1.5 Relation to Maximum Likelihood

Having specified the distribution of the error vector $\boldsymbol{\varepsilon}$, we can use the **maximum likelihood (ML) principle** to estimate the model parameters $(\boldsymbol{\beta}, \sigma^2)$.[21] In this section, we will show that $\mathbf{b}$, the OLS estimator of $\boldsymbol{\beta}$, is also the ML estimator, and the OLS estimator of σ^2 differs only slightly from the ML counterpart, when the error is normally distributed. We will also show that $\mathbf{b}$ achieves the **Cramer-Rao lower bound**.

The Maximum Likelihood Principle

Just to refresh your memory of basic statistics, we temporarily step outside the classical regression model to review the ML estimation and related concepts. The basic idea of the ML principle is to choose the parameter estimates to maximize the probability of obtaining the data. To be more precise, suppose that we observe an n-dimensional data vector $\mathbf{y} \equiv (y_1, y_2, \ldots, y_n)'$. Assume that the probability density of $\mathbf{y}$ is a member of a family of functions indexed by a finite-dimensional parameter vector $\widetilde{\boldsymbol{\theta}}$: $f(\mathbf{y}; \widetilde{\boldsymbol{\theta}})$. The set of values that $\widetilde{\boldsymbol{\theta}}$ could take is called the **parameter space** and denoted by $\boldsymbol{\Theta}$. (This is described as **parameterizing** the density function.) When the hypothetical parameter vector $\widetilde{\boldsymbol{\theta}}$ equals the true parameter vector $\boldsymbol{\theta}$, $f(\mathbf{y}; \widetilde{\boldsymbol{\theta}})$ becomes the true density of $\mathbf{y}$. We have thus specified a model, a set of possible distributions of $\mathbf{y}$. The model is said to be **correctly specified** if the parameter space $\boldsymbol{\Theta}$ includes the true parameter value $\boldsymbol{\theta}$.

The hypothetical density $f(\mathbf{y}; \widetilde{\boldsymbol{\theta}})$, viewed as a function of the hypothetical parameter vector $\widetilde{\boldsymbol{\theta}}$, is called the **likelihood function** $L(\widetilde{\boldsymbol{\theta}})$. Thus,

$$L(\widetilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y}; \widetilde{\boldsymbol{\theta}}). \tag{1.5.1}$$

The ML estimate of the unknown true parameter vector $\boldsymbol{\theta}$ is the $\widetilde{\boldsymbol{\theta}}$ that maximizes the likelihood function. The maximization is equivalent to maximizing the **log likelihood function** $\log L(\widetilde{\boldsymbol{\theta}})$ because the log transformation is a monotone transformation. Therefore, the ML estimator of $\boldsymbol{\theta}$ can be defined as

$$\text{ML estimator of } \boldsymbol{\theta} \equiv \underset{\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \ \log L(\widetilde{\boldsymbol{\theta}}). \tag{1.5.2}$$

For example, consider the so-called normal location/scale model:

Example 1.6: Suppose the sample $\mathbf{y}$ is an i.i.d. sample from $N(\mu, \sigma^2)$ (the normal distribution with mean μ and variance σ^2). With $\boldsymbol{\theta} \equiv (\mu, \sigma^2)'$, the (joint) density of $\mathbf{y}$ is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]. \tag{1.5.3}$$

We take the parameter space $\boldsymbol{\Theta}$ to be $\Re \times \Re_{++}$, where $\Re_{++}$ is a set of positive real numbers. This just means that μ can be any real number but σ^2 is constrained

[21] For a fuller treatment of maximum likelihood, see Chapter 7.

to be positive. Replacing the true parameter value $\boldsymbol{\theta}$ by its hypothetical value $\widetilde{\boldsymbol{\theta}} \equiv (\widetilde{\mu}, \widetilde{\sigma}^2)'$ and then taking logs, we obtain the log likelihood function:

$$\log L(\widetilde{\boldsymbol{\theta}}) = -\frac{n}{2}\log(2\pi\widetilde{\sigma}^2) - \frac{1}{2\widetilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \widetilde{\mu})^2. \tag{1.5.4}$$

