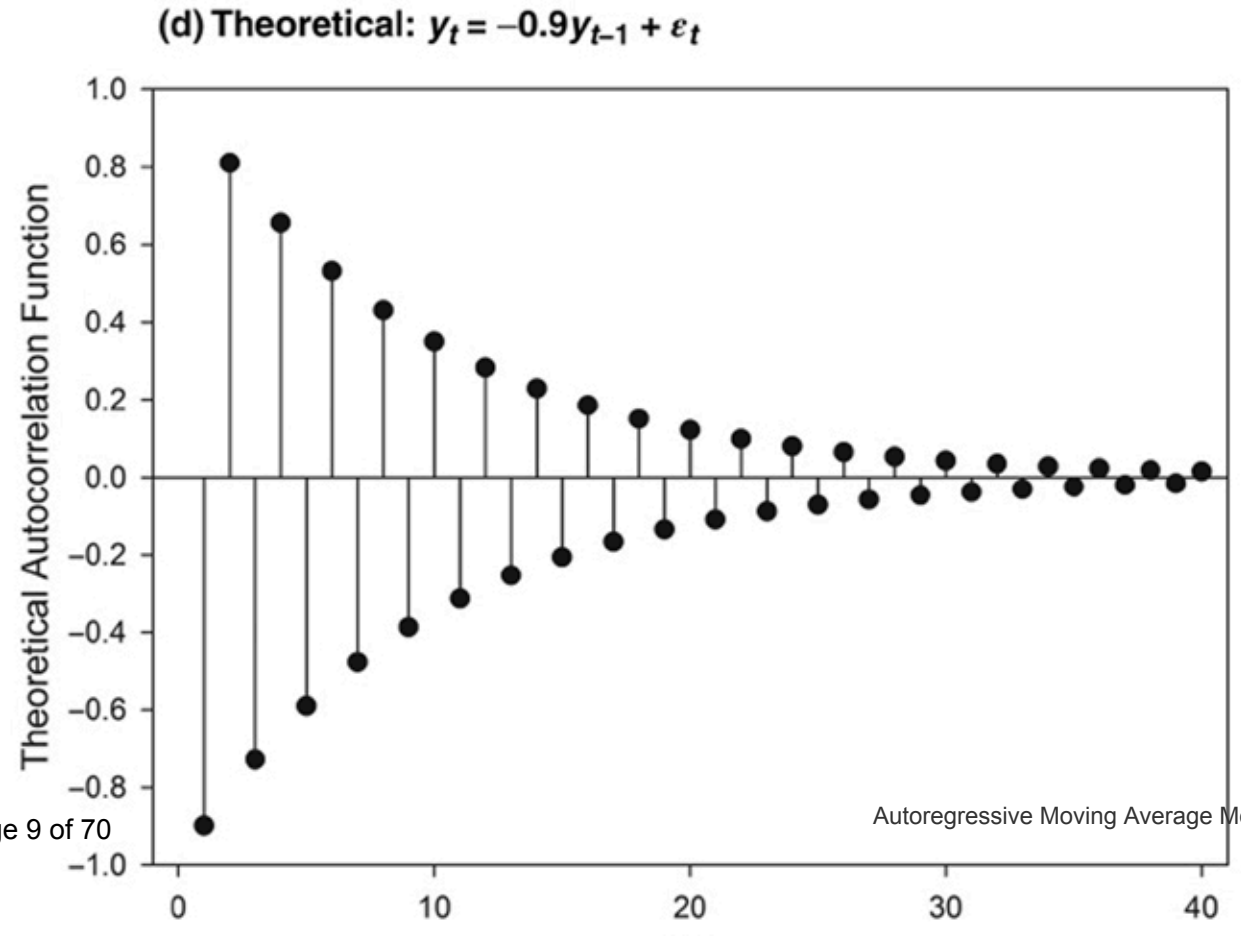
The process is assumed to be stationary, $|\alpha_1| < 1$, and so the ACF will converge to 0 exponentially, either directly if α_1 is positive or through a dampened oscillatory path if α_1 is negative. For example, consider the ACFs for data sampled from the two autoregressive data-generating processes shown in Figure 5.1. For each data-generating process, the figure includes both the theoretical ACF and the ACF for a single realization.

Figure 5.1 Autocorrelation Functions for Autoregressive Processes









If we were to estimate the ACF for a time series with an unknown data-generating process and it demonstrated one of these versions of exponential convergence to 0, we might identify the time series process as AR(1).

The statistical significance of the autocorrelations can be tested based on whether they exceed 2 standard deviations, where the standard deviation is estimated from the sample series:

$$\mathrm{Var}(\rho_s) = T^{-1}, \text{ for } s = 1.$$

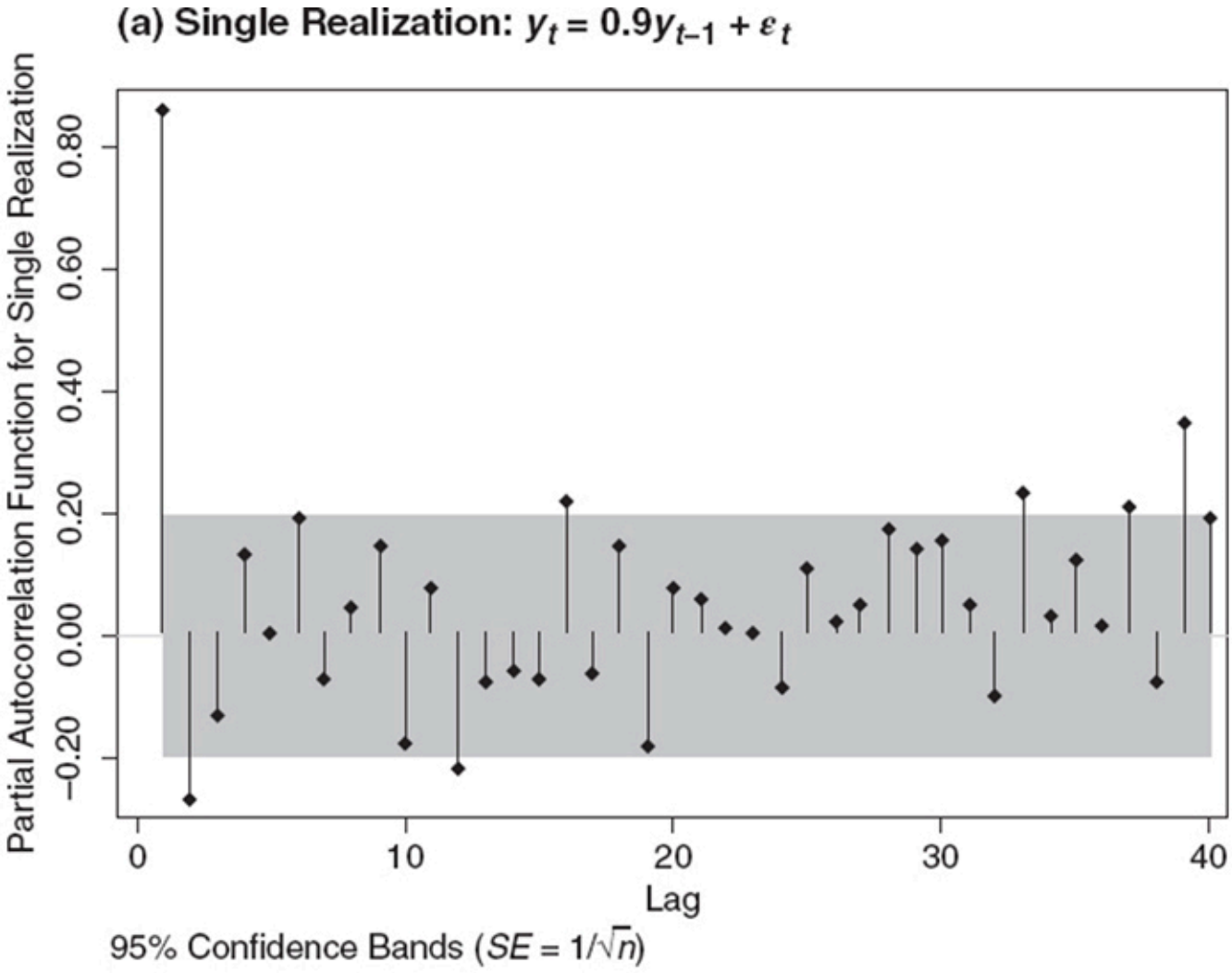
$$\mathrm{Var}\left(\rho_s\right) = T^{-1}\left(1 + 2\sum_{j=1}^{s-1}\rho_j^2\right), \text{ for } s > 1. \qquad (5.2.9)$$

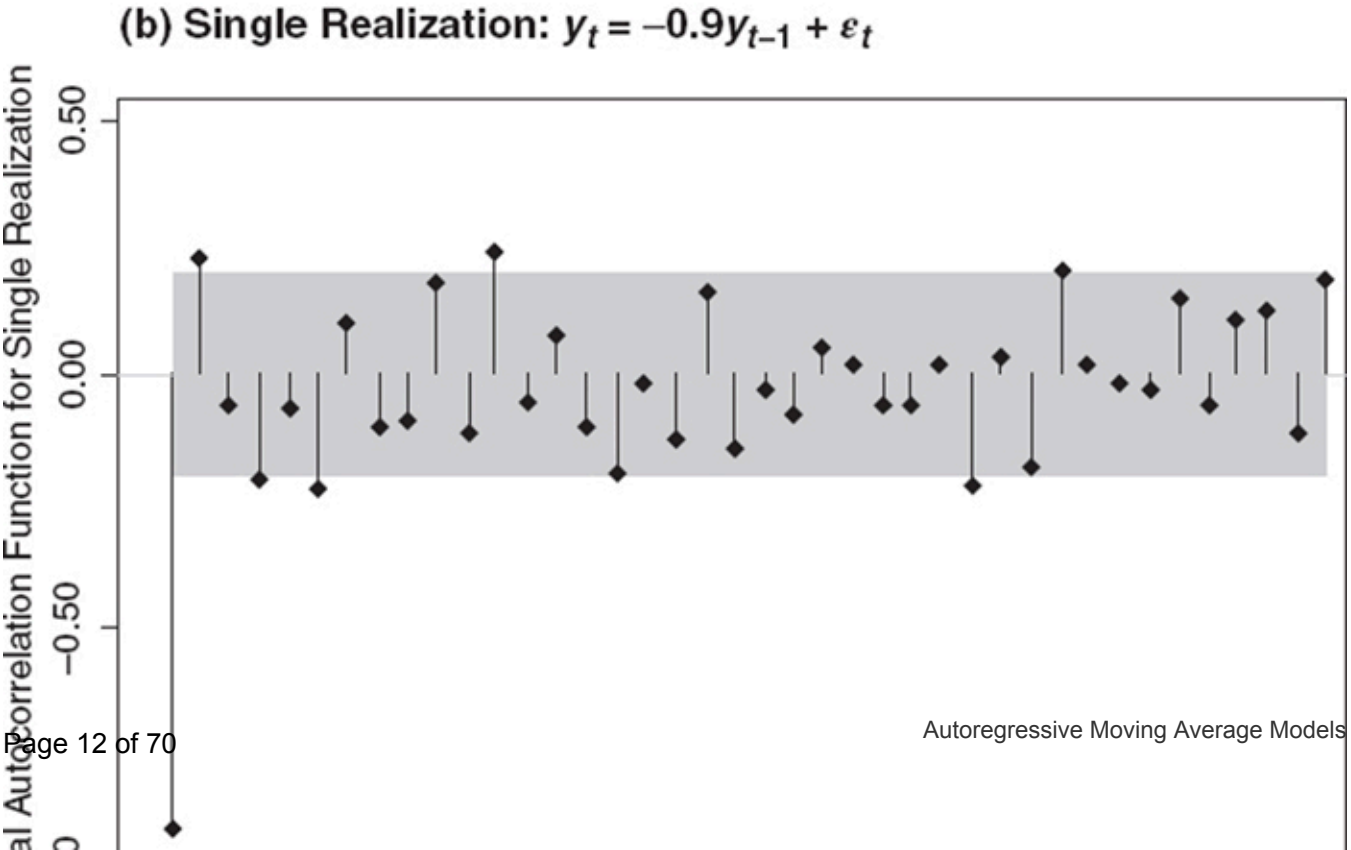
Again, we do not know ρ_s, but we can use their estimates $(\hat{\rho}_s)$ in the above equation. The confidence envelope for the ACF in Figure 5.1 is calculated using these estimates of the variance. The null hypothesis of the test is that the autocorrelation is not significantly different from 0. This is not really intended to test any individual autocorrelation. The intent is to test whether the autocorrelations are, as a whole, significantly different from what we would expect from a white noise process. In a white noise process, only 1 in 20 will be significant at the 0.05 significance level.

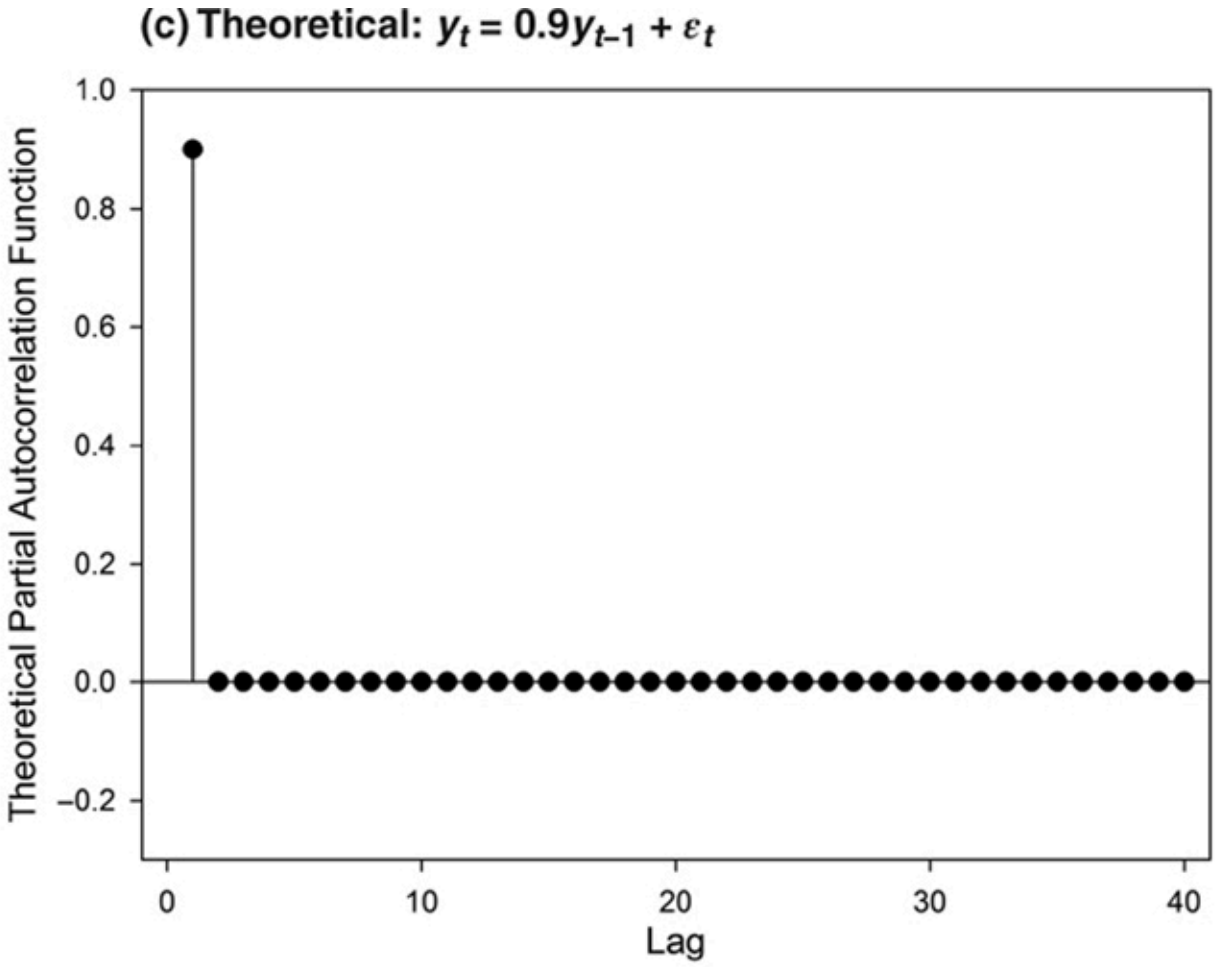
For an AR(1) process other than ρ_1, the autocorrelations are *indirect* correlations. By this, we mean that there is a correlation between y_t and y_{t-2} only because y_t and y_{t-1} are correlated and y_{t-1} and y_{t-2} are correlated. It is also possible to calculate the correlation between, for example, y_t and y_{t-2} controlling for, or partialing out, the effects of the intervening values of y_{t-1}. This can be done by regressing y_t on y_{t-1} and y_{t-2}, and using the coefficient on y_{t-2} as our estimate of the partial autocorrelation φ_2. Plotting these partial correlations, φ_s, against s is the partial autocorrelation function (PACF). We can estimate the PACF and employ this in the identification of an unknown time series process.

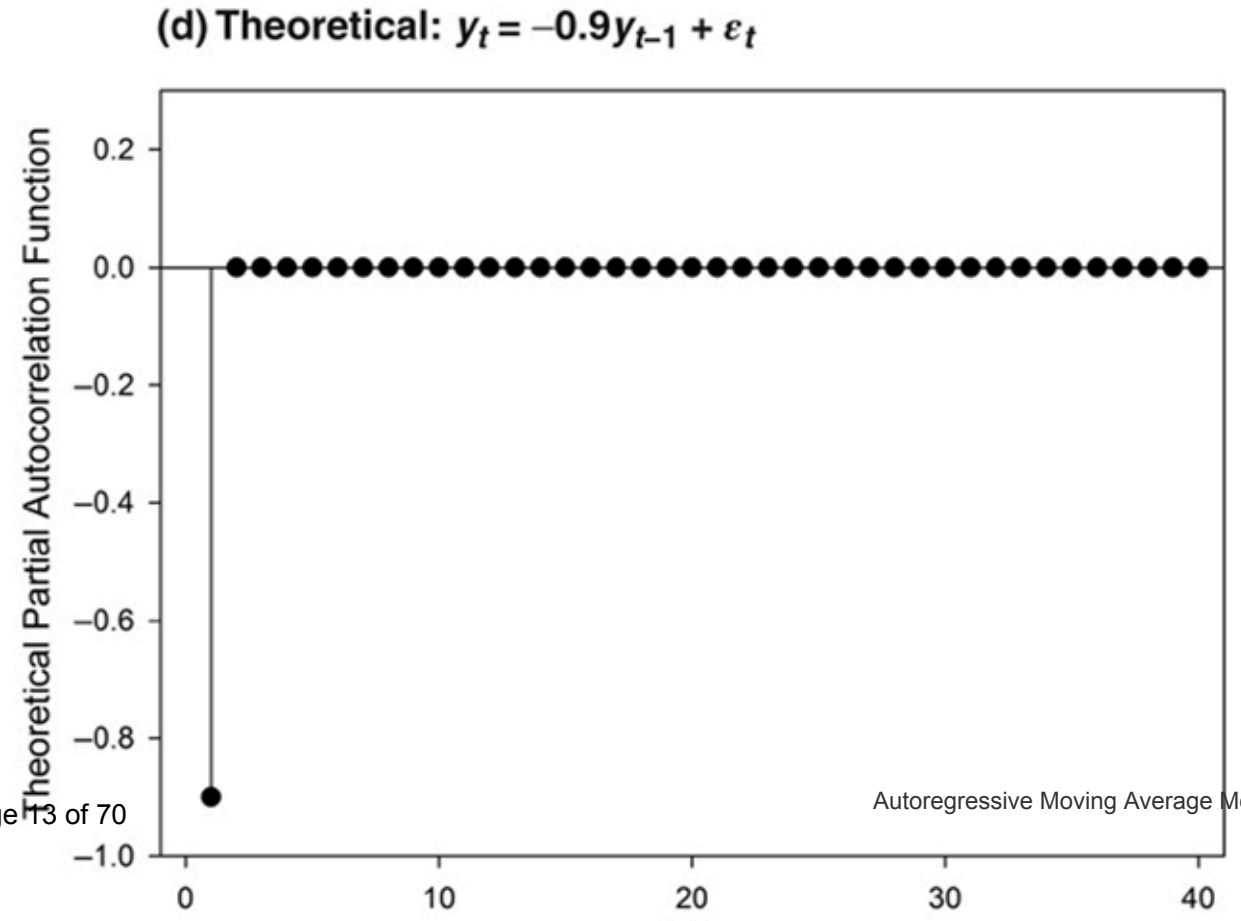
For an AR(1) process $\varphi_1 = \rho_1$ and for $s > 1$, $\varphi_s = 0$. Consider the PACFs for the same data for which we calculated the ACFs, as shown in Figure 5.2. The variance used to calculate the confidence envelopes for these partial autocorrelations is $\mathrm{Var}(\varphi_s) = T^{-1}$.

Figure 5.2 Partial Autocorrelation Functions for Autoregressive Processes









A PACF with this pattern would help us confirm that our data were generated by an AR(1) process. Generally, the patterns exhibited in the ACF and PACF of our time series data can be used to identify the process that generated them (Box & Pierce, 1970). An ACF and PACF, like the ones we have just observed, would suggest that the data-generating process contains a single autoregressive lag of order 1.

Things get a little more complicated for higher-order autoregressive processes. For example, consider an AR(p = 1, 2) process:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t. \qquad (5.2.10)$$

The autocorrelations and partial autocorrelations will be a function of the parameters α_1 and α_2.

$$\rho_1 = \frac{\alpha_1}{(1-\alpha_2)} \qquad \varphi_1 = \alpha_1.$$

$$\rho_2 = \alpha_1 \times \rho_1 + \alpha_2 \quad \varphi_2 = \alpha_2 \qquad (5.2.11)$$

For $s > 2$, $\rho_s = \alpha_1 \rho_{s-1} + \alpha_2 \rho_{s-2}$ and $\varphi_s = 0$.

