If we assume that the error terms are normally distributed, the model (10.61) becomes

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathrm{NID}(0, \omega^2). \tag{10.65}$$

We previously made the asymptotically innocuous assumption that the unobserved innovation $\varepsilon_0$ is equal to zero. Although asymptotically innocuous, that assumption is clearly false, since according to (10.65) $\varepsilon_0$ must be distributed as $N(0, \omega^2)$. The simplest way to take proper account of this fact was suggested by MacDonald and MacKinnon (1985); our treatment follows theirs.

The concentrated loglikelihood function for the model (10.65) is

$$C - \frac{n}{2}\log\Big(\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^{\top}\boldsymbol{\Delta}^{-1}(\alpha)\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)\Big) - \frac{1}{2}\log|\boldsymbol{\Delta}(\alpha)|, \tag{10.66}$$

where $\omega^2\boldsymbol{\Delta}(\alpha)$ is the covariance matrix of the vector of error terms $\boldsymbol{u}$, expression (10.60).[4] As discussed by Box and Jenkins (1976) and others, the Jacobian term $-\frac{1}{2}\log|\boldsymbol{\Delta}(\alpha)|$ is

$$\frac{1}{2}\log\big(1 - \alpha^2\big) - \frac{1}{2}\log\big(1 - \alpha^{2n+2}\big). \tag{10.67}$$

When $|\alpha| = 1$, both terms in (10.67) are undefined. In that case, by using l'Hôpital's Rule, one can show that

$$\lim_{|\alpha|\to 1}\Big(\frac{1}{2}\log\big(1 - \alpha^2\big) - \frac{1}{2}\log\big(1 - \alpha^{2n+2}\big)\Big) = -\frac{1}{2}\log(n+1).$$

This result allows the loglikelihood function (10.66) to be evaluated for any value of $\alpha$ in the invertibility region $-1 \leq \alpha \leq 1$.

It is important to be able to deal with the case in which $|\alpha| = 1$, since in practice one not infrequently obtains ML estimates with $|\hat{\alpha}| = 1$, especially when the sample size is small; see, for example, Osborn (1976) and Davidson (1981). The reason for this is that if we concentrate the loglikelihood function with respect to $\boldsymbol{\beta}$ and $\omega$ to obtain $\ell^c(\alpha)$, we will find that $\ell^c(\alpha)$ has the same value for $\alpha$ and $1/\alpha$. That, of course, is the reason for imposing the invertibility condition that $|\alpha| \leq 1$. Thus, if $\ell^c(\alpha)$ is rising as $\alpha \to 1$ or as $\alpha \to -1$, it must have a maximum precisely at $\alpha = 1$ or $\alpha = -1$. This is a distinctly undesirable feature of the model (10.65). When $|\hat{\alpha}| = 1$, one cannot make inferences about $\alpha$ in the usual way, since $\hat{\alpha}$ is then on the boundary of the parameter space. Since $\hat{\alpha}$ can equal $\pm 1$ with finite probability,

[4] In fact, expression (10.66) could be the concentrated loglikelihood function for a nonlinear regression model with error terms that follow any sort of **autoregressive moving average**, or **ARMA**, process, provided that $\boldsymbol{\Delta}(\alpha)$ were replaced by the covariance matrix for $\boldsymbol{u}$ implied by that ARMA process.

using the normal distribution to approximate its finite-sample distribution is a somewhat dubious procedure. Thus, if $\hat{\alpha}$ is equal to or even close to 1 in absolute value, the investigator should exercise care in making inferences about $\alpha$. Of course, as $n \to \infty$ the fact that $\hat{\alpha}$ is consistent means that the number of times that $|\hat{\alpha}| = 1$ tends to zero, unless $|\alpha_0| = 1$.

It is not easy to evaluate (10.66) directly; see Pesaran (1973), Osborn (1976), and Balestra (1980), among others.[5] We therefore use a trick that provides an alternative way to do so. Recall equations (10.62), in which we explicitly wrote $y_1, \ldots, y_n$ as functions of current and lagged values of $x_t(\boldsymbol{\beta})$ and lagged values of $y_t$. We may rewrite these equations, taking account of observation zero, as

$$\begin{aligned} 0 &= -\upsilon + \varepsilon_0 \\ y_1 &= x_1(\boldsymbol{\beta}) - \alpha\upsilon + \varepsilon_1 \\ y_2 &= x_2(\boldsymbol{\beta}) - \alpha\big(y_1 - x_1(\boldsymbol{\beta})\big) - \alpha^2\upsilon + \varepsilon_2 \\ y_3 &= x_3(\boldsymbol{\beta}) - \alpha\big(y_2 - x_2(\boldsymbol{\beta})\big) - \alpha^2\big(y_1 - x_1(\boldsymbol{\beta})\big) - \alpha^3\upsilon + \varepsilon_3, \end{aligned} \tag{10.68}$$

and so on. Here we have added both one observation and one parameter to equations (10.62). The extra observation is observation zero, which as written here simply says that the unknown parameter $\upsilon$ is *defined* to equal the error term $\varepsilon_0$. This unknown parameter also appears in all subsequent observations, multiplied by larger and larger powers of $\alpha$, to reflect the dependence of $y_t$ for all observations on $\varepsilon_0$. Notice that because we have added both an extra parameter and an extra observation, we have not changed the number of degrees of freedom (i.e., the number of observations minus the number of parameters estimated) at all.

If we make the definitions

$$\begin{aligned} &y_0^* = 0; \quad y_t^* = y_t + \alpha y_{t-1}^*, \ \ t = 1, \ldots, n; \\ &x_0^* = 0; \quad x_t^*(\boldsymbol{\beta}, \alpha) = x_t(\boldsymbol{\beta}) + \alpha x_{t-1}^*(\boldsymbol{\beta}, \alpha), \ \ t = 1, \ldots, n; \\ &z_0^* = -1; \quad z_t^* = \alpha z_{t-1}^*, \end{aligned}$$

we can write equations (10.68) in the form

$$y_t^*(\alpha) = x_t^*(\boldsymbol{\beta}, \alpha) + \upsilon z_t^* + \varepsilon_t, \tag{10.69}$$

making them look like very much like a nonlinear regression model. The sum of squared residuals would then be

$$\sum_{t=0}^{n} \big(y_t^*(\alpha) - x_t^*(\boldsymbol{\beta}, \alpha) - \upsilon z_t^*\big)^2. \tag{10.70}$$

[5] Another approach to the estimation of models with moving average errors has been proposed by Harvey and Phillips (1979) and by Gardner, Harvey, and Phillips (1980). It requires specialized software.