The Score, the Information Matrix, and the Cramer-Rao Bound

Having introduced the log likelihood function, we can now define some concepts related to the ML estimation. The **score function** or the **score** is simply the gradient (vector of partial derivatives) of log likelihood:

$$\text{score:}\ \ \mathbf{s}(\widetilde{\boldsymbol{\theta}}) \equiv \frac{\partial \log L(\widetilde{\boldsymbol{\theta}})}{\partial \widetilde{\boldsymbol{\theta}}}. \tag{1.5.5}$$

The **information matrix**, denoted as $\mathbf{I}(\boldsymbol{\theta})$, is defined as the matrix of second moments of the score *evaluated at the true parameter vector* $\boldsymbol{\theta}$:

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \mathrm{E}[\mathbf{s}(\boldsymbol{\theta})\,\mathbf{s}(\boldsymbol{\theta})']. \tag{1.5.6}$$

The celebrated Cramer-Rao inequality states that the variance of any unbiased estimator is greater than or equal to the inverse of the information matrix.

Cramer-Rao Inequality: *Let* $\mathbf{y}$ *be a vector of random variables (not necessarily independent) the joint density of which is given by* $f(\mathbf{y};\boldsymbol{\theta})$*, where* $\boldsymbol{\theta}$ *is an* K*-dimensional vector of parameters. Let* $L(\widetilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y};\widetilde{\boldsymbol{\theta}})$ *be the likelihood function, and let* $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ *be an unbiased estimator of* $\boldsymbol{\theta}$ *with a finite variance-covariance matrix. Then, under some regularity conditions on* $f(\mathbf{y};\boldsymbol{\theta})$ *(not stated here),*

$$\mathrm{Var}[\widehat{\boldsymbol{\theta}}(\mathbf{y})] \geq \underset{(K\times K)}{\mathbf{I}(\boldsymbol{\theta})^{-1}} \quad (\equiv \textit{Cramer-Rao Lower Bound}),$$

Also under the regularity conditions, the information matrix equals the negative of the expected value of the Hessian (matrix of second partial derivatives) of the log likelihood evaluated at the true parameter vector $\boldsymbol{\theta}$*:*

$$\mathbf{I}(\boldsymbol{\theta}) = -\,\mathrm{E}\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\theta}}\,\partial\widetilde{\boldsymbol{\theta}}'}\right]. \tag{1.5.7}$$

This is called the ***information matrix equality****.*

See, e.g., Amemiya (1985, Theorem 1.3.1) for a proof and a statement of the regularity conditions. Those conditions guarantee that the operations of differentiation and taking expectations can be interchanged. Thus, for example,

$$\mathrm{E}[\partial L(\boldsymbol{\theta})/\partial\widetilde{\boldsymbol{\theta}}] = \partial\,\mathrm{E}[L(\boldsymbol{\theta})]/\partial\widetilde{\boldsymbol{\theta}}.$$

Conditional Likelihood

The theory of maximum likelihood presented above is based on the hypothetical density $f(\mathbf{y};\widetilde{\boldsymbol{\theta}})$. You may have noticed that all that is required for the above results to hold is that $f(\mathbf{y};\widetilde{\boldsymbol{\theta}})$ is *a* density, so the whole discussion can be easily adapted to a conditional density. That is, if the data are $(\mathbf{y},\mathbf{X})$ rather than $\mathbf{y}$, we can develop the conditional version of the theory of ML and related concepts, based on the hypothetical conditional density $f(\mathbf{y} \mid \mathbf{X};\widetilde{\boldsymbol{\theta}})$, as follows.

- The likelihood function $L(\widetilde{\boldsymbol{\theta}})$ is now the conditional likelihood

$$L(\widetilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y} \mid \mathbf{X}; \widetilde{\boldsymbol{\theta}}). \tag{1.5.8}$$

- Just as above, the ML estimator is the $\widetilde{\boldsymbol{\theta}}$ that maximizes this likelihood function. The estimator is called the **conditional ML estimator**. The score function is defined just as above, as the gradient of the log likelihood function.

- The definition of the information matrix is the same as above, except that the expectation is conditional on $\mathbf{X}$:

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \mathrm{E}\left[\mathbf{s}(\boldsymbol{\theta})\, \mathbf{s}(\boldsymbol{\theta})' \mid \mathbf{X}\right]. \tag{1.5.9}$$

  The inverse of this matrix is the Cramer-Rao lower bound.