Consider an AR(p = 1, 2) data-generating process with $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$:

$$y_t = 0.5 y_{t-1} + 0.3 y_{t-2} + \varepsilon_t.$$

$$\rho_1 = \frac{0.5}{(1-0.3)} = 0.7 \qquad \varphi_1 = 0.5$$

$$\rho_2 = 0.5 \times 0.7 + 0.3 = 0.65 \quad \varphi_2 = 0.3.$$

For $s > 2$, $\varphi_s = 0$, and the ACF decays at a rate slower than that of an AR(1) process.

We have seen what the ACF and PACF look like for an autoregressive process, but the time series process may also contain moving average components. Consider the MA(1) data-generating process:

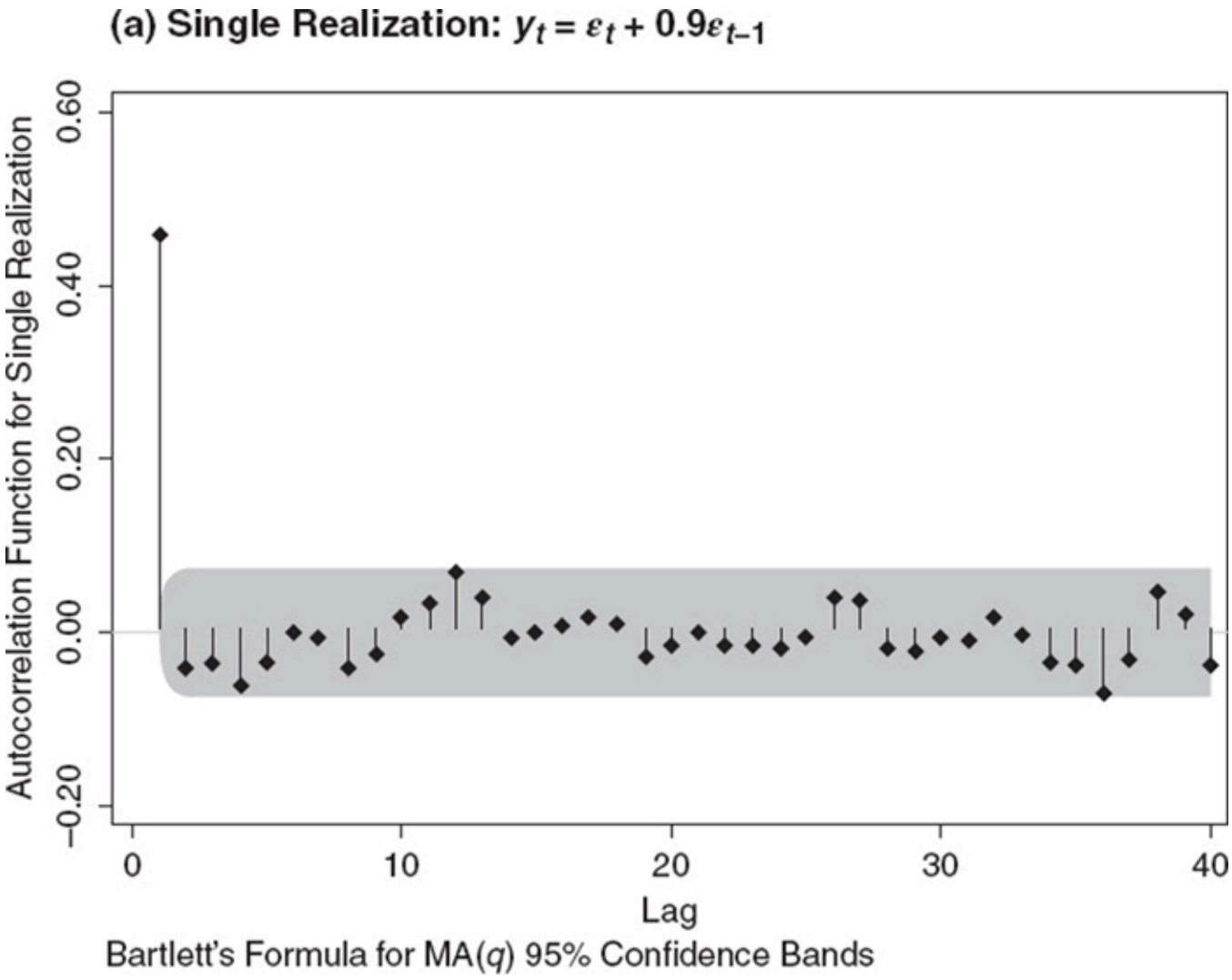
$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1}. \qquad (5.2.12)$$

It can be shown that for the ACF,

$$\rho_1 = \frac{\phi_1}{(1+\phi_1^2)} \text{ and } \rho_s = 0 \forall s > 1. \qquad (5.2.13)$$

It can also be shown that the PACF will decay geometrically if $\varphi_1 < 0$ and will decay through an oscillatory path if φ_1 <i» 0. Consider the ACFs and PACFs for data generated by two MA(1) processes, one with $\varphi_1 = 0.9$ and the other with $\varphi_1 = -0.9$, as depicted in Figures 5.3 and 5.4.

Figure 5.3 Autocorrelation Functions for Moving Average Processes



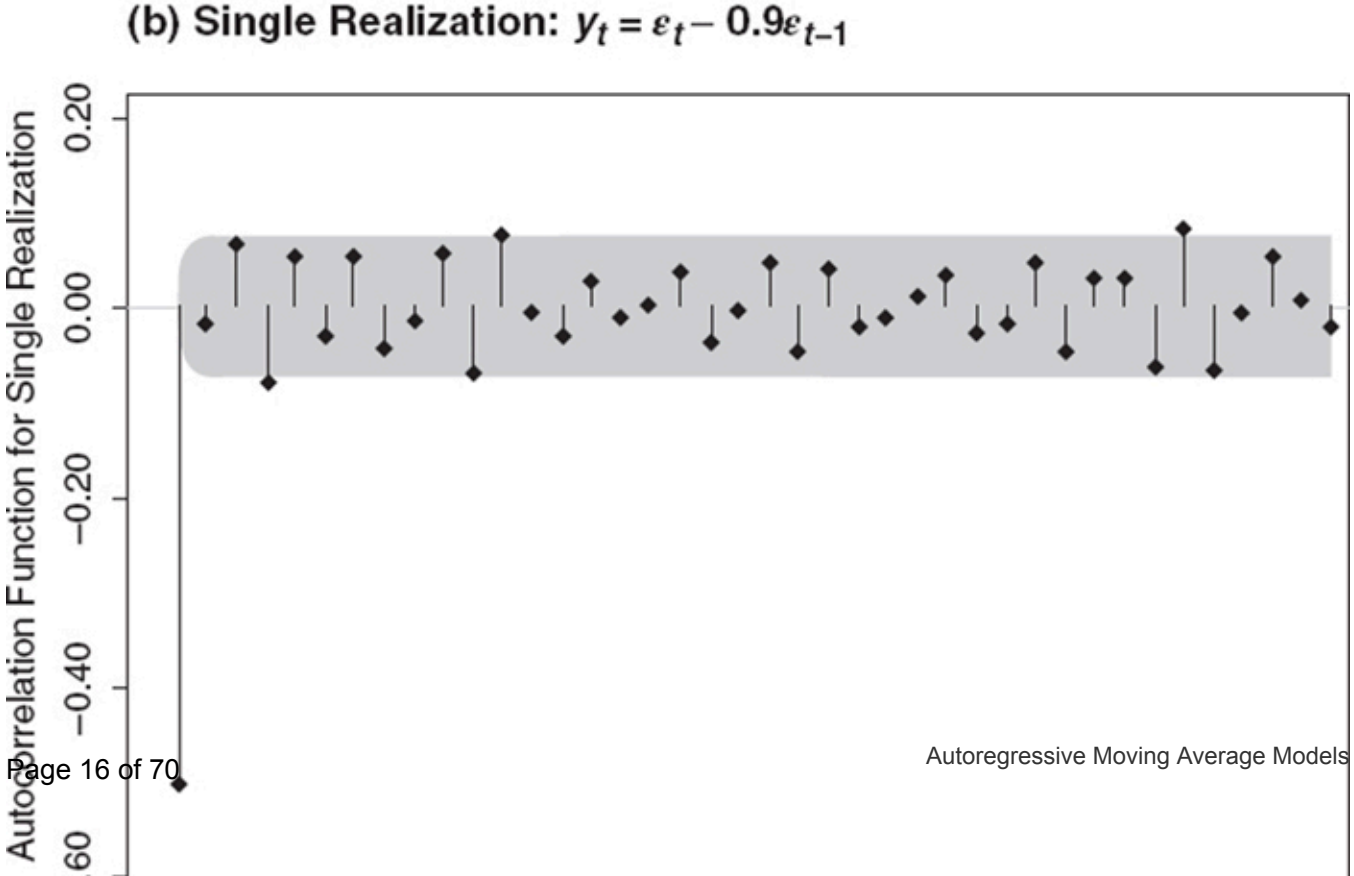
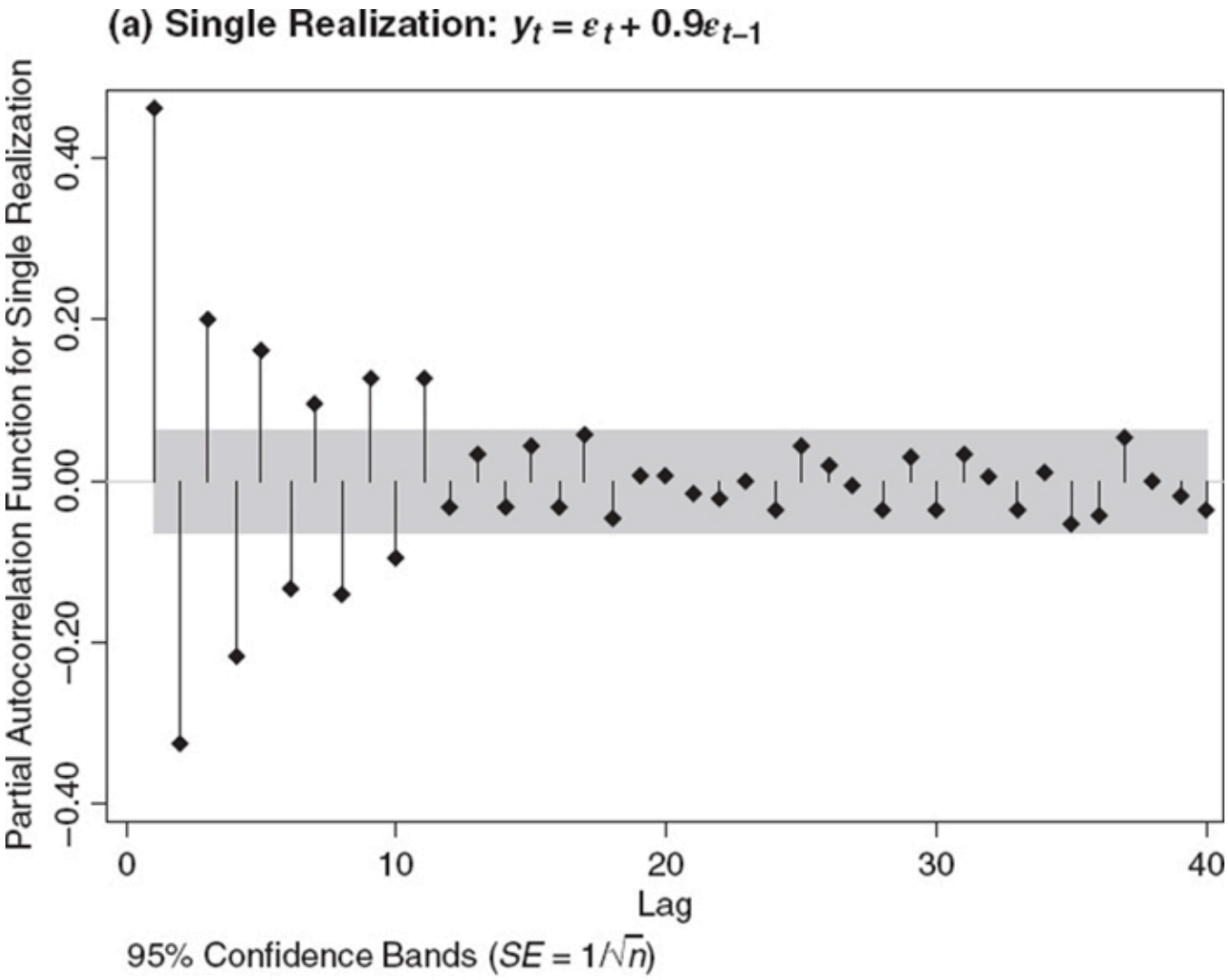
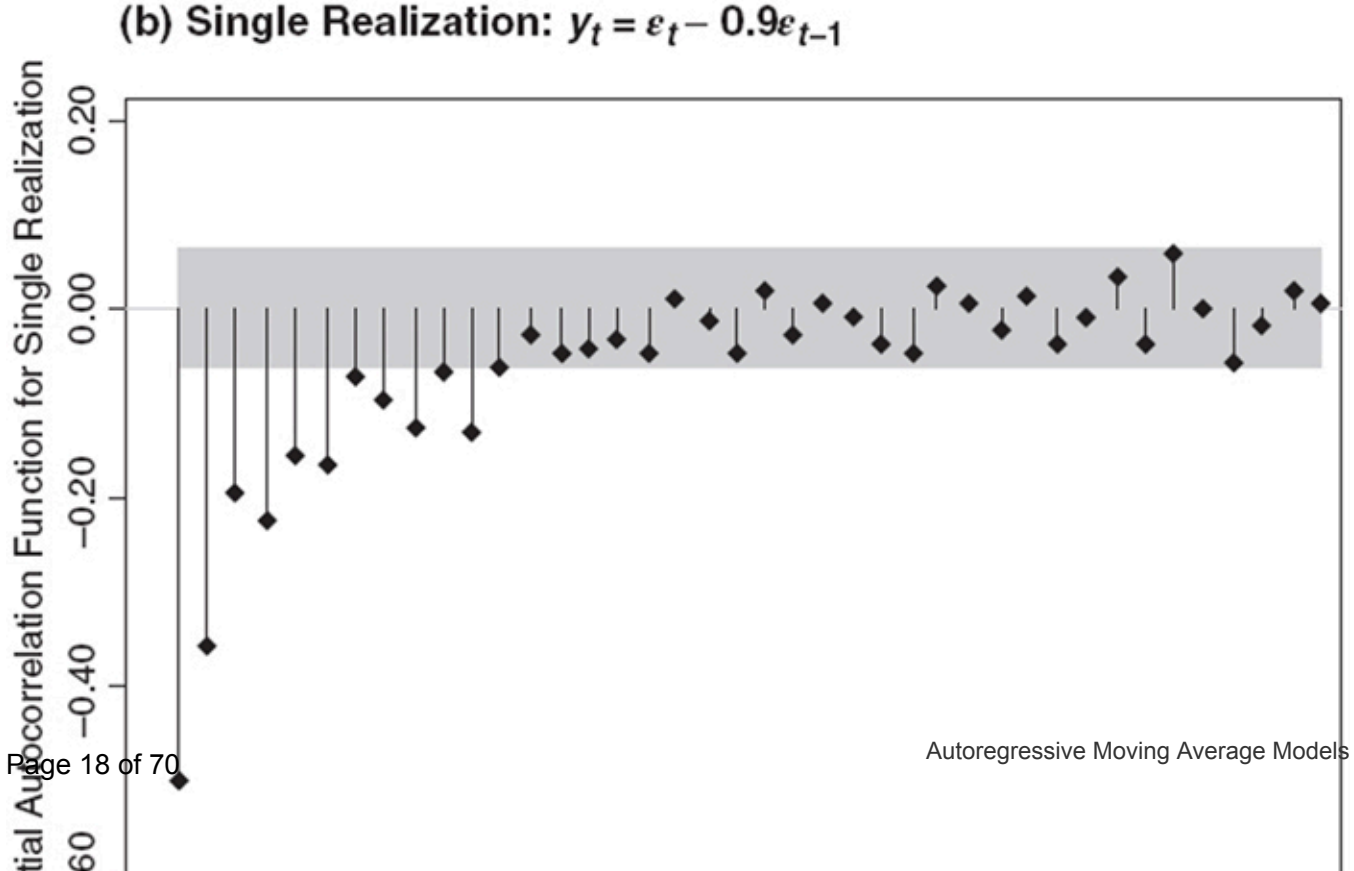


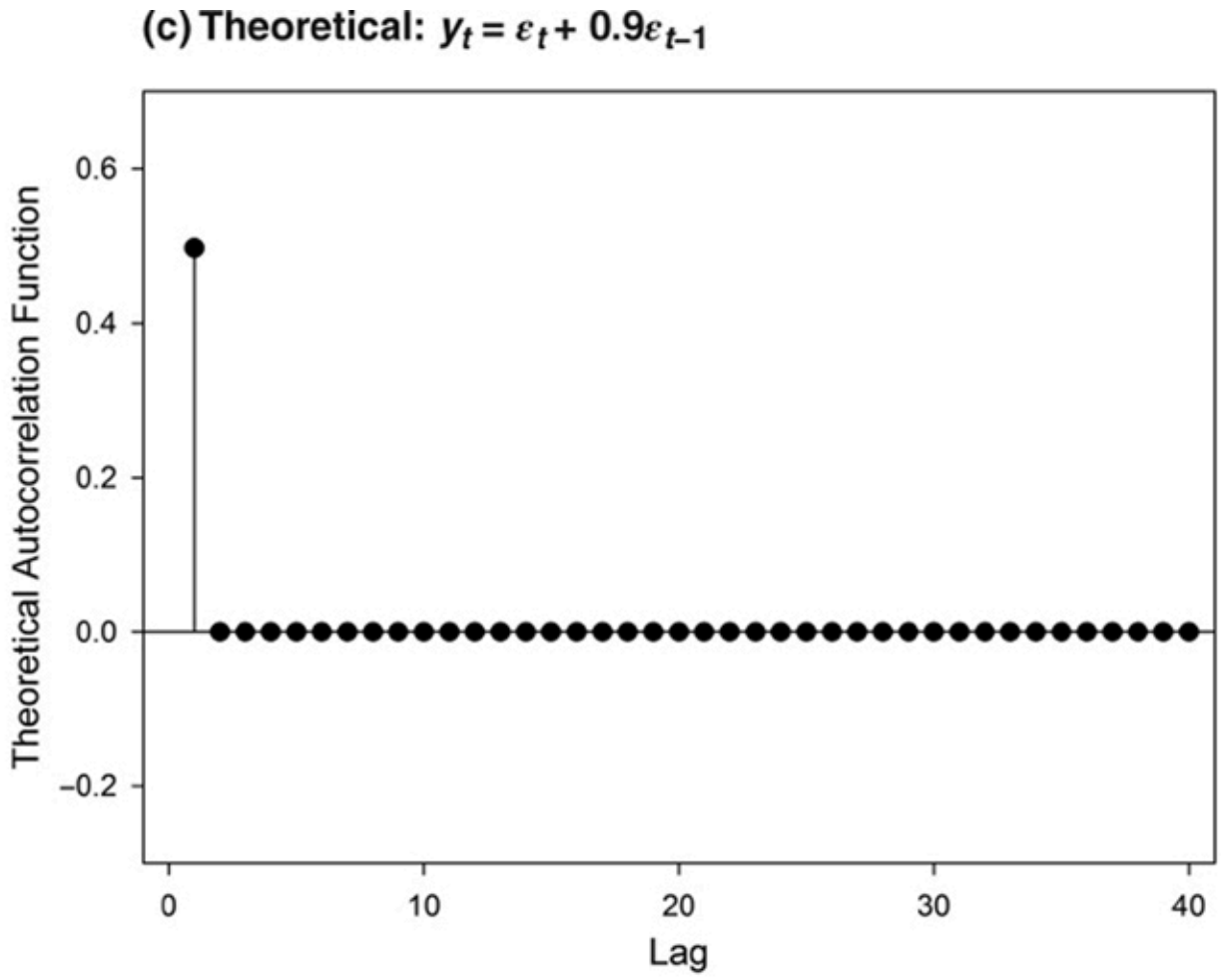
Figure 5.4 Partial Autocorrelation Functions for Moving Average Processes

