simple example. More generally, as discussed by Davidson and MacKinnon (1987), we can allow for drifting DGPs that do not lie within the alternative hypothesis, but that drift toward some fixed DGP in the null hypothesis. It then turns out that, for drifting DGPs that are, in an appropriate sense, equally distant from the null, the noncentrality parameter is maximized by those DGPs that do lie within the alternative hypothesis. This result justifies the intuition that, for a given number of degrees of freedom, tests against an alternative which happens to be true should have more power than tests against other alternatives.

10.7 ML Estimation of Models with Autoregressive Errors

In Section 7.8, we discussed several methods based on generalized or nonlinear least squares for estimating linear regression models with error terms that follow an autoregressive process. An alternative approach is to use maximum likelihood. If it is assumed that the innovations are normally distributed, ML estimation is quite straightforward. With the normality assumption, the model (7.40) considered in Sections 7.7 and 7.8 can be written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2), \tag{10.86}$$

in which the error terms follow an AR(1) process with parameter ρ that is assumed to be less than 1 in absolute value. If we omit the first observation, this model can be rewritten as in equation (7.41). The result is just a nonlinear regression model, and so, as we saw in Section 10.2, the ML estimates of $\boldsymbol{\beta}$ and ρ must coincide with the NLS ones.

Maximum likelihood estimation of (10.86) is more interesting if we do not omit the first observation, because, in that case, the ML estimates no longer coincide with either the NLS or the (feasible) GLS estimates. For observations 2 through n, the contributions to the loglikelihood can be written as in (10.09):

$$\begin{aligned}\ell_t(\boldsymbol{y}^t, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) = \\ -\frac{1}{2}\log 2\pi - \log\sigma_\varepsilon - \frac{1}{2\sigma_\varepsilon^2}(y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta})^2.\end{aligned} \tag{10.87}$$

As required by (10.24), this expression is the log of the density of y_t conditional on the lagged dependent variable y_{t-1}.

For the first observation, the only information we have is that

$$y_1 = \boldsymbol{X}_1\boldsymbol{\beta} + u_1,$$

since the lagged dependent variable y_0 is not observed. However, with the normality assumption, we know from Section 7.8 that $u_1 \sim \mathrm{N}\big(0, \sigma_\varepsilon^2/(1-\rho^2)\big)$. Thus the loglikelihood contribution from the first observation is the log of the density of that distribution, namely,

$$\begin{aligned} &\ell_1(y_1, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) = \\ &\quad -\tfrac{1}{2}\log 2\pi - \log\sigma_\varepsilon + \tfrac{1}{2}\log(1-\rho^2) - \frac{1-\rho^2}{2\sigma_\varepsilon^2}(y_1 - \boldsymbol{X}_1\boldsymbol{\beta})^2. \end{aligned} \tag{10.88}$$

Of course, we are assuming here that $\boldsymbol{X}_1$ is exogenous and therefore uncorrelated with u_1; see the discussion in Section 7.8.

The loglikelihood function for the model (10.86) based on the entire sample is obtained by adding the contribution (10.88) to the sum of the contributions (10.87), for $t = 2, \ldots, n$. The result is

$$\begin{aligned} \ell(\boldsymbol{y}, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) &= -\frac{n}{2}\log 2\pi - n\log\sigma_\varepsilon + \frac{1}{2}\log(1-\rho^2) \\ &- \frac{1}{2\sigma_\varepsilon^2}\Big((1-\rho^2)(y_1 - \boldsymbol{X}_1\boldsymbol{\beta})^2 + \sum_{t=2}^{n}(y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta})^2\Big). \end{aligned} \tag{10.89}$$

The term $\frac{1}{2}\log(1-\rho^2)$ that appears in (10.89) plays an extremely important role in ML estimation. Because it tends to minus infinity as ρ tends to ± 1, its presence in the loglikelihood function ensures that there must be a maximum within the **stationarity region** defined by $|\rho| < 1$. Therefore, maximum likelihood estimation using the full sample is guaranteed to yield an estimate of ρ for which the AR(1) process is stationary. This is not the case for any of the estimation techniques discussed in Section 7.8.

Let us define $u_t(\boldsymbol{\beta})$ as $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ for $t = 1, \ldots, n$, and let $\hat{u}_t = u_t(\hat{\boldsymbol{\beta}})$. Then, from the first-order conditions for the maximization of (10.89), it can be seen that the ML estimators $\hat{\boldsymbol{\beta}}$, $\hat{\rho}$, and $\hat{\sigma}_\varepsilon^2$ satisfy the following equations:

$$\begin{aligned} (1-\hat{\rho}^2)\boldsymbol{X}_1^{\top}\hat{u}_1 + \sum_{t=2}^{n}(\boldsymbol{X}_t - \hat{\rho}\boldsymbol{X}_{t-1})^{\top}(\hat{u}_t - \hat{\rho}\hat{u}_{t-1}) &= \boldsymbol{0}, \\ \hat{\rho}\hat{u}_1^2 - \frac{\hat{\rho}\hat{\sigma}_\varepsilon^2}{1-\hat{\rho}^2} + \sum_{t=2}^{n}\hat{u}_{t-1}(\hat{u}_t - \hat{\rho}\hat{u}_{t-1}) &= 0, \text{ and} \\ \hat{\sigma}_\varepsilon^2 = \frac{1}{n}\Big((1-\hat{\rho}^2)\hat{u}_1^2 + \sum_{t=2}^{n}(\hat{u}_t - \hat{\rho}\hat{u}_{t-1})^2\Big)&. \end{aligned} \tag{10.90}$$

The first two of these equations are similar, but not identical, to the estimating equations (7.71) developed in Section 7.8 for iterated feasible GLS or NLS with account taken of the first observation. In Exercise 10.21, an artificial regression is developed which makes it quite easy to solve equations (10.90). This approach is simpler than the better-known algorithm for finding ML estimates that was proposed by Beach and MacKinnon (1978).