- Similarly, the expectation in the information matrix equality is conditional on $\mathbf{X}$:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\Big[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \widetilde{\boldsymbol{\theta}}\, \partial \widetilde{\boldsymbol{\theta}}'} \mid \mathbf{X}\Big]. \tag{1.5.10}$$

Given the data $(\mathbf{y}, \mathbf{X})$, we could also construct the theory based on the joint density of $(\mathbf{y}, \mathbf{X})$. A natural question that arises is: what is the relationship between the conditional ML estimator, which is based on the density of $\mathbf{y}$ given $\mathbf{X}$, and the (full or joint) ML estimator based on the joint density of $(\mathbf{y}, \mathbf{X})$? This issue will be briefly discussed at the end of this section and more fully in Chapter 7.

Conditional ML Estimation of the Classical Linear Regression Model

We now return to the classical regression model and derive the conditional ML estimator and the Cramer-Rao lower bound.

The Log Likelihood

As already observed, Assumption 1.5 (the normality assumption) together with Assumptions 1.2 and 1.4 imply that the distribution of $\boldsymbol{\varepsilon}$ conditional on $\mathbf{X}$ is $N(\mathbf{0}, \sigma^2\, \mathbf{I}_n)$ (see (1.4.1)). But since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by Assumption 1.1, we have

$$\mathbf{y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\, \mathbf{I}_n). \tag{1.5.11}$$

Thus, the conditional density of $\mathbf{y}$ given $\mathbf{X}$ is[22]

$$f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big]. \tag{1.5.12}$$

[22]Recall from basic probability theory that the density function for an n-variate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is

$$(2\pi)^{-n/2}\, |\boldsymbol{\Sigma}|^{-1/2} \exp\Big[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\Big].$$

To derive (1.5.12), just set $\mu = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$.

Replacing the true parameters $(\boldsymbol{\beta}, \sigma^2)$ by their hypothetical values $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$ and taking logs, we obtain the log likelihood function:

$$\begin{aligned}\log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widetilde{\sigma}^2) - \frac{1}{2\widetilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widetilde{\sigma}^2) - \frac{1}{2\widetilde{\sigma}^2}SSR(\widetilde{\boldsymbol{\beta}}), \qquad (1.5.13)\end{aligned}$$

where $SSR(\widetilde{\boldsymbol{\beta}}) \equiv (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$ is the sum of squared residuals cosidered in Section 1.2.

ML via Concentrated Likelihood

It is instructive to maximize the log likelihood in two stages. First, maximize over $\widetilde{\boldsymbol{\beta}}$ for any given $\widetilde{\sigma}^2$. The $\widetilde{\boldsymbol{\beta}}$ that maximizes the objective function could (but does not, in the present case of Assumptions 1.1–1.5) depend on $\widetilde{\sigma}^2$. Second, maximize over $\widetilde{\sigma}^2$ taking into account that the $\widetilde{\boldsymbol{\beta}}$ obtained in the first stage could depend on $\widetilde{\sigma}^2$. The log likelihood function in which $\widetilde{\boldsymbol{\beta}}$ is constrained to be the value from the first stage is called the **concentrated log likelihood function** (concentrated with respect to $\widetilde{\beta}$). For the normal log likelihood (1.5.13), the first stage amounts to minimizing the sum of squares $(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$. The $\widetilde{\boldsymbol{\beta}}$ that does it is none other than the OLS estimator $\mathbf{b}$, and the minimized sum of squares is $\mathbf{e}'\mathbf{e}$. Thus, the concentrated log likelihood is

$$\text{concentrated log likelihood} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widetilde{\sigma}^2) - \frac{1}{2\widetilde{\sigma}^2}\mathbf{e}'\mathbf{e}. \qquad (1.5.14)$$

This is a function of $\widetilde{\sigma}^2$ alone, and the $\widetilde{\sigma}^2$ that maximizes the concentrated likelihood is the ML estimate of σ^2. The maximization is straightforward for the present case of the classical regression model, because $\mathbf{e}'\mathbf{e}$ is not a function of $\widetilde{\sigma}^2$ and so can be taken as a constant. Still, taking the derivative with respect to $\widetilde{\sigma}^2$, rather than with respect to $\widetilde{\sigma}$, can be tricky. This can be avoided by denoting $\widetilde{\sigma}^2$ by $\widetilde{\gamma}$. Taking the derivative of (1.5.14) with respect to $\widetilde{\gamma}$ $(\equiv \widetilde{\sigma}^2)$ and setting it to zero, we obtain the following result.