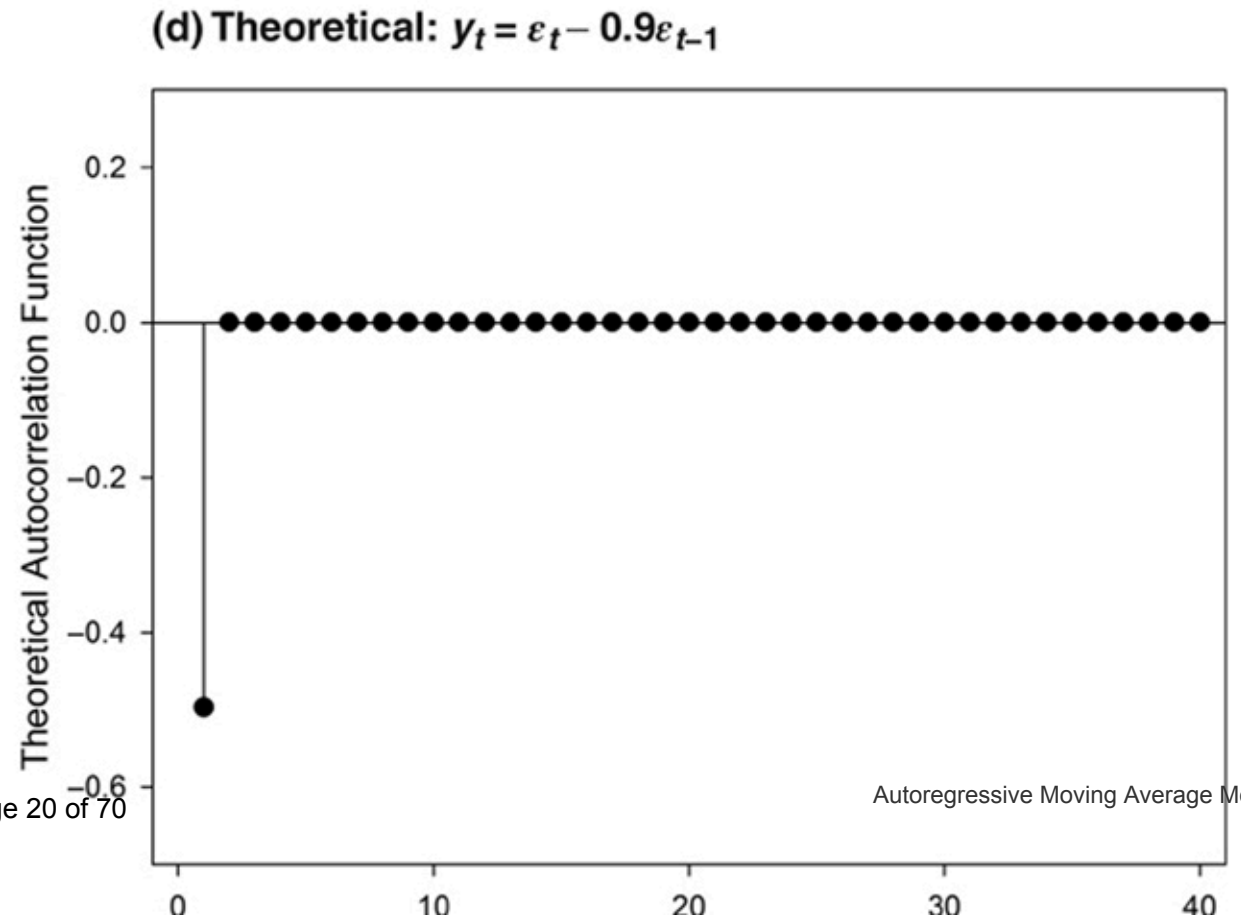


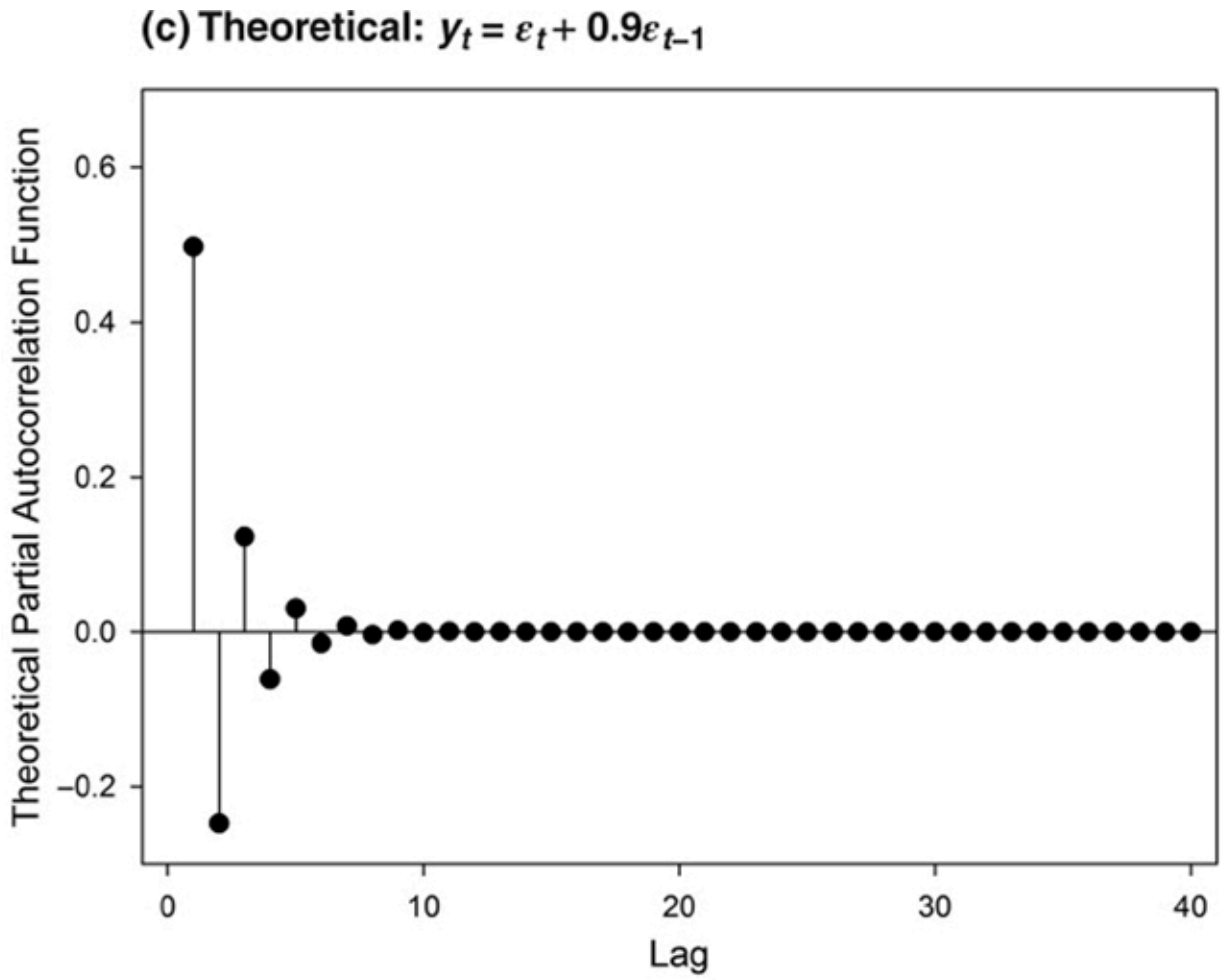


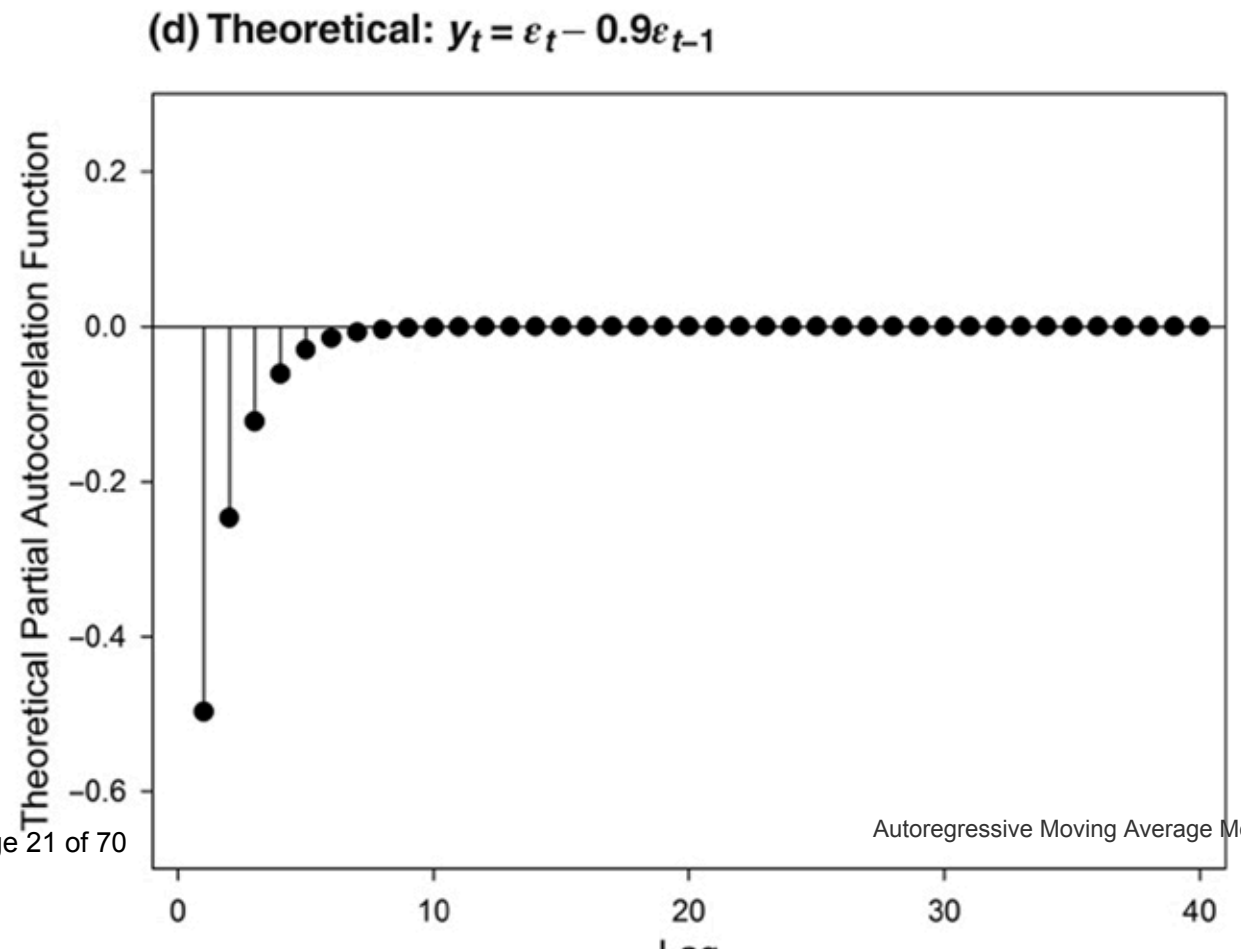
With $\varphi_1 = 0.9$ in the data-generating process, the first autocorrelation is

$$\rho_1 = \frac{\phi_1}{\left(1+\phi_1^2\right)} = \frac{0.9}{\left(1+0.9^2\right)} = 0.5,$$









which is what we observe for $\hat{\rho}_1$ in the ACF. With $\varphi_1 = -0.9$ in the data-generating process, the first autocorrelation is -0.5, which again is what we observe for $\hat{\rho}_1$ in the ACF. As for the PACFs, they decay approximately geometrically—directly with $\varphi_1 = -0.9$ and through an oscillating path when $\varphi_1 = 0.9$.

Generally, the pattern of the ACF and PACF indicates the order of *p* and *q* in the data-generating process. This can be used to inform the order of *p* and *q* required in our ARMA(*p*,*q*) model in order to capture the time series dynamics.

Some basic guidelines for identifying some common processes based on their ACF and PACF are described in Table 5.1 (Enders, 2004, pg. 66).

Table 5.1 Guidelines for Identifying Autoregressive and Moving Average Processes

Table 5.1 Guidelines for Identifying Autoregressive and Moving Average Processes

Process	*ACF*	*PACF*
White noise	All $\rho_s = 0, s \neq 0$.	All $\varphi_s = 0$.
AR($p = P$)	Decay toward zero. Coefficients may oscillate (decay is geometric if $P = 1$).	Spikes at s = P. All $\varphi_s = 0$ for $s > P$.
MA($q = Q$)	Spike at $s = Q$ and $\rho_s = 0$ $\forall\, s \neq Q$.	Decay toward zero (either direct or oscillatory).
ARMA ($p = P$, $q = Q$)	Decay (either direct or oscillatory) beginning at lag Q.	Decay (either direct or oscillatory) beginning after lag P.

NOTE: ACF = autocorrelation function, AR = autoregressive, ARMA = autoregressive moving average, MA = moving average, PACF = partial autocorrelation function.

For complex processes, the correct combination of *p* and *q* may not be clear. In fact, more than one combination may fit the same data-generating process. In selecting the best ARMA data model, Box-Jenkins set out model selection criteria. The goal is to select a data model that accounts for the autocorrelation in the error term (i.e., once we include autoregressive and moving average terms, we are left with residuals that are a white noise process) while following the principle of parsimony. We choose the model that fits the data best, but (a) if two alternative models fit the data approximately equally well, we choose the model with fewer coefficients, and (b) each coefficient should be significantly different from 0 at our chosen significance level—typically 0.05.

When testing how well a model fits the data, a number of goodness-of-fit measures are at our disposal. Some of the most common are the Akaike Information Criterion (AIC) and the Schwartz Bayesian Information Criterion (BIC). These are calculated as follows (Enders, 2004):

$$\text{AIC} = T\ln(\text{sum of squared residuals}) + 2(k),$$

$$\text{BIC} = T\ln(\text{sum of squared residuals}) = ln(T)(k). \qquad (5.2.14)$$

For both formulae, k is the number of parameters in the model including the intercept. Models with smaller (including more negative) AIC and BIC values fit better. The BIC, which applies a greater penalty for additional parameters, will select a more parsimonious model and has better properties when T is large. The AIC, however, may be superior with a small T (Harvey, 1993).

When comparing BIC values from two different models, we can use the rough guidelines given in Table 5.2 (Raftery, 1995) to decide if there is evidence that one fits better than the other.

Table 5.2 Guidelines for Comparing Models Using BIC Values

Difference	*Evidence*
0–2	Weak
2–6	Positive
6–10	Strong
<p»10	Very strong
NOTE: BIC = Schwartz Bayesian Information Criterion.	

We can also use a likelihood ratio (LR) test to compare two nested models.

$$\text{LR} = -2\left(\text{Likelihood for null model} - \text{Likelihood for alternative model}\right).$$

This statistic has a chi-squared distribution, with the degrees of freedom equal to the difference in the number of parameters between the two nested models, and so the usual methods of inference can be used.

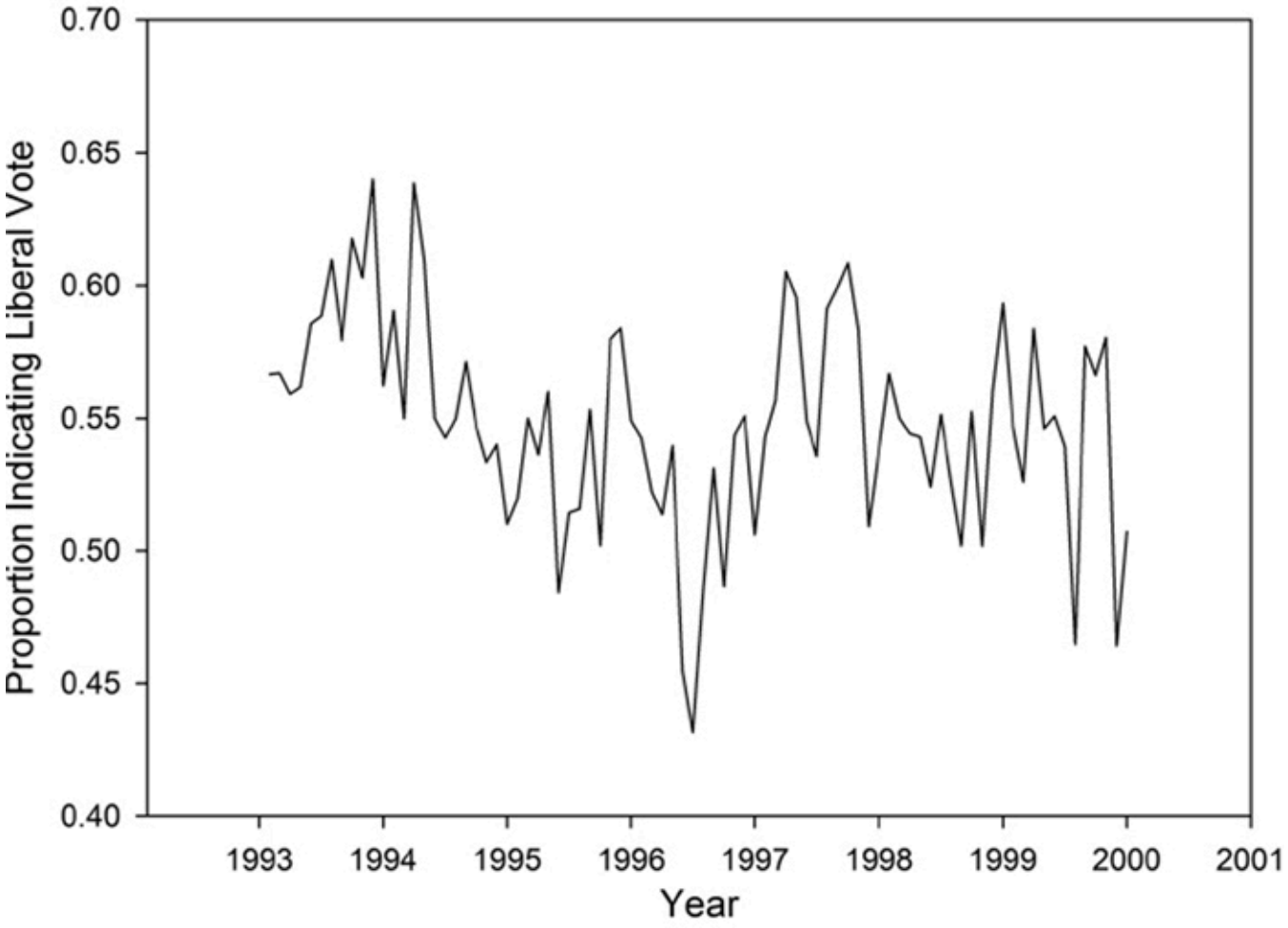
It is important to keep T constant when comparing models. This should be kept in mind as including an additional autoregressive lag, in practice, means eliminating a data point. This data point should also be eliminated for all the models that are being considered.

When comparing models, it is important to always keep in mind that the calculation of the ACF and PACF is based on the assumption that the sequence $\{y_t\}$ is stationary. We should therefore be suspicious of any model that produces estimates of the coefficients that suggest instability (e.g., for an AR(1) model, $|\alpha_1| \geq 1$).

Let us now consider an example of model identification using the Box-Jenkins approach for time series data with an unknown data-generating process. For this example, we shall look at monthly vote intention data for the Canadian federal government (the proportion of survey respondents who indicated that they would vote for the party currently in government if an election were held tomorrow). The period we will examine is

1993 to 2000, during which there was a Liberal federal government (Figure 5.5). We will also want to include exogenous regressors in our model. We are interested in determining whether economic conditions affected vote intention for the governing party (gross domestic product [GDP], inflation, and unemployment)—this will be a classic economic popularity model.

Figure 5.5 Vote Intention for the Liberal Party of Canada, 1993 to 2000



We begin by plotting the vote intention variable: the proportion of respondents indicating that they would vote for the Liberal governing party if an election were held.

Visually, there is some evidence of trending, and (more important) there are theoretical reasons to believe that there is a downward trend. After an initial honeymoon, new Canadian governments generally lose popularity over their term in office as they are forced to make difficult choices that inevitably upset some of those who voted for the governing party. As demonstrated in Chapter 2, we can test for the trend by regressing the vote intention data on a time variable. This produces the results presented in Table 5.3.

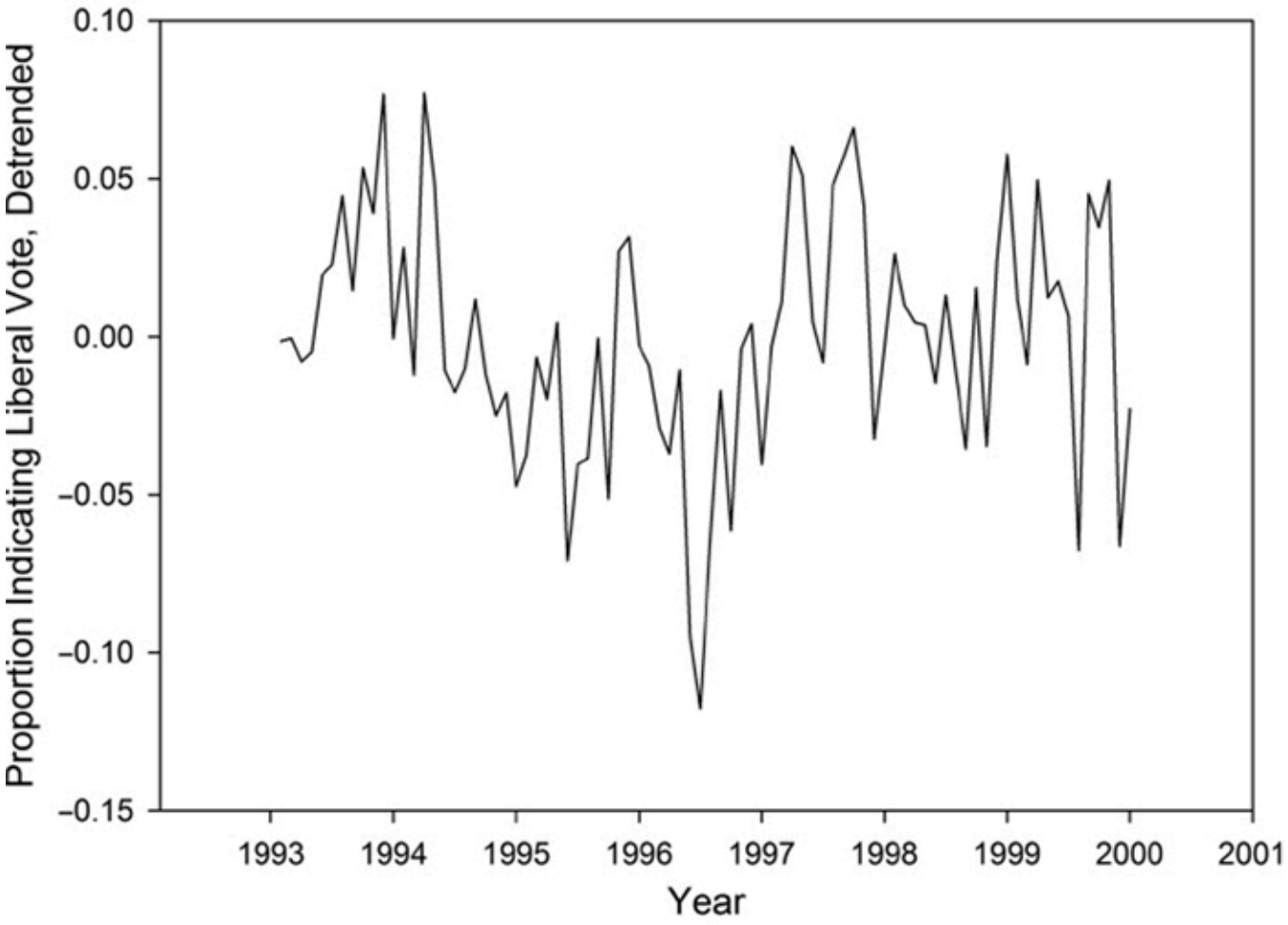
Table 5.3 Trending in Vote Intention for the Incumbent Government in Canada

Vote	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
Trend	–0.00046	0.00017	–2.63	0.010
Intercept	0.57	0.0085	67.902	<0.001

NOTE: $R^2 = 0.078$, $T = 84$; T = number of time points.

