over all t and then taking the logarithm yields the Jacobian term that appears in (8.92).

Concentrating the loglikelihood function with respect to σ yields

$$\begin{aligned}\ell^c(\boldsymbol{\beta},\gamma) = C - \tfrac{n}{2}\log\biggl(\sum_{t=1}^{n}\bigl(y_t^{\gamma} - \beta_0 - \beta_1 x_t\bigr)^2\biggr) \\ + n\log|\gamma| + (\gamma-1)\sum_{t=1}^{n}\log(y_t).\end{aligned} \tag{8.93}$$

Maximizing this with respect to γ and $\boldsymbol{\beta}$ is straightforward. If a suitable nonlinear optimization program is not available, one can simply do a one-dimensional search over γ, calculating β_0 and β_1 conditional on γ by means of least squares, so as to find the value $\hat{\gamma}$ that maximizes (8.93). Of course, one cannot use the OLS covariance matrix obtained in this way, since it treats $\hat{\gamma}$ as fixed. The information matrix is *not* block-diagonal between $\boldsymbol{\beta}$ and the other parameters of (8.91), so one must calculate and invert the full information matrix to obtain an estimated covariance matrix.

ML estimation works in this case because of the Jacobian term that appears in (8.92) and (8.93). It vanishes when $\gamma = 1$ but plays an extremely important role for all other values of γ. We saw in Section 8.1 that if one applied NLS to (8.01) and all the y_t's were greater than unity, one would end up with an infinitely large and negative estimate of γ. That will not happen if one uses maximum likelihood, because the term $(\gamma-1)\sum_{t=1}^{n}\log(y_t)$ will tend to minus infinity as $\gamma \to -\infty$ much faster than $-n/2$ times the logarithm of the sum-of-squares term tends to plus infinity. This example illustrates how useful ML estimation can be for dealing with modified regression models in which the dependent variable is subject to a transformation. We will encounter other problems of this type in Chapter 14.

ML estimation can also be very useful when it is believed that the error terms are nonnormal. As an extreme example, consider the following model:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \alpha\varepsilon_t, \quad f(\varepsilon_t) = \frac{1}{\pi(1+\varepsilon_t^2)}, \tag{8.94}$$

where $\boldsymbol{\beta}$ is a k–vector and $\boldsymbol{X}_t$ is the t^{th} row of an $n \times k$ matrix. The density of ε_t here is the Cauchy density (see Section 4.6) and ε_t therefore has no finite moments. The parameter α is simply a scale parameter, *not* the standard error of the error terms; since the Cauchy distribution has no moments, the error terms do not have a standard error.

If we write ε_t as a function of y_t, we find that

$$\varepsilon_t = \frac{y_t - \boldsymbol{X}_t\boldsymbol{\beta}}{\alpha}.$$