We can see from the results that although the magnitude of the trend is small, it is statistically significant at the 0.05 significance level. Next, we may choose to produce a detrended vote intention variable by using the estimated equation to calculate the residuals from the above regression. The resulting time series is the vote intention time series with the trend partialed out, as we can see in the plot of this time series shown in Figure 5.6.

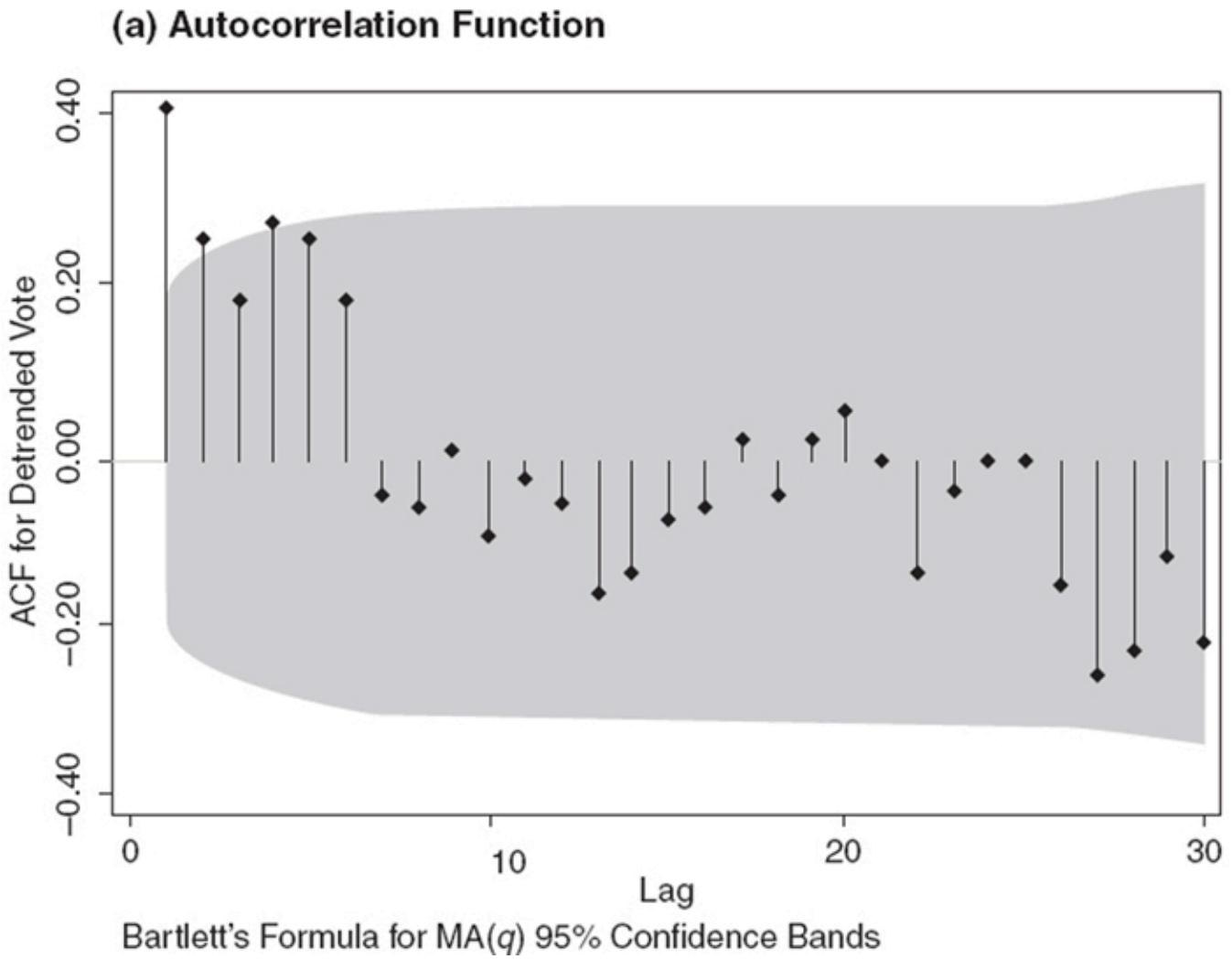
Figure 5.6 Detrended Vote Intention for the Liberal Party of Canada, 1993 to 2000



As we will discuss in Chapter 6, we might ask ourselves whether the original data are integrated (defined in Chapter 2), rather than containing a deterministic linear trend. We will leave this issue for now and assume that they are not and that they are stationary once the deterministic linear trend is removed.

Having addressed the issue of stationarity, we next estimate and examine the ACF and PACF for the detrended vote intention variable. These are displayed in Figure 5.7.

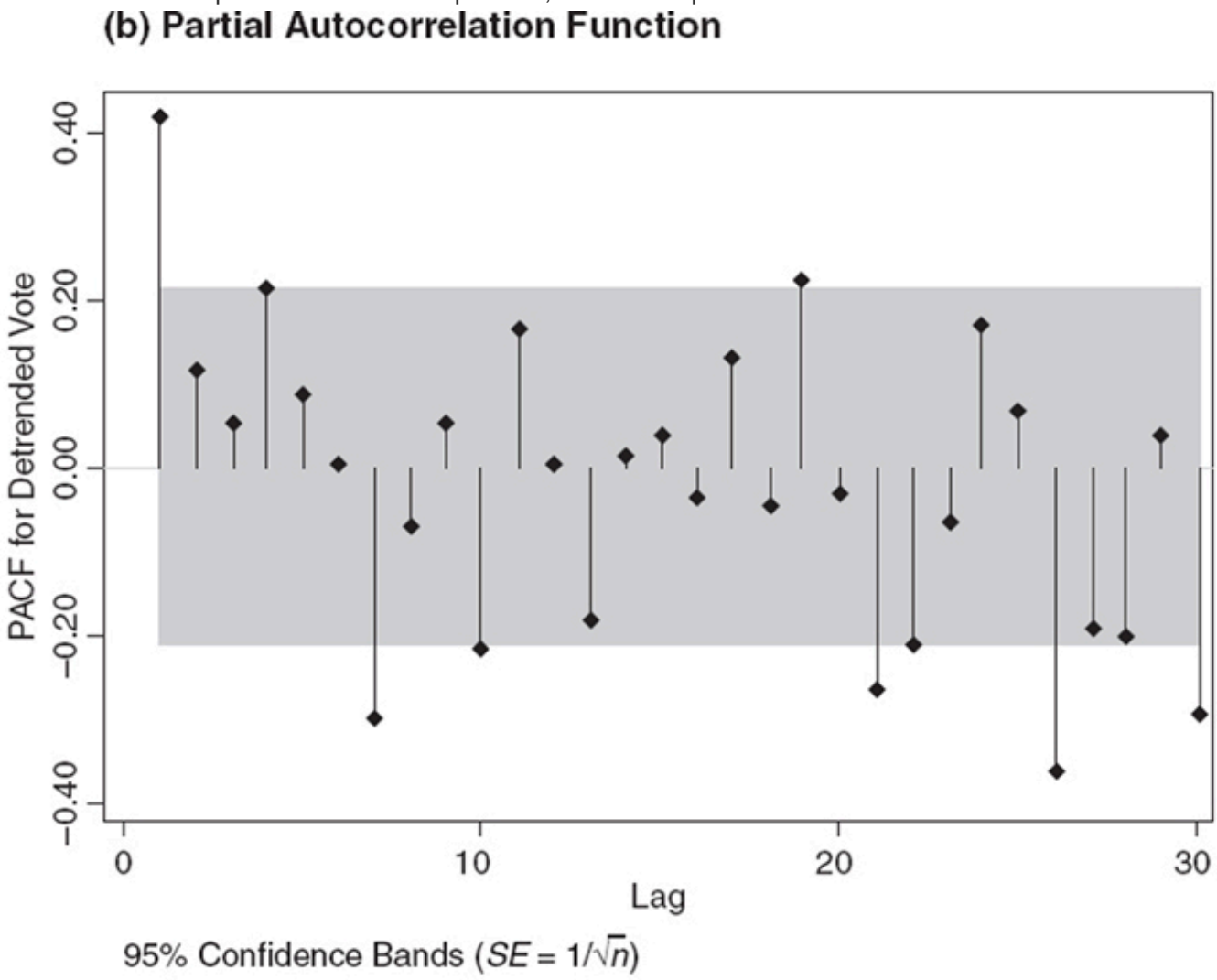
Figure 5.7 Autocorrelation and Partial Autocorrelation Functions for Detrended Vote Intention



Examining the ACF, it appears that the autocorrelations begin to decay at $s = 1$ and again at $s = 4$. Meanwhile, the PACF has spikes at $s = 1$ and $s = 4$. The PACF also appears to have spikes at $s = 7$ and higher, but there are no corresponding significant autocorrelations in the ACF. As a general rule, it is best to focus on the autocorrelations and partial autocorrelations at the lowest s, and worry about the higher values of s only if we are unable to come up with an ARMA model that accounts for all autocorrelation within the residuals. This is based on the principle of parsimony, which is key to the Box-Jenkins approach.

The patterns found in the ACF and PACF suggest that we might try including an AR(1) and an AR($p = 4$) term

in our model. We would denote this as an ARMA (p = 1,4) model. Having identified the ARMA(p = 1,4) model that we think will capture the time series process, the next step is estimation.



NOTE: ACF = autocorrelation function, MA = moving average, PACF = partial autocorrelation function, *SE* = standard error.

Box-Jenkins Approach: Estimation

Once an ARMA model is selected, the α_is and φ_js can be estimated. This is the second stage of the Box-Jenkins approach. Estimation is done using maximum likelihood estimation. Recall the two-component representation of the ARMA process:

$$y_t = \beta_0 + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t, \qquad (5.2.15)$$

where the first equation is called the structural component and the second is the disturbance component. This is the representation often used for the purposes of model estimation and interpretation of the estimation results. The two-component representation of the ARMA(p = 1,4) data-generating process is

$$y_t = \beta_0 + \mu_t,$$

$$\mu_t = \alpha_1 \mu_{t-1} + \alpha_4 \mu_{t-4} + \varepsilon_t. \qquad (5.2.16)$$

Applied to the vote intention data, the maximum likelihood estimates of the data model parameters are presented in Table 5.4.

Table 5.4 Liberal Government Vote Intention—AR(1,4)

Vote	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
AR(1)	0.37	0.090	4.15	<0.001
AR(4)	0.21	0.088	2.42	0.015
Intercept	–0.00036	0.0091	–0.04	0.969

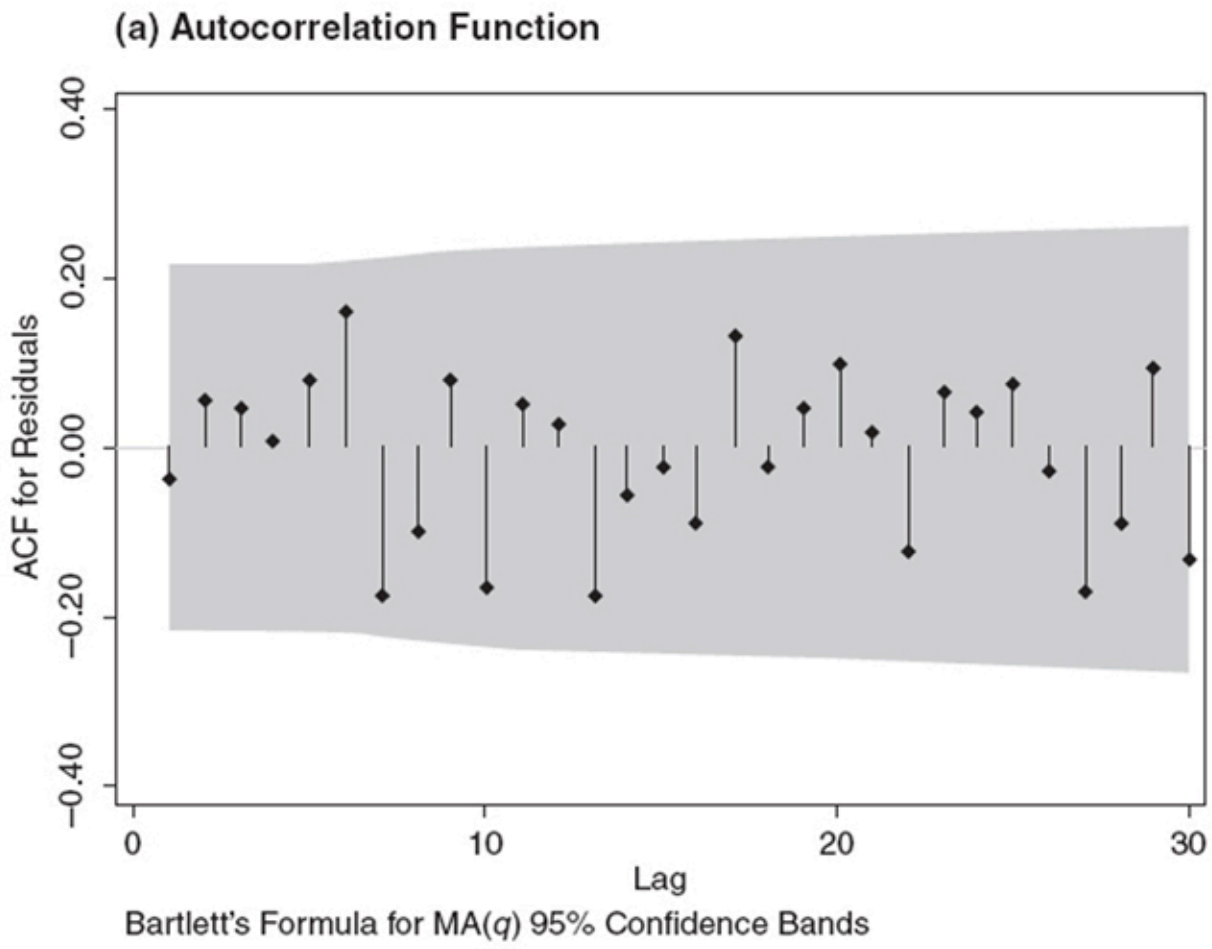
NOTE: Log likelihood = 165.7978, $T = 84$; T = number of time points, AR = autoregressive.

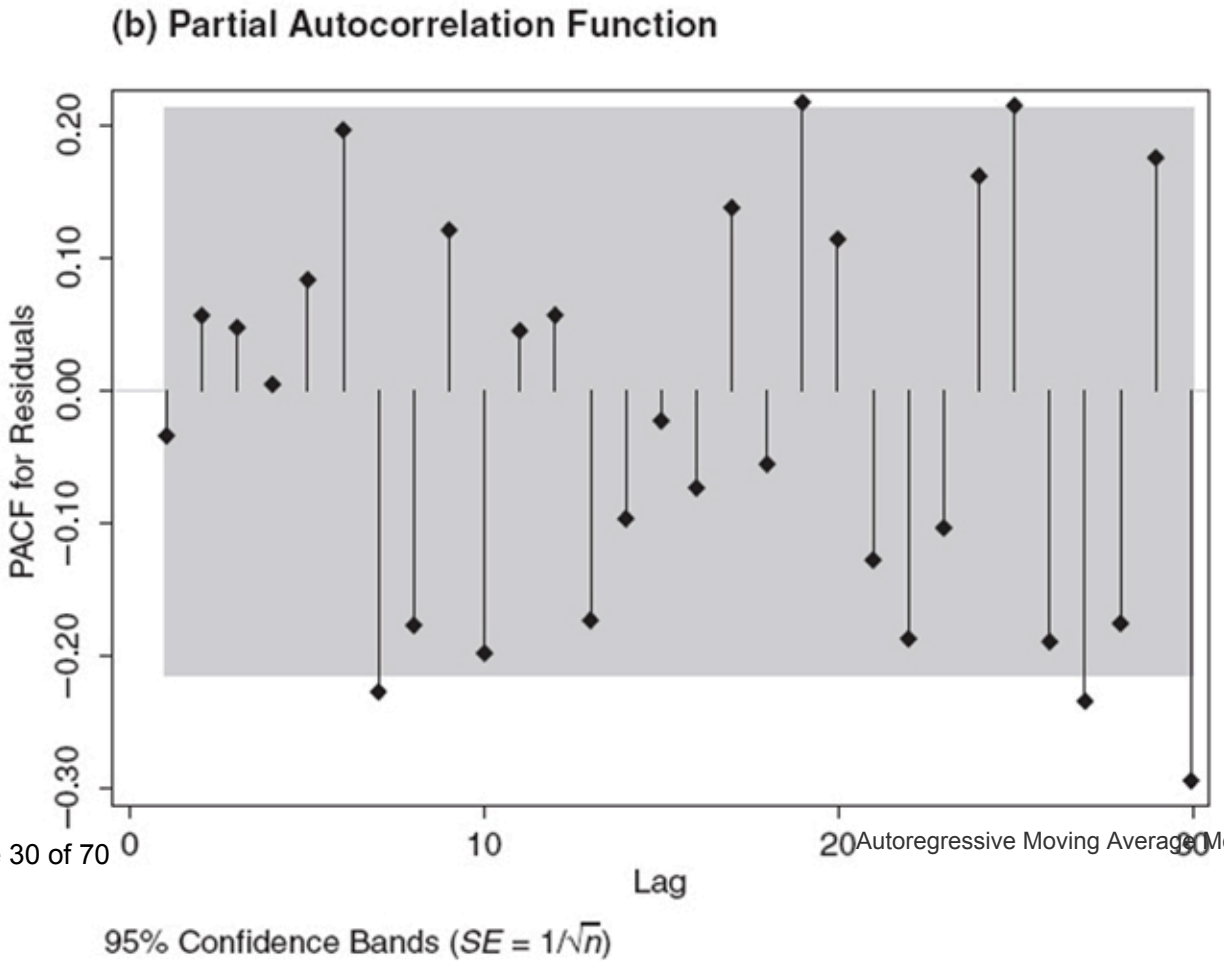
We can see from the estimation results that both AR(1) and AR(4) coefficients are statistically significant at the 0.05 significance level. The next step is diagnostics.

Box-Jenkins Approach: Diagnostic Checking

The first diagnostic check is to determine whether the estimated errors are free from serial correlation or any other unwanted patterns. In other words, we test the residuals to determine if they are a white noise process. This can be done using the Portmanteau Q statistic for white noise, along with the ACF and PACF. The Q statistic is 47.91 and is chi-squared distributed with 40 degrees of freedom. This gives us a P value of 0.18. We are not able to reject the null hypothesis of a white noise process for the residuals. This suggests that the estimated errors do not include serial correlation, trending, or periodicity. The ACF and PACF of the estimated errors can also help us check if the correct model was specified (Figure 5.8).

Figure 5.8 Autocorrelation and Partial Autocorrelation Functions of Liberal Government Vote Model Residuals





We can see from the PACF that there may still be some autocorrelation in the seventh lag. Based on this information, we may try estimating an ARMA(p = 1,4,7) model (Table 5.5).

Table 5.5 Liberal Government Vote Intention—AR(1,4,7)

Vote	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
AR(1)	0.40	0.090	4.48	<0.001
AR(4)	0.26	0.095	2.75	0.006
AR(7)	–0.20	0.11	–1.75	0.081
Intercept	–0.00056	0.0068	–0.08	0.935

NOTE: Log likelihood = 167.678, T = 84; T = number of time points, AR = autoregressive.

Looking at these results, we see that the AR(7) term is not statistically significant at the 0.05 significance level. Based on the Box-Jenkins approach, we would not include this term. Looking at the previous PACF (Figure 5.7), it is not clear whether the autocorrelation at the seventh lag is due to an AR(7) or an MA(7) term. On this basis, we might try estimating an ARMA(p = 1, 4, q = 7) model (Table 5.6).

Table 5.6 Liberal Government Vote Intention—AR(1,4) MA(7)

Vote	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
AR(1)	0.39	0.091	4.35	<0.001
AR(4)	0.25	0.096	2.61	0.009
MA(7)	–0.25	0.12	–2.13	0.033
Intercept	–0.00028	0.0077	–0.040	0.971

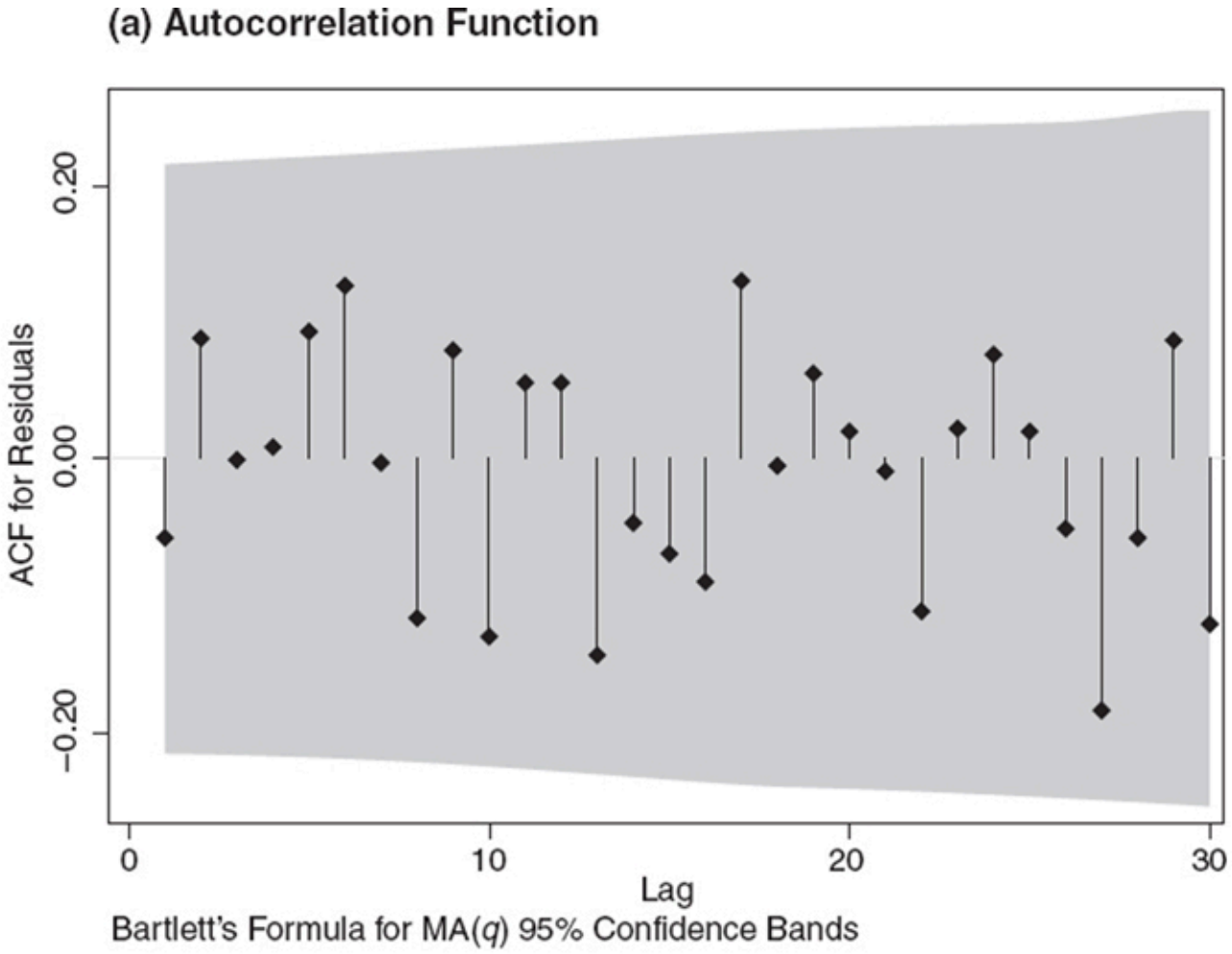
NOTE: Log likelihood = 167.742, T = 84; T = number of time points, AR = autoregressive, MA = moving average.

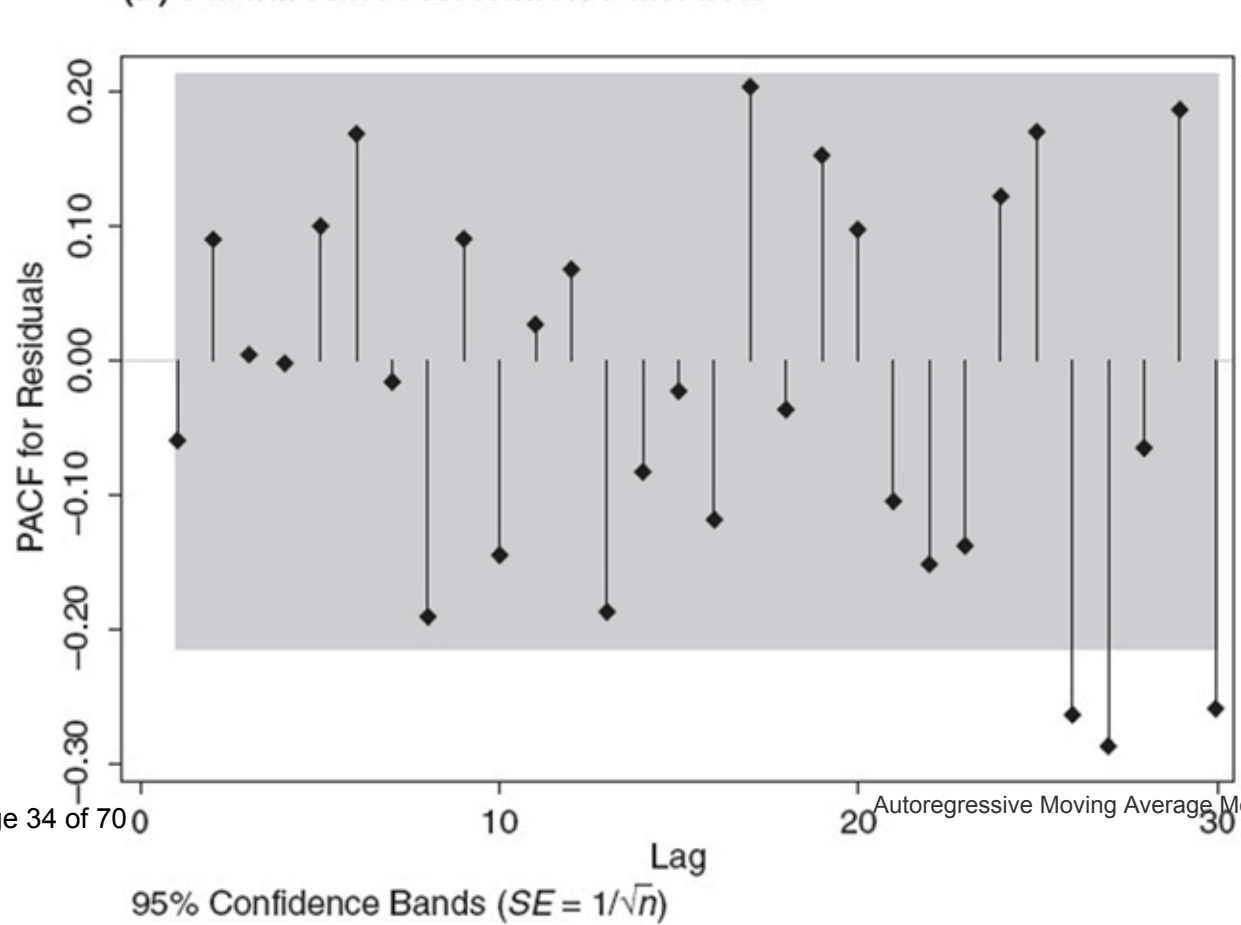
All terms in this model are significant at the 0.05 significance level. Again, we can test the residuals against the null of a white noise process and examine their ACF and PACF. The Q statistic is 43.64, and the corresponding P value is 0.32, again indicating no evidence that serial correlation remains within the residuals.

The ACF and PACF for the residuals (Figure 5.9) show no signs of any remaining autocorrelation. It appears that we have specified a model that correctly accounts for the autocorrelation in the residuals. If

autocorrelations are still found in the estimated errors, the model can again be respecified and reestimated.[1]

Figure 5.9 Autocorrelation and Partial Autocorrelation Functions of Liberal Government Vote Model Residuals





Finally, we may want to compare the fit of ARMA(p = 1, 4, q = 7) with the simpler ARMA(p = 1, 4). These models are nested, and so we can do this with the LR test. We get an LR value of 3.89. This is chi-squared distributed with 1 degree of freedom, giving us a P value of 0.049. We can reject the null hypothesis that the ARMA(p = 1, 4, q = 7) model fits no better than the ARMA(p = 1, 4) model.

5.3 Autoregressive Moving Average (ARMA) Models with Exogenous Regressors

Independent variables (exogenous regressors) and/or their lags can be added to the first component of the two-component representation of the ARMA model:

$$y_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t, \qquad (5.3.1)$$

where $\sum_{j=1}^{k} \beta_j x_j$ represents the exogenous regressors (including lags) that we wish to include in the model. Returning to our example, we want to include economic conditions as variables. Specifically, we include the third lag (recall that the data are monthly) of year-over-year change in GDP and inflation and the sixth lag of unemployment. The lags included are based on theoretical considerations (e.g., the timing of the release of such economic figures). This produces the following estimates (Table 5.7).

Table 5.7 Canadian Economic Voting Model—AR(1,4) MA(7)

Vote	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
L3. GDP	0.0085	0.0034	2.51	0.012
L3. Inf	−0.023	0.0065	−3.53	<0.001
L6. Unemployment	−0.0053	0.0051	−1.05	0.296
Constant	0.059	0.059	1.00	0.319
AR(1)	0.19	0.11	1.63	0.103
AR(4)	0.083	0.15	0.56	0.573
MA(7)	−0.33	0.12	−2.71	0.007

NOTE: Log likelihood = 161.704, $T = 78$; T = number of time points, AR = autoregressive, GDP = gross domestic product, MA = moving average, L3 = third lag, L6 = sixth lag.

We see from the results that the AR(1) and AR(4) terms are no longer statistically significant. This sometimes happens once the independent variables are included. This may also occur because we have lost six data points by including the sixth lag of unemployment. Given the Box-Jenkins guidelines to include only the autoregressive and moving average terms that reach statistical significance, we may now want to rerun the model without the AR(1) and AR(4) terms and then test the errors. We might also want to test if either of these terms is statistically significant when entered individually. If we did this, we would find that they are not. The results from the model without these terms are given in Table 5.8.

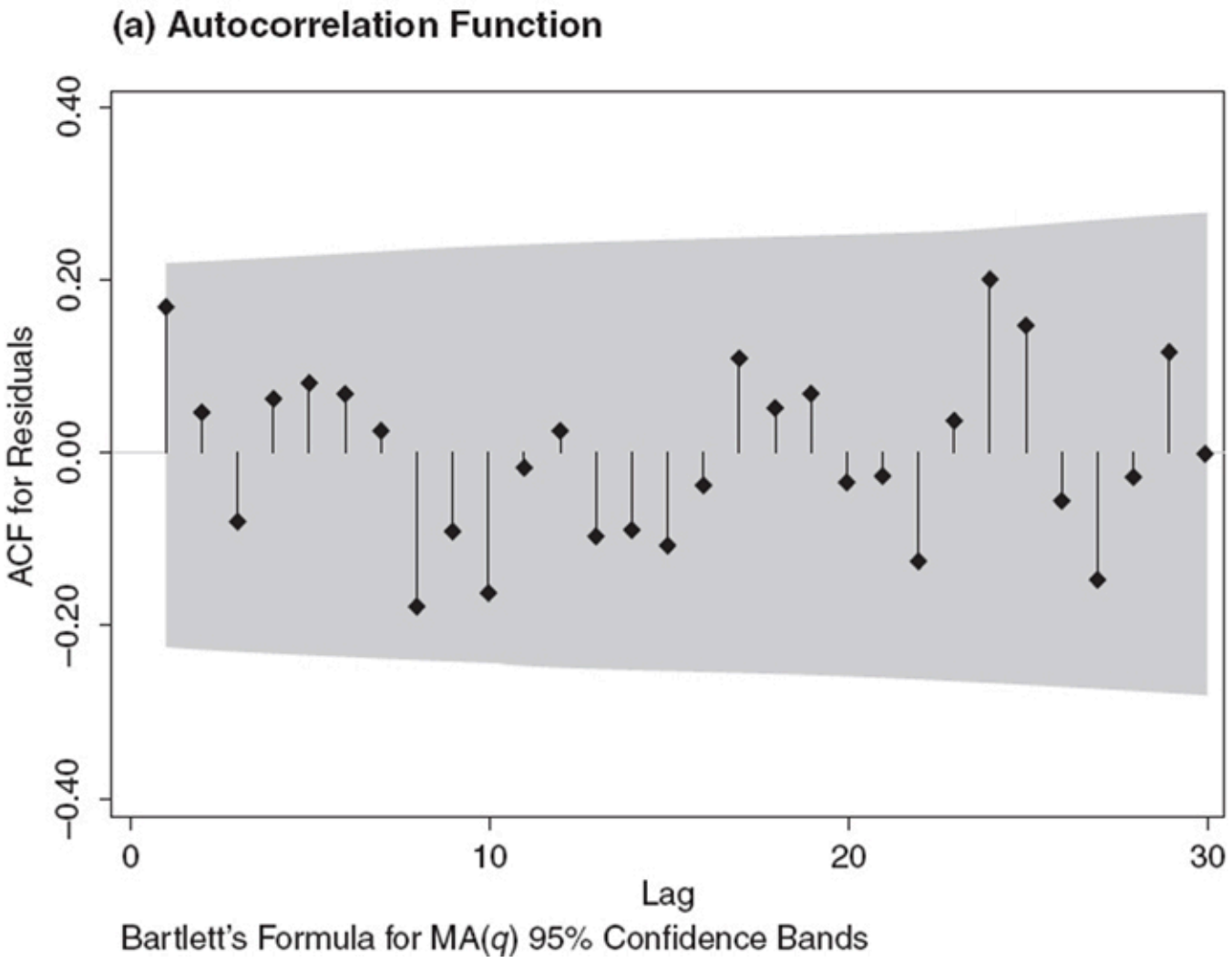
Table 5.8 Canadian Economic Voting Model—MA(7)

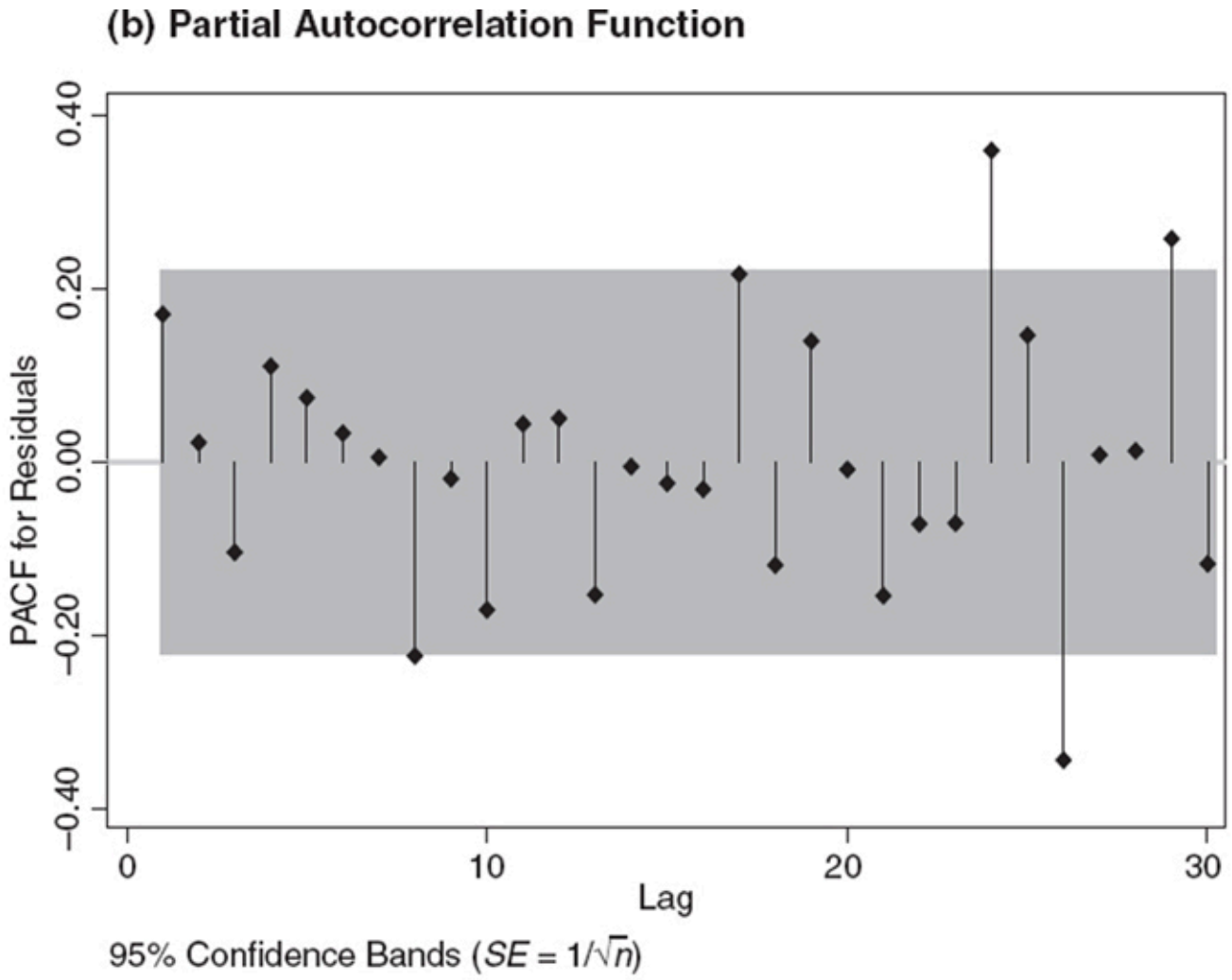
Vote	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
L3. GDP	0.0082	0.0027	3.05	0.002
L3. Inf	–0.023	0.0052	–4.37	<0.001
L6. Unemployment	–0.0057	0.0039	–1.47	0.143
Constant	0.063	0.045	1.38	0.168
MA(7)	–0.38	0.12	–3.22	0.001

NOTE: Log likelihood = 160.22, $T = 78$; T = number of time points, GDP = gross domestic product, MA = moving average, L3 = third lag, L6 = sixth lag.

We can now predict the errors and calculate the *Q* statistic to test if they follow a white noise process. The *Q* statistic is 41.26, and the corresponding *P* value is 0.29. Based on these results, we cannot reject the null hypothesis that the errors follow a white noise process. We can also examine the ACF and PACF for the residuals (Figure 5.10).

Figure 5.10 Autocorrelation and Partial Autocorrelation Functions for Residuals





NOTE: ACF = autocorrelation function, MA = moving average, PACF = partial autocorrelation function, *SE* = standard error.

No evidence of autocorrelation remaining in the errors can be found in the ACF. The PACF shows little sign of autocorrelation. This evidence, in combination with the *Q* test, indicates that we have specified a model that correctly accounts for the autocorrelation in the residuals.

Again, we may want to compare model fits. Specifically, we might want to compare the ARMA(p = 1, 4, q = 7) model without exogenous regressors with the ARMA(q = 7) with exogenous regressors. We might do this using BIC values.[2] For a fair comparison, we must compare the BIC values for the two models estimated on the same data. This requires us to reestimate the ARMA(p = 1, 4, q = 7) model without the first six points. The BIC for this model is -284.29, and for the ARMA(q = 7) model with exogenous regressors, it is -294.31. The BIC for the model with exogenous regressors is 10 smaller than the BIC for the model without. Referring to Table 5.2, this is strong to very strong evidence that the model with exogenous regressors fits better.

We now turn to interpreting the estimated effects of the exogenous regressors on vote intention. The coefficients on the independent variables in the ARMA are the long-run effects of a permanent one-unit

increase in the independent variable. However, they are also the immediate effects. Just like the static model in Chapter 3, the independent variables in an ARMA model have their full effect immediately. This is different from an LDV model where the independent variables have an initial short-run effect, which then builds into a long-run effect. We can see why this is so by considering the ARMA(1,1) process with a single exogenous regressor:

$$y_t = \beta_0 + \beta_1 x_t + \mu_t, \qquad (5.3.2)$$

$$\mu_t = \alpha_1 \mu_{t-1} + \phi_1 \varepsilon_{t-1} + \varepsilon_t. \qquad (5.3.3)$$

By rearranging Equation 5.3.2,

$$\mu_t = y_t - \beta_0 - \beta_1 x_t, \qquad (5.3.4)$$

and lagging one period,

$$\mu_{t-1} = y_{t-1} - \beta_0 - \beta_1 x_{t-1}. \qquad (5.3.5)$$

We can now insert Equations 5.3.4 and 5.3.5 into Equation 5.3.3,

$$y_t - \beta_0 - \beta_1 x_t = \alpha_1 \left(y_{t-1} - \beta_0 - \beta_1 x_{t-1} \right) + \phi_1 \varepsilon_{t-1} + \varepsilon_t,$$

and rearrange the terms:

$$y_t = \alpha_1 y_{t-1} - \alpha_1 \beta_0 - \alpha_1 \beta_1 x_{t-1} + \beta_0 + \beta_1 x_t + \phi_1 \varepsilon_{t-1} + \varepsilon_t,$$

$$y_t = \alpha_1 y_{t-1} + \left(1 - \alpha_1\right) \beta_0 + \beta_1 x_t - \alpha_1 \beta_1 x_{t-1} + \phi_1 \varepsilon_{t-1} + \varepsilon_t. \qquad (5.3.6)$$

This is the ADL(1,1) model from Chapter 4, with the addition of a moving average error and the constraint that the coefficient for the first lag of x_t is equal to the autoregressive parameter times the coefficient on x_t times -1: $-\alpha_1 \beta_1$. For an ADL(1,1), the immediate effect of a one-unit increase in x_t is β_1, and the long-run effect of x_t is

$$\frac{\beta_1 - \alpha_1 \beta_1}{\left(1 - \alpha_1\right)} = \beta_1. \qquad (5.3.7)$$

The constraint on the coefficient for the first lag of x_t guarantees that the long-run effect is equal to the initial effect: β_1. We can only include a long-run effect that differs from the immediate effect in the ARMA model by adding lags of the independent variables (like the FDL model in Chapter 3).

Returning to our estimates and using the 0.05 significance level, it would appear that year-over-year growth in GDP has a small but positive effect on vote intention for the party in government, while inflation has a reasonably large and negative effect: An increase of 1 percentage point in inflation reduces government popularity by 2.3 percentage points.

We can apply the formula for calculating the equilibrium of an ADL to Equation 5.3.6:

$$\frac{(1-\alpha_1)\beta_0}{(1-\alpha_1)} = \beta_0.$$

The β_0 in the ARMA is the long-run equilibrium, when growth, inflation, and unemployment are all zero. In reality, this is not a particularly likely scenario.

Having covered the basics of the Box-Jenkins approach to identifying, estimating, and testing ARMA models, in the next chapter, we will examine data with a trend of a different sort than we have encountered so far. We will examine the possibility that our data are a unit root process and, therefore, not stable. We will extend the ARMA approach to be able to model such data—the autoregressive integrated moving average (ARIMA) model. We will also consider how to account for periodicity in our data with such models. In the following section, we will look further at including exogenous variables, as we discuss transfer functions and intervention analysis.

5.4 Interventions and Transfer Functions

In Chapter 3, we estimated a model of (the log of) the number of drivers in the United Kingdom who were killed or seriously injured (KSI) in traffic accidents, using monthly data between January 1969 and December 1984 (Harvey & Durbin, 1986). We included variables that controlled for and estimated the magnitudes of seasonality and a structural break due to the introduction of a new seatbelt law at t = 170 (February 1983). The variable "seatbelt law" is coded "0" before t = 170 and "1" at t = 170 and afterward. The inclusion of this variable is a basic form of intervention analysis. The seatbelt law is a policy intervention that is expected to affect KSIs. We assume that the effect is to produce a change in the mean level of the dependent variable, "KSI," at time point t = 170. This change is assumed to be permanent. This is called a step function. We allow the magnitude of the shift in the mean to be estimated. The results shown in Table 3.8 indicate that the magnitude of the mean shift is a 15% reduction in KSIs per month.

This type of intervention is an example of a level change (Pourahmadi, 2001) or a deterministic-step change (Tsay, 1988). Such an intervention could also include additional dynamics such as building in magnitude over time (Box & Tiao, 1975). Other types of interventions include additive outliers, innovation outliers, transient changes, or any combination of these (Pourahmadi, 2001). An additive outlier is a change in the mean at a particular time point with an immediate reversion to the original mean at the next time point. The residual effect of the outlier may last for some time if the model is dynamic. Such an intervention can be represented in a model by including a variable that is coded "0" before the intervention, "1" at the time of the intervention, and "0" afterward. This is called a pulse function.

An innovation outlier is similar to an additive outlier except that it affects the innovations and would be modelled by including a variable representing a pulse function as multiplicative heteroskedasticity in a autoregressive conditional heteroscedasticity (ARCH) process. The included variable is the same as that for additive outliers, except that it is included in the model of the error variance. A transient change is also like

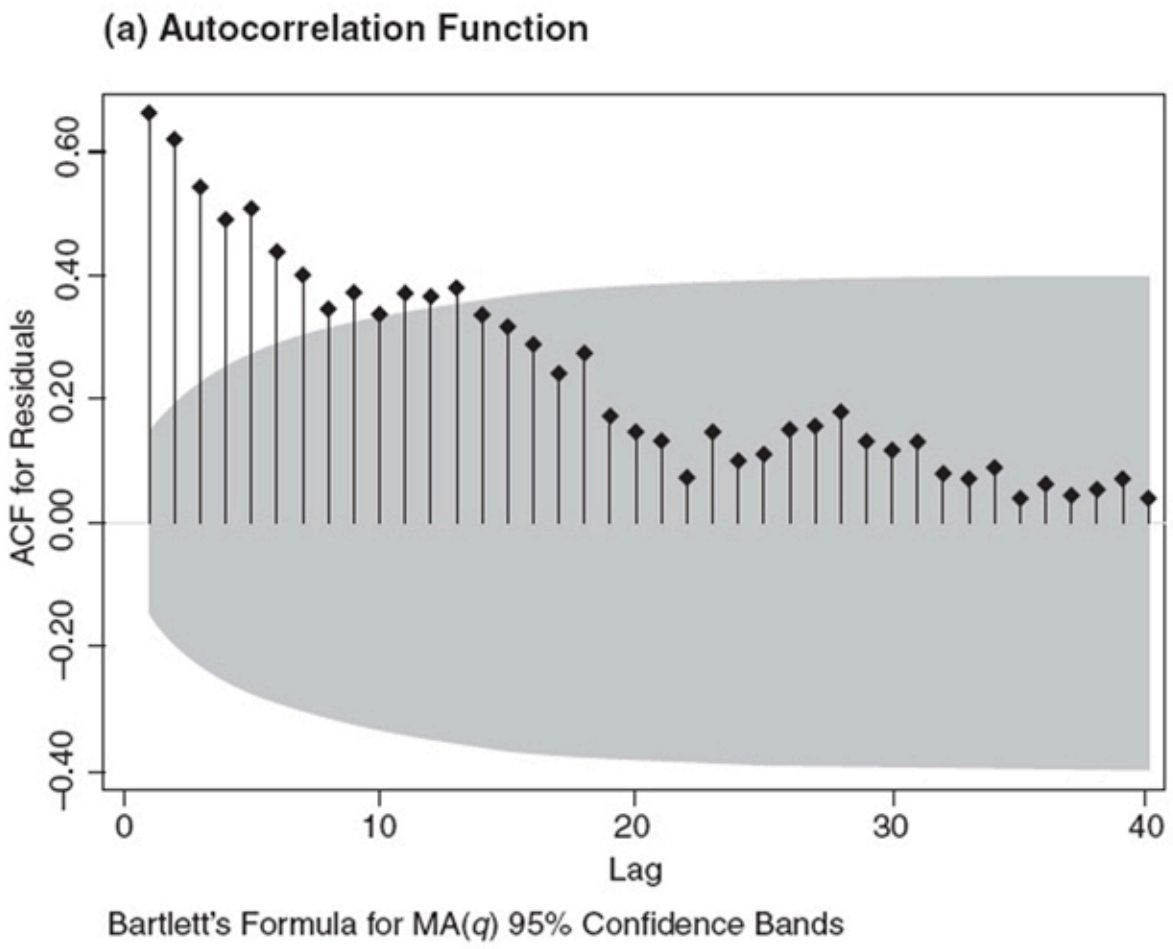
an additive outlier, except that the change in the mean of the series due to an intervention dissipates slowly with time. These can take many different forms, and we will examine them shortly, when we discuss transfer functions.

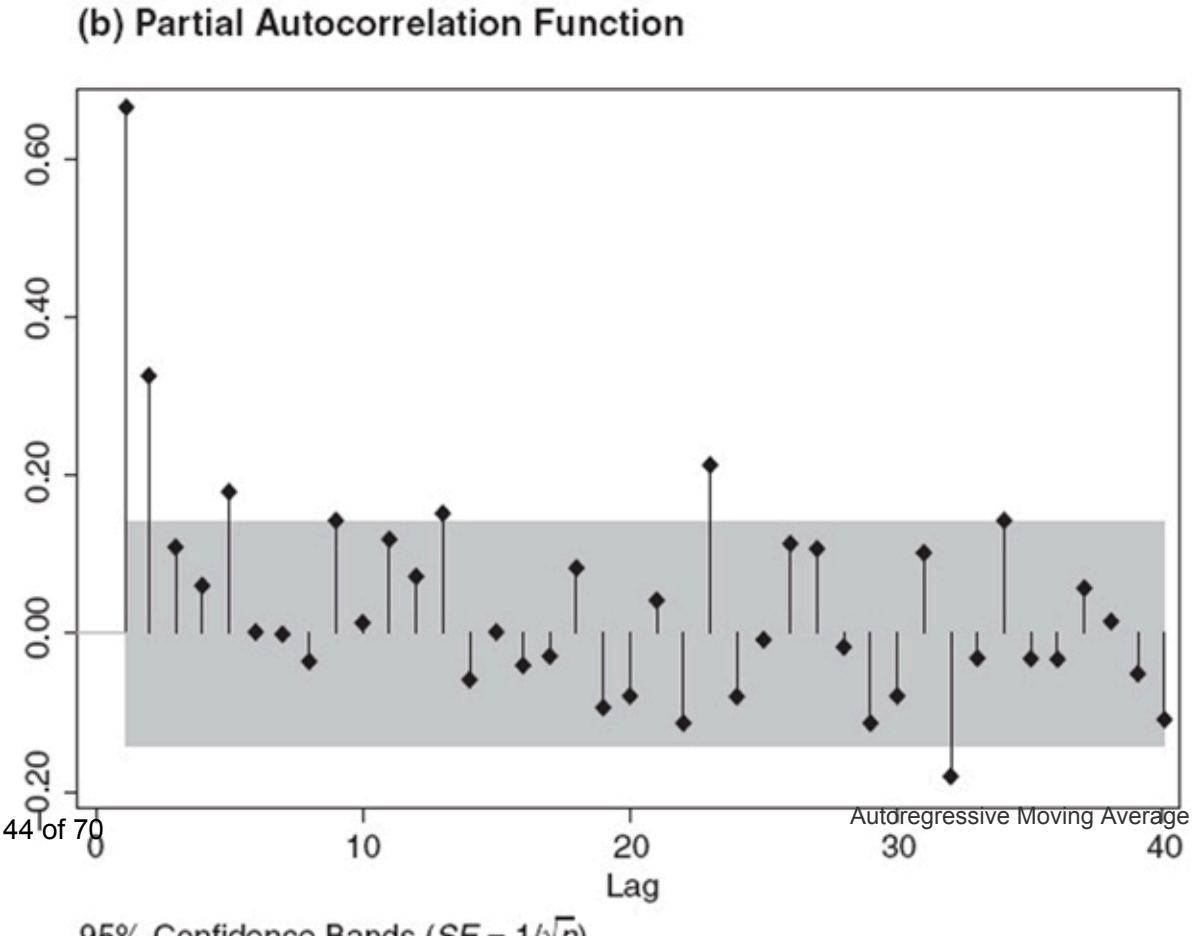
We could also model a structural break in the covariance between variables. For example, the effect of petrol prices on KSI may be suspected of changing after the introduction of the new seatbelt law. This can be modelled by including an interaction between the seatbelt law step function and petrol prices.

In some instances, we are uncertain regarding the exact timing or nature of an intervention. In this case, we may want to conduct a more exploratory analysis. This can be done by examining the residuals from an out-of-sample forecast, as illustrated in the following example. Let us say we believe that an intervention of some form occurred after time point 169. We begin by estimating a model using just the data up to time point 169—prior to the potential intervention.

Previously, we did not examine the ACF or PACF for the residuals to determine if any autoregressive or moving average terms should be included in the model. If we follow the Box-Jenkins approach to model specification, we might determine that the appropriate model for ln(KSI) is ARMA(1,1). To see this, estimate the model from Chapter 3 using data up to time point 169. This model includes the monthly dummy variables and the log of petrol prices. It does not include the seatbelt law variable, as this occurred after time point 169. We next examine the autocorrelation and partial autocorrelations from the resulting residuals (Figure 5.11).

Figure 5.11 Autocorrelation and Partial Autocorrelation Functions for Residuals





The ACF suggests the presence of an AR(1) process, while the PACF suggests the presence of an MA(1) process. The possibility of an MA(1) process is suggested by the fact that the spike at a lag of 1 is followed by a shorter spike at a lag of 2 and even shorter spikes at lags of 3 and 4—that is, decaying partial autocorrelations. The results from estimating ARMA(1,1) model are given in Table 5.9.

Table 5.9 ln(KSI) in U.K. Traffic Accidents, January 1969 to February 1983

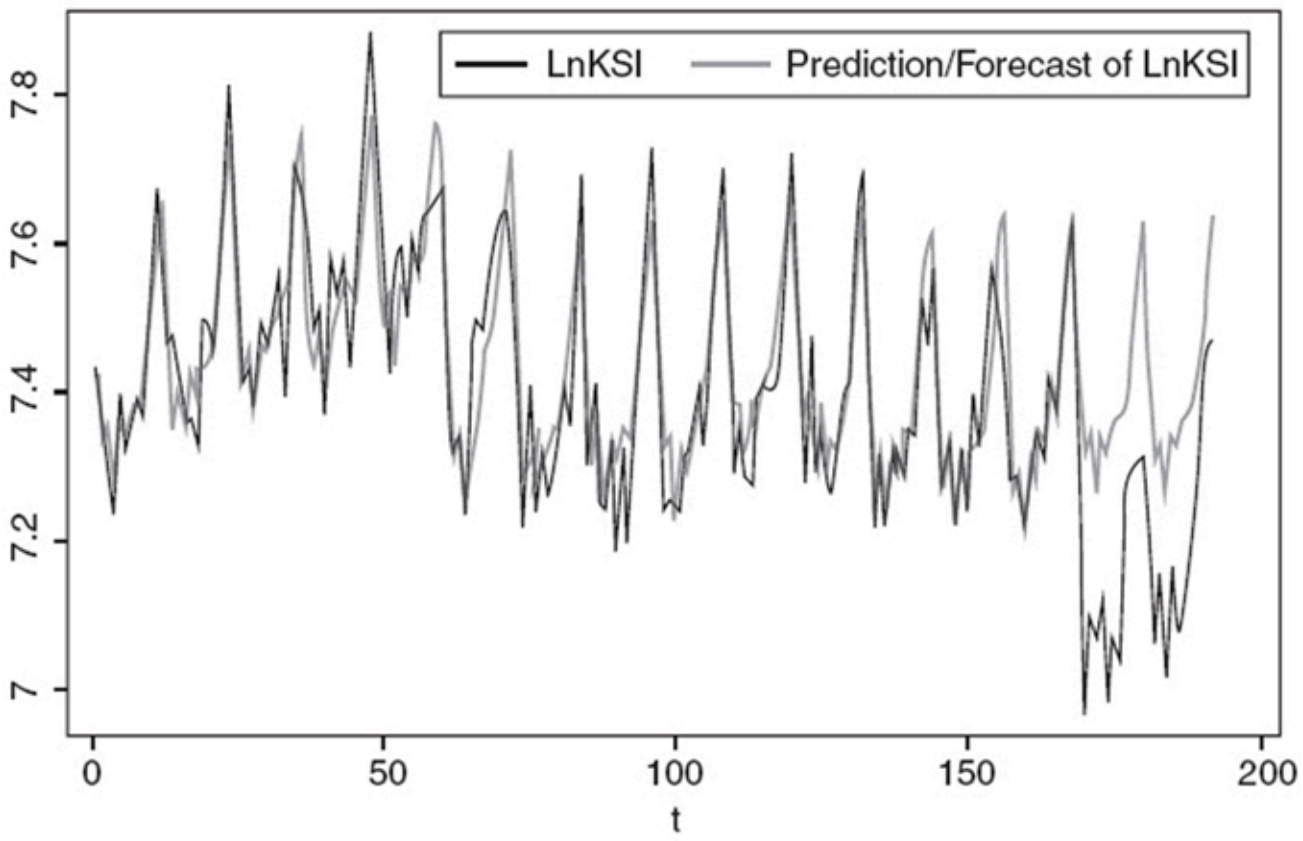
ln(KSI)	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
January	–0.24	0.019	–12.48	<0.001
February	–0.34	0.021	–16.34	<0.001
March	–0.31	0.023	–13.64	<0.001
April	–0.39	0.033	–11.79	<0.001
May	–0.30	0.025	–12.11	<0.001
June	–0.33	0.024	–13.53	<0.001
July	–0.28	0.028	–10.1	<0.001
August	–0.27	0.023	–11.57	<0.001
September	–0.25	0.023	–11.11	<0.001
October	–0.18	0.023	–7.68	<0.001
November	–0.061	0.021	–2.86	0.004
Petrol—ln(£)	–0.33	0.13	–2.57	0.01
Constant	6.93	0.29	23.74	<0.001
AR(1)	0.92	0.045	20.29	<0.001
MA(1)	–0.64	0.093	–6.89	<0.001

NOTE: Log likelihood = 211.317, $T = 169$; T = number of time points, AR = autoregressive, MA = moving average, KSI = killed or seriously injured.

Next we produce predicted values for ln(KSI) using the parameters from the model estimated using only the data up to time point 169.

After this time point, these are out-of-sample forecasts. We next plot the predicted values and forecasts and compare them to the raw ln(KSI) data (Figure 5.12).

Figure 5.12 Predicted and Forecast ln(KSI)



NOTE: KSI = killed or seriously injured.

We can see the consequence of not accounting for the seatbelt law after time point 169. This also gives us a nice visual representation of the estimated effect of the seatbelt law. At time point 170, the forecasted value of ln(KSI) clearly deviates from the observed value. This deviation appears to remain to the end of the observed time series. It would appear that a level change intervention occurring at time point 170 is appropriate. Therefore, we reestimate our model for the full period for which we have data and include the seatbelt law intervention as a step function (Table 5.10).

Table 5.10 ln(KSI) in U.K. Traffic Accidents, January 1969 to December 1984

ln(KSI)	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
January	–0.23	0.018	–12.93	<0.001
February	–0.35	0.020	–17.44	<0.001
March	–0.31	0.022	–13.99	<0.001
April	–0.38	0.030	–12.95	<0.001
May	–0.30	0.023	–12.67	<0.001

ln(KSI)	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
June	–0.33	0.022	–14.92	<0.001
July	–0.28	0.026	–10.88	<0.001
August	–0.27	0.022	–12.39	<0.001
September	–0.24	0.020	–11.71	<0.001
October	–0.16	0.021	–7.71	<0.001
November	–0.055	0.020	–2.75	0.006
Petrol—ln(£)	–0.30	0.12	–2.46	0.014
Seatbelt law—step function	–0.22	0.041	–5.32	<0.001
Constant	7.01	0.28	25.43	<0.001
AR(1)	0.92	0.041	22.31	<0.001
MA(1)	–0.66	0.087	–7.55	<0.001

NOTE: Log likelihood = 240.945, $T = 192$; T = number of time points, AR = autoregressive, MA = moving average, KSI = killed or seriously injured.

We next apply the *Q* test to the residuals from the model. The Portmanteau (*Q*) statistic is 54.72 and is chi-squared distributed with 40 degrees of freedom. The corresponding *P* value is 0.060. The Portmanteau statistic indicates that we cannot reject the null hypothesis of the errors being a white noise process. Note that including additional autoregressive and/or moving average processes would produce a Portmanteau statistic with a larger *P* value, but following the principle of parsimony, we use the model with the fewest such components that produces residuals for which we cannot reject the null of a white noise process at our chosen significance level.

The estimated coefficient for the seatbelt law step function is -0.22, indicating that the introduction of the new seatbelt law produced a 22% reduction in KSIs. As this is an ARMA model, it is also the long-run effect of the seatbelt law.

What if we had used a pulse function for the seatbelt law variable? Table 5.11 displays the results of the estimates from such a model. The immediate effect of the seatbelt law is -0.21. This is of the same magnitude as before. However, because the seatbelt law is a pulse function, it drops to 0 after time point 170. The consequence of including a pulse function is to model an additive outlier intervention. This means that the effects of the seatbelt law change disappear after $T = 170$. Given the effect of the seatbelt law plotted in Figure 5.12, this seems inappropriate. This brings us to transfer functions.

Table 5.11 ln(KSI) in U.K. Traffic Accidents, January 1969 to December 1984

ln(KSI)	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
January	–0.23	0.019	–12.39	<0.001
February	–0.34	0.020	–17.06	<0.001
March	–0.31	0.021	–15.28	<0.001
April	–0.39	0.030	–13.02	<0.001
May	–0.30	0.025	–12.01	<0.001
June	–0.34	0.023	–14.74	<0.001
July	–0.28	0.027	–10.6	<0.001
August	–0.28	0.023	–11.99	<0.001
September	–0.24	0.021	–11.31	<0.001
October	–0.16	0.022	–7.51	<0.001
November	–0.056	0.019	–3.01	0.003
Petrol—ln(£)	–0.32	0.16	–1.96	0.05
Seatbelt law—pulse function	–0.21	0.050	–4.23	<0.001
Constant	6.93	0.37	18.74	<0.001
AR(1)	0.94	0.034	27.79	<0.001
MA(1)	–0.55	0.079	–6.98	<0.001

NOTE: Log likelihood = 234.871, $T = 192$; T = number of time points, AR = autoregressive, MA = moving average, KSI = killed or seriously injured.

Transfer Functions

In the last example, we assumed a lag structure for the independent variable, “petrol.” This lag structure was based on theoretical expectations. In the social sciences, we generally have strong theories to suggest when an independent variable will have its effect on our dependent variable. However, there is sometimes some ambiguity about whether an independent variable has an effect that is immediate or has some degree of lag.

If an independent variable x_t is white noise, as in Equation 5.4.2, it is rather straightforward to examine the cross-correlations between y_t and x_t to determine the lag structure of x_t for the model of y_t.

$$y_t = \alpha_{0,1} + \alpha_{1,1} y_{t-1} + \beta_{1,1} x_t + \varepsilon_t, \qquad (5.4.1)$$

$$x_t = \alpha_{0,2} + \nu_t. \qquad (5.4.2)$$

The cross-correlations are the correlations between y_t and the lags of x_t: x_t-1, x_t-2, …. A plot of these cross-correlations is called a cross-correlogram. The use of a cross-correlogram to determine the lag structure in this manner makes the important assumption that there is no feedback from y_t to x_t. This means that our data-generating process for x_t (Equation 5.4.2) does not contain a current or lagged value of y_t. If such an assumption is not feasible for the contemporaneous values of x_t, we need to consider more advanced multivariate time series approaches, such as vector autoregression. This is an advanced topic, which receives excellent treatment in Brandt and Williams (2006) and Enders (2004).

As an example of the use of a cross-correlogram to determine the lag structure of an independent variable, consider a model of media tone toward the governing party as a function of errors in published vote intention polls. The data are derived from media coverage of the 2006 Canadian federal election campaign. The media content analysis involves a sample of mainstream newspaper, radio, television, and Internet coverage of the election during the campaign period.[3] Codes for media tone—*positive, negative*, or *neutral*—were assigned for all major parties. A party received a -1, 0, or +1 tone score each time it appeared in a story. Net tone, determined by the proportion of positive stories minus the proportion of negative stories, indicates the relative weight of positive news over negative news for a given party over all stories on a given campaign day. For this example, we use the net tone of media coverage for the governing Liberal party.

The poll error (in percentage points) is based on the average of the differences between the poll results published each day of the campaign and estimates of true vote intention for the period the polls were in the field (Pickup, Andrew, Cutler, & Matthews, 2014). Again, we use the estimated poll error for the governing Liberal party. It is the change in this estimate of poll error that is used as the regressor. Polling error is random and (theoretically) independent from one poll to the next, and so it should be white noise, as in Equation 5.4.2, as should its first difference. The cross-correlogram is presented in Table 5.12.

Table 5.12 Cross-Correlogram: Media Tone and Poll Error

Lag	*Cross-Correlations*
0	0.33

1	0.16
2	−0.057
3	0.0057
4	−0.087
5	0.025

The cross-correlogram indicates cross-correlations at the zero and first lags of poll error. On this basis, media tone is regressed on these lags of poll error. The results are presented in Table 5.13. The significant and positive coefficients for poll error and its lag suggest that errors that overstate the popularity of the governing Liberal party tend to produce positive news coverage the day the poll is published and the following day. A change of 1 percentage point in the error in favor of the Liberal party results in a positive shift of 2 to 2.5 percentage points in net tone of the media coverage of the Liberal party each day.

Table 5.13 Media Tone and Poll Error

Tone	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
D. Poll error	0.026	0.010	2.6	0.014
LD. Poll error	0.021	0.0088	2.42	0.021
Constant	–0.14	0.010	–13.94	<0.001

NOTE: $R^2 = 0.19$, $T = 38$; T = number of time points, L = lag of variable, D = difference of variable.

If x_t is not a white noise process, the appropriate lag structure for x_t in a model of y_t is more difficult to determine. To see why this might be the case, let the data-generating process for y_t and x_t be

$$y_t = \alpha_{0,1} + \alpha_{1,1} y_{t-1} + \beta_{1,1} x_t + \varepsilon_t, \qquad (5.4.3)$$

$$x_t = \alpha_{0,2} + \alpha_{1,2} x_{t-1} + \nu_t. \qquad (5.4.4)$$

If we plug Equation 5.4.4 into Equation 5.4.3, we get the following:

$$y_t = \alpha_{0,1} + \alpha_{1,1} y_{t-1} + \beta_{1,1}\left(\alpha_{0,2} + \alpha_{1,2} x_{t-1} + \nu_t\right) + \varepsilon_t,$$

$$y_t = (\alpha_{0,1} + \beta_{1,1}\alpha_{0,2}) + \alpha_{1,1} y_{t-1} + \beta_{1,1}\alpha_{1,2} x_{t-1} + \beta_{1,1}\nu_t + \varepsilon_t. \qquad (5.4.5)$$

It might seem that the appropriate model for y_t includes a lag of x_t. There is nothing wrong with the estimation of such a model, but it would be incorrect to interpret this to mean that x_t causes y_t with a lag. If Equations 5.4.3 and 5.4.4 do represent the data-generating process, then the lag of x_t only has an effect on y_t through

its effect on x_t. When the dynamic structure of y_t and x_t becomes more complicated, there is an increasing number of possible models for y_t and no clear-cut method for distinguishing between them.

In response to this challenge, Box and Jenkins (1976) proposed an inductive method for selecting the optimal lag structure for the purposes of forecasting. This requires the transfer function representation of the ARMA model (Harvey, 1993, subsection 5.8). Consider the two-component representation of the ARMA model with a single exogenous regressor, x_t:

$$y_t = \beta_0 + f_t + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t. \qquad (5.4.6)$$

The transfer function f_t is a function of x_t and determines how movements in the independent variable are translated into movements in the dependent variable. For example, a change in x_t might simply have an immediate and permanent effect, in which case f_t would simply be

$$f_t = \beta_1 x_t.$$

Of course, a change in x_t might have a more complicated effect, such as an immediate short-run effect that builds into a permanent long-run effect,

$$f_t = \varrho f_{t-1} + \beta_1 x_t,$$

assuming that $|\eta| < 1$. Unfortunately, the transfer function representation does not lend itself to a systematic approach when there is more than one explanatory variable (Harvey, 1990, subsection 7.5). Furthermore, the Box-Jenkins approach is designed to optimize the value of the model for forecasting and not for testing hypotheses. Therefore, while it is important to consider alternative lag structures for the independent variables, the possibilities considered are best driven by theoretical considerations. In the context of intervention analysis, where the lag structure of an intervention is often less clear and may not be suggested to us by theory, the transfer function is a useful concept. Therefore, we examine the idea of a transfer function in that context.

Let the transfer function f_t have the following form:

$$f_t = \varrho f_{t-1} + \beta_1 x_t. \qquad (5.4.7)$$

The η parameter can be assumed to be 1, assumed to be 0, or estimated assuming that $|\eta| < 1$. The *indicator variable* x_t could be either a step function or a pulse function and indicates a disturbance at a particular time $t = d$. If f_t is deterministic, $x_t = 0$ for $t < d$, $x_t = 1$ for $t = d$, and depending on whether x_t is a step or pulse $x_t = 1$ or $x_t = 0$ for $t > d$. If f_t is stochastic, $x_t = 0$ for $t < d$, $x_t \sim N(0, \sigma^2_V)$ for $t = d$ (Tsay, 1988), and depending on whether x_t is a step or pulse $x_t \sim N(0, \sigma^2_V)$ or $x_t = 0$ for $t > d$ (Box & Tiao, 1975).

For example, if we wish to model a deterministic step (aka level) change intervention, x_t is a deterministic step function and η is assumed to be 0:

$$f_t = \beta_1 x_t. \qquad (5.4.8)$$

Alternatively, η could be estimated, and f_t would model a deterministic dynamic step change:

$$f_t = \varrho f_{t-1} + \beta_1 x_t. \qquad (5.4.9)$$

If, instead, η is assumed to be 1, f_t produces a deterministic ramp response:

$$f_t = f_{t-1} + \beta_1 x_t. \qquad (5.4.10)$$

A ramp response builds over time, increasing by β_1 each time point. If x_t in Equation 5.4.9 is a pulse function, instead of a step function, f_t produces an effect that decays to zero at a rate determined by η. This last intervention is an example of a transient change intervention. To produce more complicated interventions, multiple transfer functions can be combined. If we wish the intervention to produce an effect that decays to a nonzero value, we can include two transfer functions in Equation 5.4.6:

$$f_{1,t} = \varrho f_{1,t-1} + \beta_1 x_t.$$

$$f_{2,t} = f_{2,t-1} + \beta_2 x_t. \qquad (5.4.11)$$

Assuming that $|\eta| < 1$ and with x_t defined as a pulse function, the initial effect will have a magnitude of $\beta_1 + \beta_2$, and this will decay to a magnitude of β_2. While each of these interventions has been deterministic, they can be modelled as stochastic by using the stochastic form of x_t.

If we combine Equation 5.4.6 with Equation 5.4.11 or any other transfer function, we have what is called a state-space model. This is an advanced topic, and the interested reader is referred to Commandeur and Koopman (2007).[4] Such a model is called a structural model as the included terms and parameters reflect the theorized structure of the data-generating process.

$$y_t = \beta_0 + f_{1,t} + f_{2,t} + \mu_t,$$

$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t.$$

$$f_{1,t} = \varrho f_{1,t-1} + \beta_1 x_t.$$

$$f_{2,t} = f_{2,t-1} + \beta_2 x_t. \qquad (5.4.12)$$

If we wish to estimate such a model as an ARMA or ADL model, we would need to transform it. If our model contains only a single autoregressive term and the transfer function from Equation 5.4.9, an appropriate transformation would be as follows:

$$y_t = \beta_0^* + \alpha_1 y_{t-1} + \beta_1^* x_t + \varepsilon_t,$$

where $\beta^*_0 \equiv (1 - \alpha_1)\beta_0$ and $\beta^*_1 \equiv \alpha_1/\eta\beta_1$. The transformation is straightforward but involved and so is not

included here. Such a model could be estimated by OLS, but note that we do not estimate the structural parameters.

In the last section of this chapter, we extend our discussion of ARCH models from Chapter 4.

5.5 Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Models

In Chapter 4, we described an LDV(1) model with an ARCH process as follows:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_1 x_t + \varepsilon_t,$$

$$\varepsilon_t = \sqrt{h_t}\upsilon_t. \qquad (5.5.1)$$

where v$_t$ is a white noise process with a zero mean and unit variance.

$$E(\upsilon_t) = 0.$$

$$E\left(\upsilon_t \upsilon_{t-s}\right) = \begin{cases} 1 \text{ for } s = 0 \\ 0 \text{ otherwise} \end{cases}.$$

The unconditional variance of ε_t is constant, and the conditional variance of ε_t is h_t—a function of past ε^2_t (Hamilton, 1994). Specifically,

$$h_t = \zeta + \phi_1 \varepsilon^2_{t-1} + \phi_2 \varepsilon^2_{t-2} + \cdots + \phi_m \varepsilon^2_{t-m}. \qquad (5.5.2)$$

A generalization of the ARCH model (GARCH) is to include lags of the conditional variance in Equation 5.5.2 (Bollerslev, 1986):

$$h_t = \zeta + \phi_1 \varepsilon^2_{t-1} + \phi_2 \varepsilon^2_{t-2} + \cdots + \phi_m \varepsilon^2_{t-m} + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \cdots + \delta_r h_{t-r}. \qquad (5.5.3)$$

From this, the squared errors can be represented as follows (Hamilton, 1994):

$$\varepsilon^2_t = \zeta + \left(\delta_1 + \phi_1\right)\varepsilon^2_{t-1} + \left(\delta_2 + \phi_2\right)\varepsilon^2_{t-2} + \cdots + \left(\delta_p + \phi_p\right)\varepsilon^2_{t-p}$$

$$+ \omega_t - \delta_1 \omega_{t-1} - \delta_2 \omega_{t-2} - \cdots - \delta_r \omega_{t-r}, \qquad (5.5.4)$$

where p is the larger of m and r and $\omega_t = \varepsilon^2_t - h_t$. We denote the order of the GARCH as GARCH(r, m). In this representation, ε^2_t follows an ARMA(p, r) process (Bollerslev, 1986). This fact is useful for determining the order of the GARCH model to use, as we will see momentarily. One potential advantage of this extension of the ARCH model is that a higher-order ARCH process can often be represented by a lower-order GARCH process, thereby providing gains in efficiency.

As in the ARCH process, the conditional and unconditional means of ε_t are zero, and the unconditional variance of ε_t is a constant. In particular,

$$E(\varepsilon_t^2) = E(h_t v_t^2) = E(h_t)$$

$$= \frac{\zeta}{1-(\delta_1+\phi_1)-(\delta_2+\phi_2)-\cdots-(\delta_p+\phi_p)}. \qquad (5.5.5)$$

The covariance stationarity requirement for this process is that

$$(\delta_1+\phi_1)+(\delta_2+\phi_2)+\cdots+(\delta_p+\phi_p)<1. \qquad (5.5.6)$$

Just as an ARCH or GARCH process can be included within an LDV model, they can also be included within an ARMA model. Using the two-component representation of the ARMA model (Equation 5.3.1), we can include a GARCH(1,1) process in an ARMA(*p,q)* model as follows:

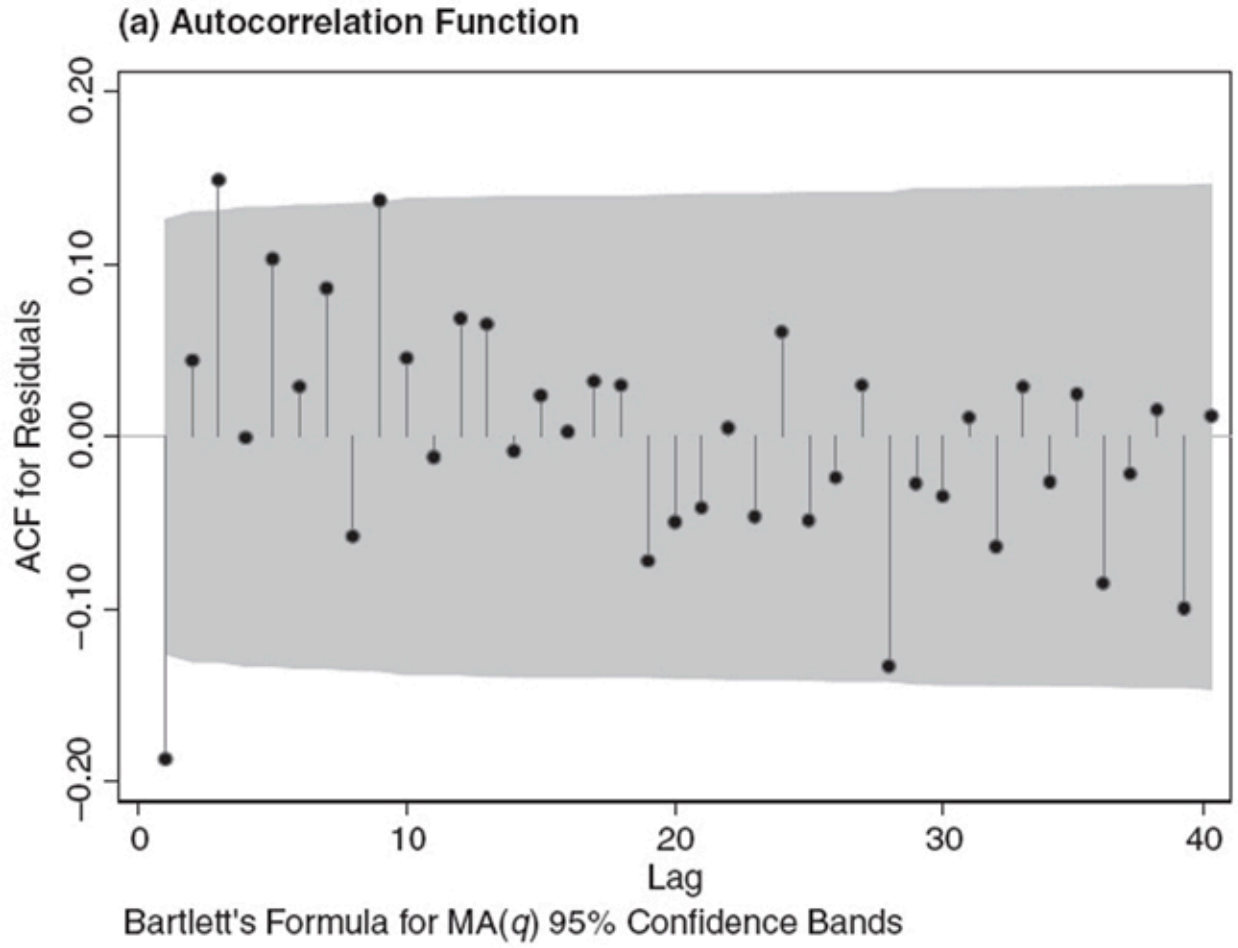
$$y_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \mu_t,$$

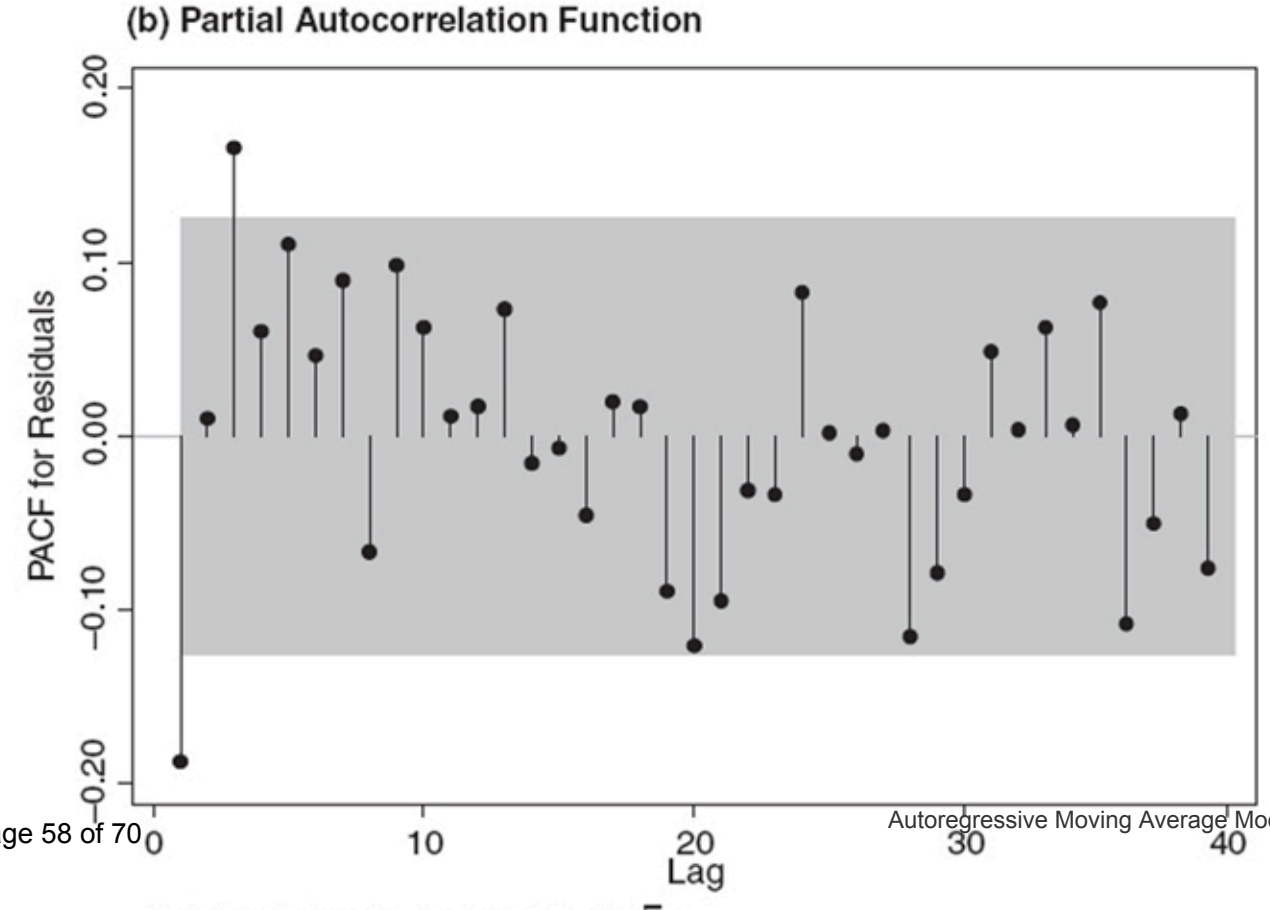
$$\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} + \varepsilon_t.$$

$$\varepsilon_t^2 = \zeta + (\delta_1+\phi_1)\varepsilon_{t-1}^2 + \omega_t + \delta_1 \omega_{t-1}. \qquad (5.5.7)$$

To demonstrate the specification, estimation, and testing of a GARCH model, we return to the German economic approval model we used in Chapter 4 to demonstrate the ARCH model. After estimating the ADL(1,1) model, we might have produced the ACF and PACF for model residuals, as shown in Figure 5.13.

Figure 5.13 Autocorrelation and Partial Autocorrelation Functions for German Economic Approval





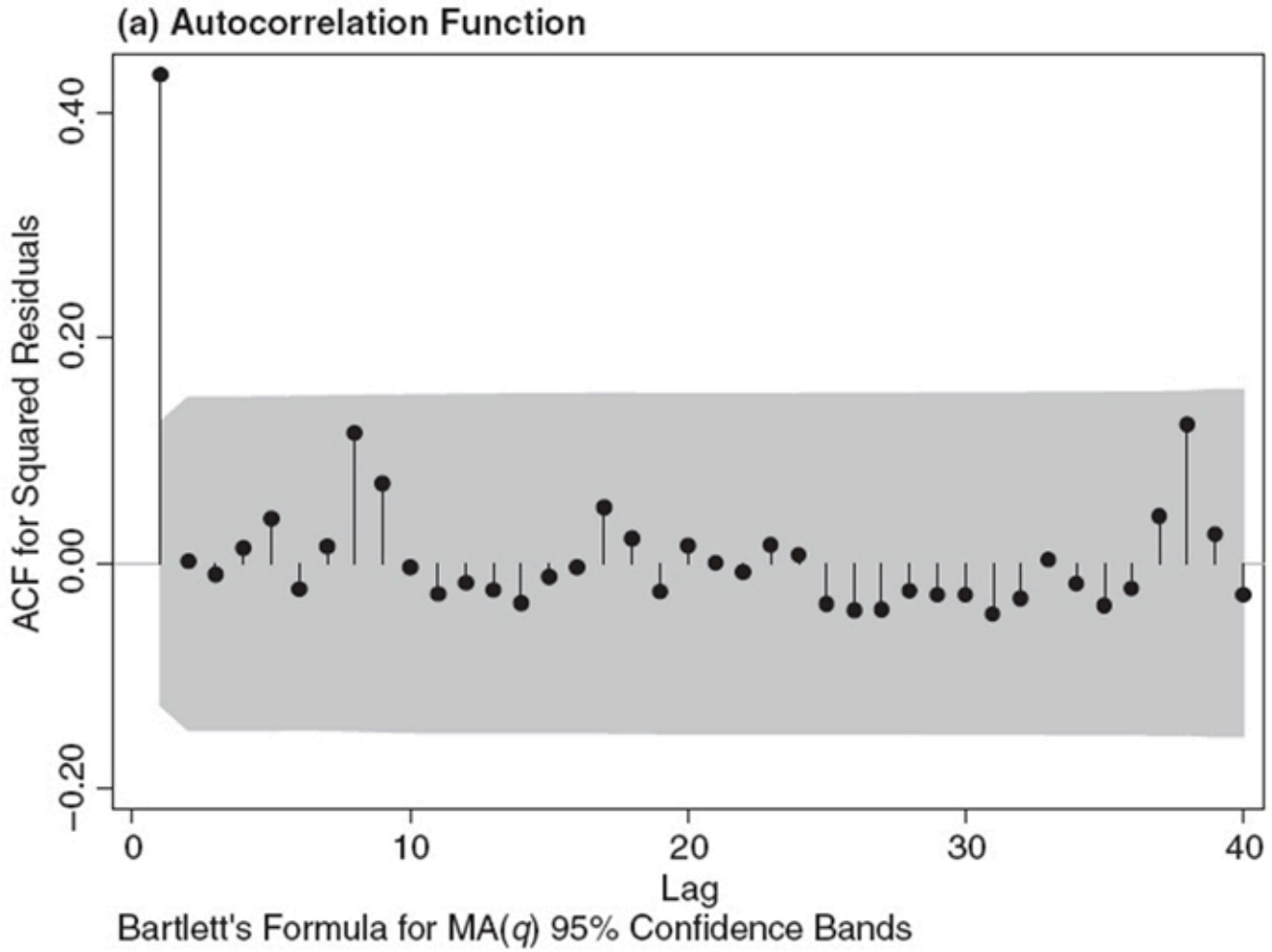
The ACF and PACF suggest that we might have wanted to include either an AR(1) or an MA(1) term in our model, but this model is already autoregressive by including a lagged dependent variable, so we proceed with the ADL(1,1) MA(1) model. The estimates of this model are presented in Table 5.14. This specification produces the desired white noise errors, according to the Portmanteau *Q* statistic. Next, we calculate the autocorrelation and partial autocorrelation functions for the squared residuals of the model to help us determine the specification of the GARCH process. These are shown in Figure 5.14.

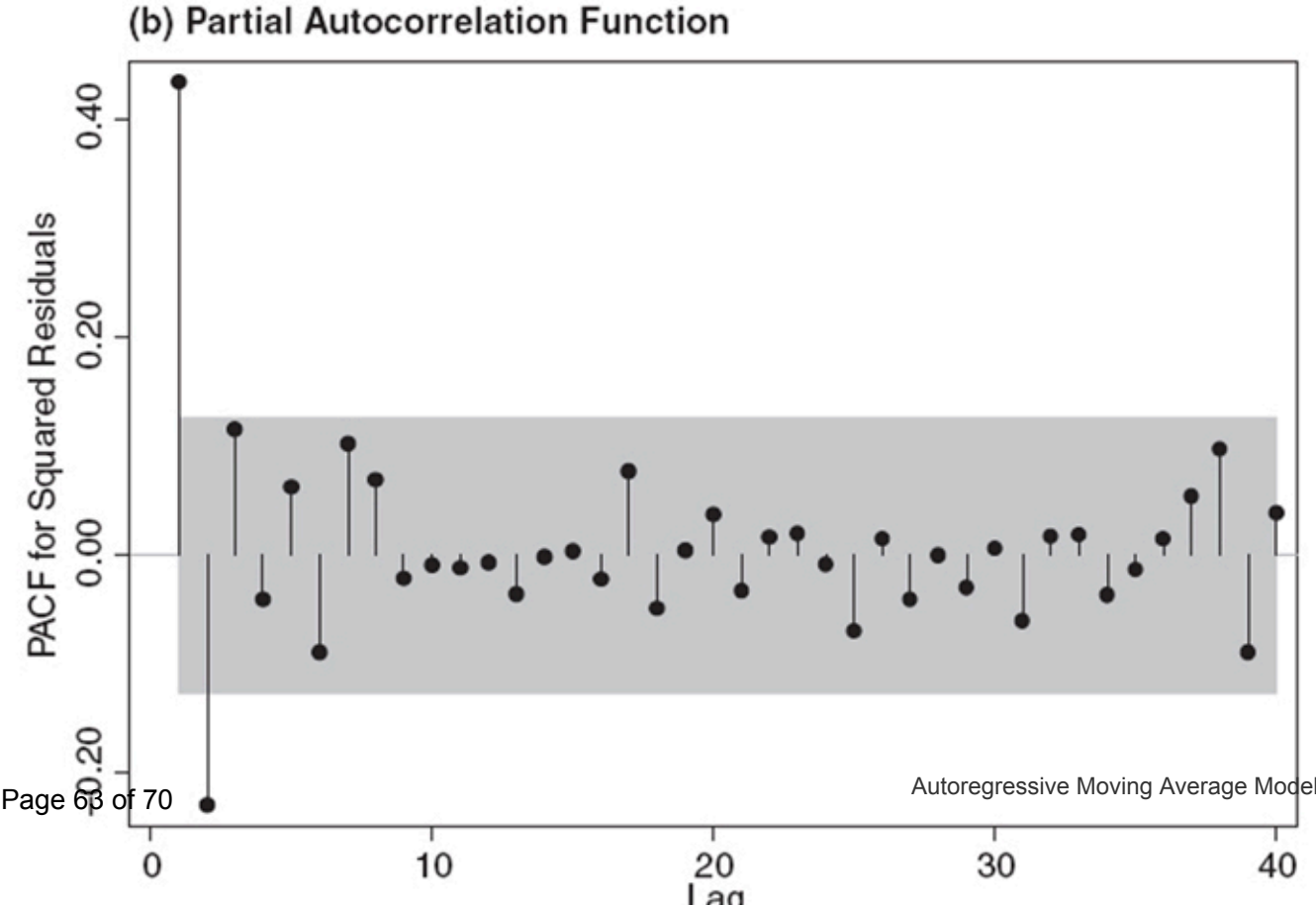
Table 5.14 ADL(1,1) MA(1) German Approval Model With GARCH(1,1) Process

ADL(1,1) MA(1)				ADL(1,1) MA(1) GARCH(1,1)			
Coefficient	*Standard Error*	*z Statistic*	*P Value*	*Coefficient*	*Standard Error*	*z Statistic*	*P Value*
0.90	0.039	23.06	<0.001	0.87	0.044	19.55	<0.001
0.41	0.27	1.54	0.123	0.52	0.26	1.99	0.046
–0.36	2.32	–0.16	0.876	–0.25	3.17	–0.08	0.936
0.13	0.95	0.14	0.888	0.42	0.76	0.56	0.578
–0.36	0.27	–1.36	0.173	–0.46	0.25	–1.83	0.067
0.11	2.33	0.05	0.963	–0.15	3.17	–0.05	0.962
–0.40	0.94	–0.43	0.67	–0.87	0.75	–1.15	0.249
2.20	1.15	1.91	0.056	3.76	1.23	3.07	0.002
–0.33	0.055	–6	<0.001	–0.18	0.098	–1.85	0.064
—	—	—	—	0.17	0.056	3.11	0.002
—	—	—	—	0.62	0.11	5.49	<0.001
—	—	—	—	2.94	1.28	2.30	0.021
Log likelihood = −669.7142				Log likelihood = −653.3762			

= probability, ADL = autoregressive distributed lag, ARCH = autoregressive conditional heteroskedasticity, GARCH = generalized
ional heteroskedasticity, MA = moving average, L1 = first lag.

Figure 5.14 Autocorrelation and Partial Autocorrelation Functions for the Squared Residuals



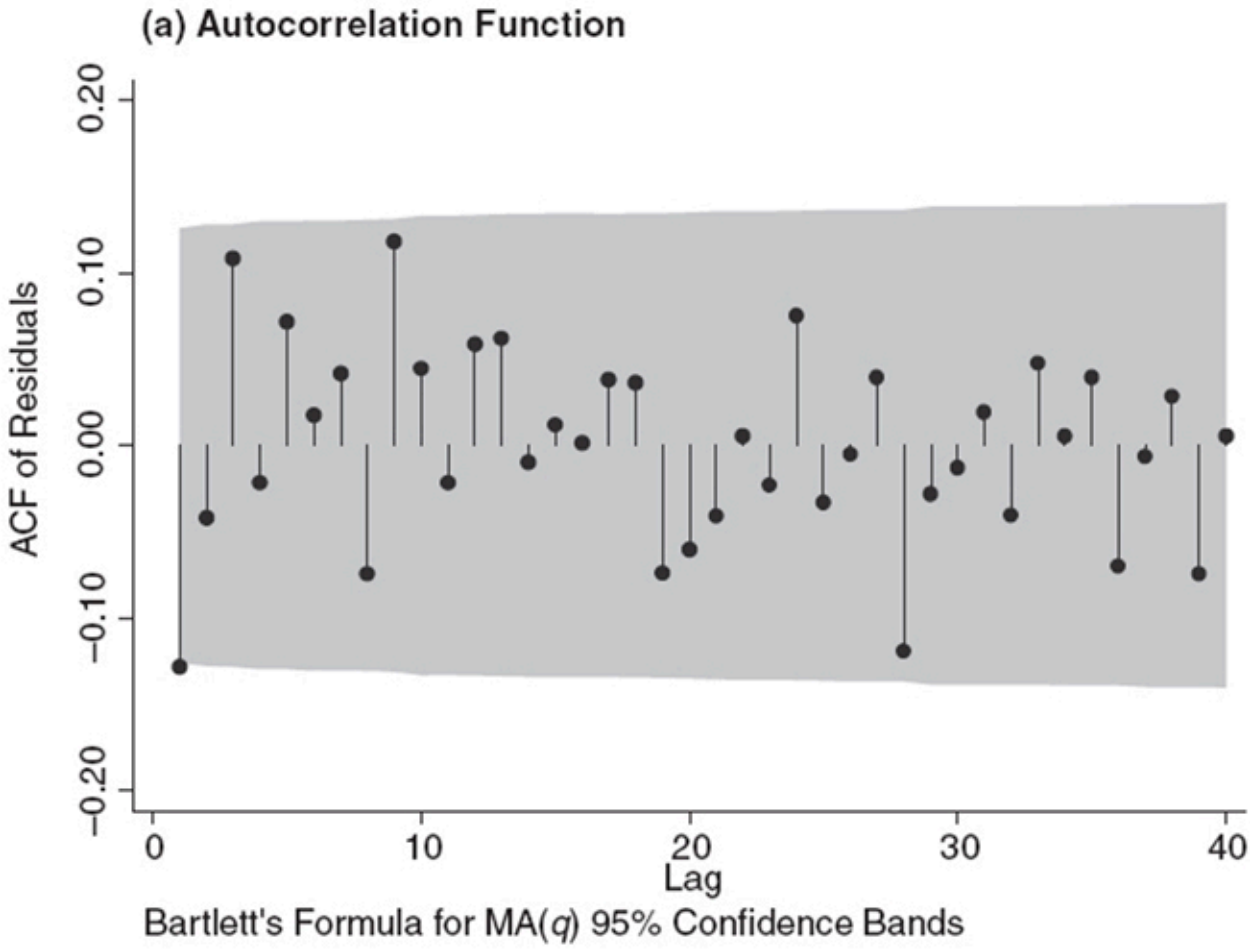


The squared residuals of a GARCH(*r,m*) follow an ARMA(*p,r*) process with *p* = max{*m,r*}. The ACF and PACF of the squared residuals indicate that there is an MA(1) process. Accordingly, we can deduce that *r* = 1. Since *p* = max{*m,r*} and *r* = 1, we would expect *p* to be at least 1. This means that we would expect to see at least an AR(1) component in the squared residuals. This is not evident in Figure 5.14, but it may just be too small to be observed, and/or the MA(1) process in the squared residuals might be obscuring evidence of it.

The value of *m* is more ambiguous in this case. If we assume that *m* = 0, we have the ARCH(1) process we used in our model in Chapter 4. However, there is nothing in Figure 5.14 that rules out the possibility that *m* = 1 and there is no evidence it is any larger. Therefore, we will include a GARCH(1,1) process and compare (Table 5.14).

Testing the residuals, the *Q* statistic is 30.79 with a *P* value of 0.85. We cannot reject the null of a white noise process. The ACF and PACF for the residuals (Figure 5.15) also indicate the residuals are white noise. We next test the adequacy of the specification of the GARCH process by examining the standardized residuals. These should also be a white noise process. We calculate standardized residuals by dividing the estimated errors by the square root of the estimated conditional variance of those errors: $\hat{s}_t = \hat{\varepsilon}_t / \sqrt{\hat{h}_t}$. The ACF and PACF for the standardized residuals are shown in Figure 5.16.

Figure 5.15 Autocorrelation and Partial Autocorrelation Functions for GARCH Residuals



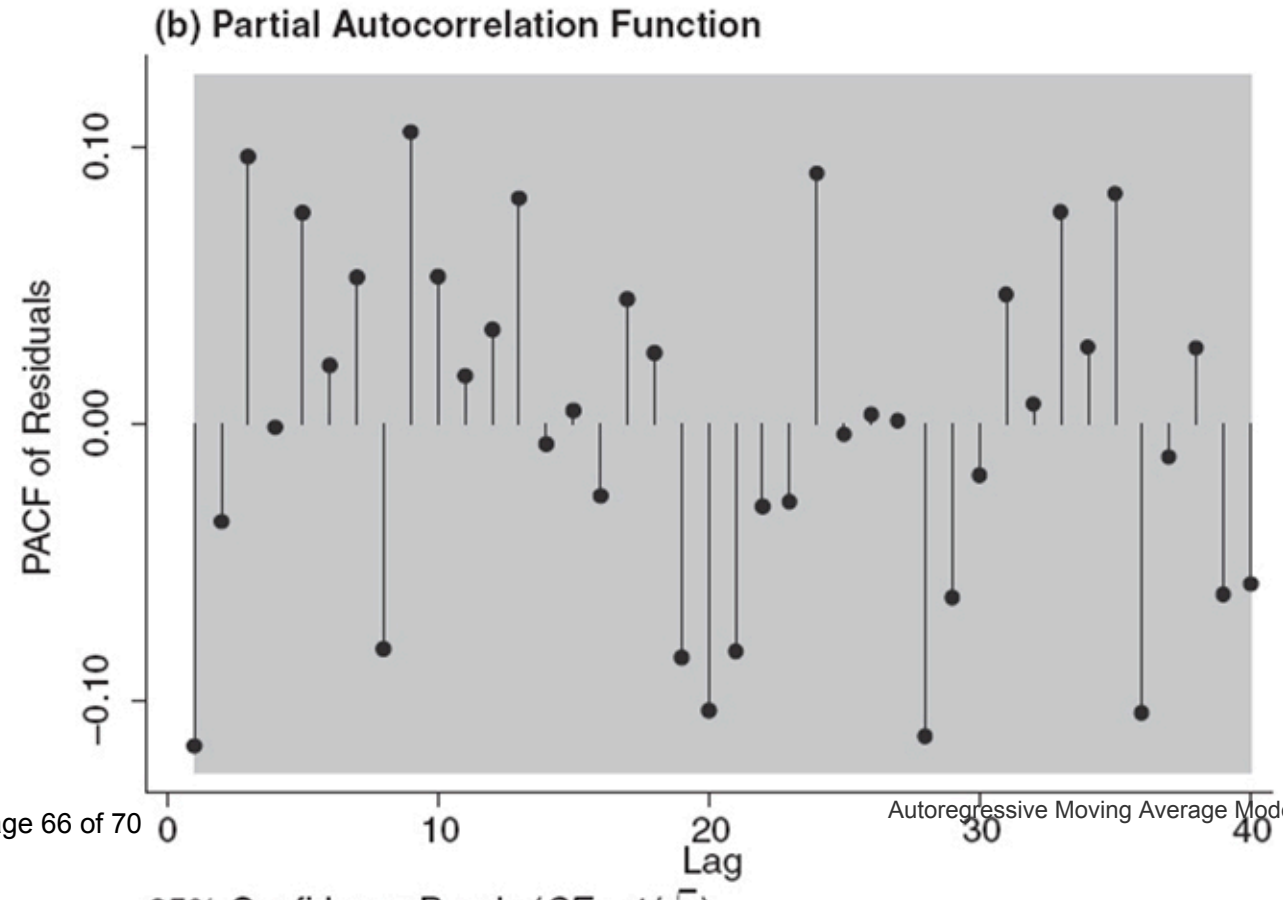
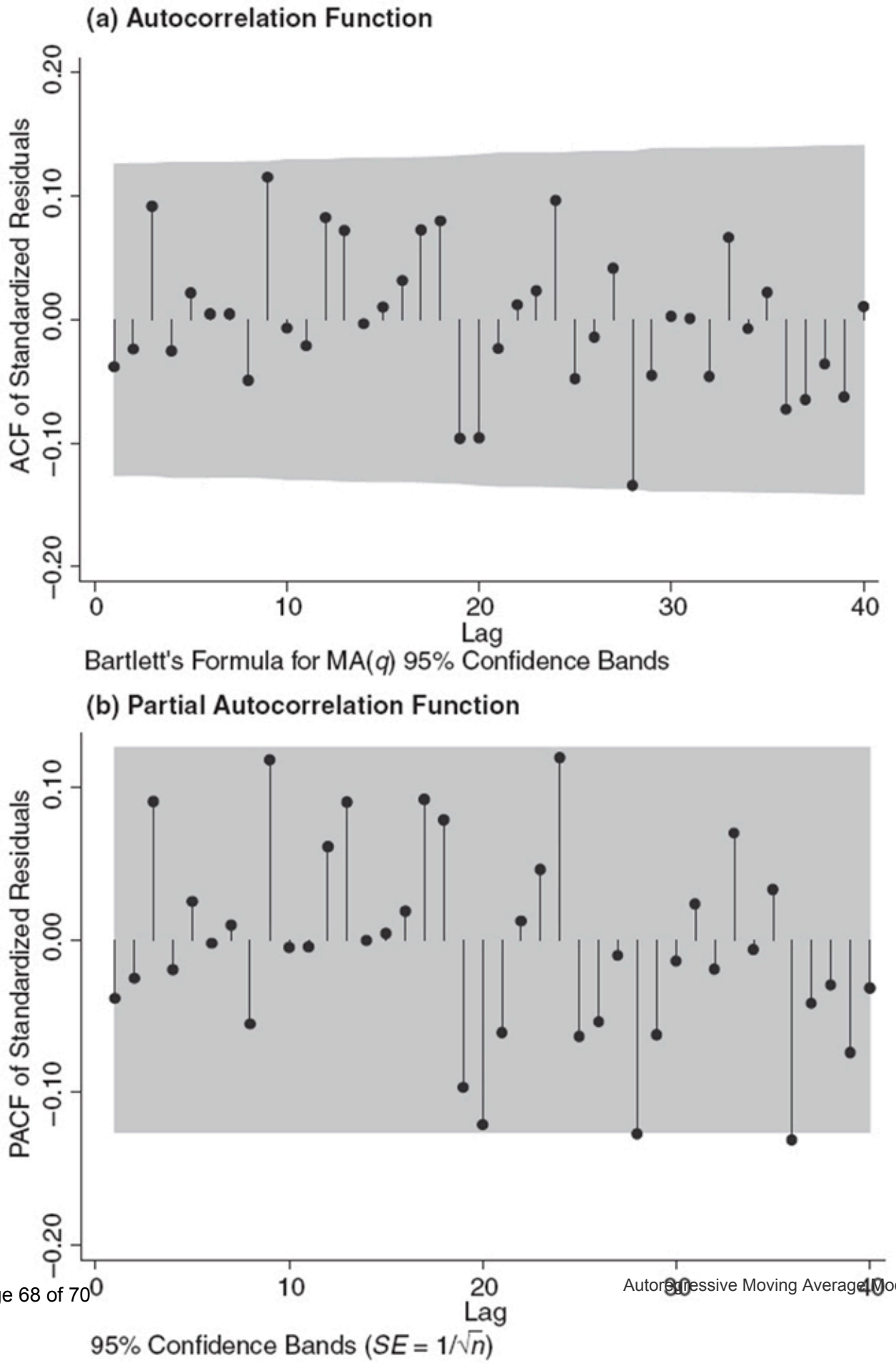


Figure 5.16 Autocorrelation and Partial Autocorrelation Functions for the Standardized Residuals



There is nothing in the ACF and PACF for the standardized residuals to suggest that they are anything but a white noise process. Furthermore, the *Q* statistic is 34.08 with a *P* value of 0.73. If there had been a pattern that suggested dynamics other than white noise, we could use the ACF and PACF for the squared residuals to inform the respecification of the GARCH process.

Having determined that the specification of the GARCH process is adequate, we can compare the results from the ADL(1,1) MA(1) GARCH(1,1) and the ADL(1,1) MA(1) (Table 5.14). Neither of the growth in GDP coefficients are statistically significant (at the 0.5 significance level) in the ADL(1,1) MA(1) model. This is also true for the estimated long-run effect of growth. (This is calculated in the same manner as for any ADL model results.) In the same model, neither inflation nor unemployment has a statistically significant short-run or long-run effect. In the ADL(1,1) MA(1) GARCH(1,1) model, the contemporaneous effect of growth is statistically significant, although the estimated long-run effect is not. Inflation and unemployment do have statistically significant long-run effects of -3.46 and -3.08, respectively. The inclusion of the GARCH(1,1) process in the model reveals that there is a short-lived effect of growth in GDP on government approval. It also reveals the long-run effects of inflation and unemployment.

We now compare the fit of the ADL(1,1) MA(1) GARCH(1,1) with the fit of the ADL(1,1) MA(1) ARCH(1). The second is nested within the first, so we can use the LR test. The test statistic is chi-squared distributed with 1 degree of freedom. It is 3.27, with a corresponding *P* value of 0.0704. We cannot reject the null hypothesis that the ADL(1,1) MA(1) GARCH(1,1) fits no better than the ADL(1,1) MA(1) ARCH(1) at the 0.05 significance level, although we could at the 0.1 significance level. If we tested the standardized residuals from the ADL(1,1) MA(1) ARCH(1) model the tests would suggest that the standardized residuals are a white noise process and that the ARCH(1) process is an adequate specification. The evidence suggests that either model is adequate.

Summary

In this chapter, you have been introduced to the Box-Jenkins approach to model selection. This approach is distinct from the general-to-specific approach discussed in Chapter 4. The Box-Jenkins approach places an emphasis on parsimony and building up a minimal model that meets the estimation assumptions for the data. In contrast, the general-to-specific approach starts with the most general model and only reduces the model through the placement of restrictions if they can be justified by both theory and data. The GARCH model discussed in the last section of this chapter can be incorporated into either approach.

[1] The tests of the null hypothesis of no skewness and of the null hypothesis of no kurtosis (relative to the normal) have *P* values of 0.2807 and 0.699, respectively. We cannot reject the null hypotheses that the skewness and kurtosis of the residuals do not deviate from what is expected for a normal distribution.

[2] The LR test is inappropriate here as the models are not nested.

[3] For full methodological details of the content analysis, see Blake (2010).

[4] Some examples of the use of state-space models in the social sciences are Martin and Quinn (2002), Pickup and Johnston (2008), Armstrong (2008), Jackman (2005), Beck (1989), McAvoy (1998), Brandt and Williams (2001), and Kellstedt, McAvoy, and Stimson (1996).

http://dx.doi.org/10.4135/9781483390857.n5