Proposition 1.5 (ML Estimator of $(\boldsymbol{\beta}, \boldsymbol{\sigma^2})$): *Suppose Assumptions 1.1–1.5 hold. Then the ML estimator of $\boldsymbol{\beta}$ is the OLS estimator* $\mathbf{b}$ *and*

$$\textit{ML estimator of } \sigma^2 = \frac{1}{n}\mathbf{e}'\mathbf{e} = \frac{SSR}{n} = \frac{n-K}{n}s^2. \qquad (1.5.15)$$

We know from Proposition 1.2 that s^2 is unbiased. Since s^2 is multiplied by a factor $(n-K)/n$ which is different from 1, the ML estimator of σ^2 is biased, although the bias becomes arbitrarily small as the sample size n increases for any given fixed K.

For later use, we calculate the maximized value of the likelihood function. Substituting (1.5.15) into (1.5.14), we obtain

$$\text{maximized log likelihood} = -\frac{n}{2}\log\Big(\frac{2\pi}{n}\Big) - \frac{n}{2} - \frac{n}{2}\log(SSR),$$

so that the maximized likelihood is

$$\max_{\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2} L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) = \Big(\frac{2\pi}{n}\Big)^{-n/2} \cdot \exp\Big(-\frac{n}{2}\Big) \cdot (SSR)^{-n/2}. \qquad (1.5.16)$$

The Cramer-Rao Bound

Now, for the classical regression model (of Assumptions 1.1–1.5), the likelihood function $L(\widetilde{\boldsymbol{\theta}})$ in the Cramer-Rao inequality is the conditional density (1.5.12), so the variance in the inequality is the variance conditional on $\mathbf{X}$. It can be shown that those regularity conditions are satisfied for the normal density (1.5.12) (see, e.g., Amemiya, 1985, Sections 1.3.2 and 1.3.3). We now calculate the information matrix for the classical regression model. The parameter vector $\boldsymbol{\theta}$ is $(\boldsymbol{\beta}', \sigma^2)'$. So $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}', \widetilde{\gamma})'$ and the matrix of second derivatives we seek to calculate is

$$\underset{((K+1)\times(K+1))}{\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\theta}}\,\partial\widetilde{\boldsymbol{\theta}}'}} = \begin{bmatrix} \underset{(K\times K)}{\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\beta}}\,\partial\widetilde{\boldsymbol{\beta}}'}} & \underset{(K\times 1)}{\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\beta}}\,\partial\widetilde{\gamma}}} \\ \underset{(1\times K)}{\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\gamma}\,\partial\widetilde{\boldsymbol{\beta}}'}} & \underset{(1\times 1)}{\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\gamma}^2}} \end{bmatrix}. \tag{1.5.17}$$

The first and second derivatives of the log likelihood (1.5.13) with respect to $\widetilde{\boldsymbol{\theta}}$, evaluated at the true parameter vector $\boldsymbol{\theta}$, are

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\beta}}} = \frac{1}{\gamma}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{1.5.18a}$$

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial\widetilde{\gamma}} = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.5.18b}$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\beta}}\,\partial\widetilde{\boldsymbol{\beta}}'} = -\frac{1}{\gamma}\mathbf{X}'\mathbf{X}, \tag{1.5.19a}$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\gamma}^2} = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{1.5.19b}$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\beta}}\,\partial\widetilde{\gamma}} = -\frac{1}{\gamma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.5.19c}$$

Since the derivatives are evaluated at the true parameter value, we have $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\varepsilon}$ in these expressions. Substituting (1.5.19) into (1.5.17) and using $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$ (Assumption 1.2), $\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \mathbf{X}) = n\sigma^2$ (implication of Assumption 1.4), and recalling $\gamma = \sigma^2$, we can easily derive

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\sigma^4} \end{bmatrix}. \tag{1.5.20}$$

Here, the expectation is conditional on $\mathbf{X}$ because the likelihood function (1.5.12) is a conditional density conditional on $\mathbf{X}$. This block diagonal matrix can be inverted to obtain the Cramer-Rao bound:

$$\text{Cramer-Rao bound} \equiv \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\sigma^4}{n} \end{bmatrix}. \tag{1.5.21}$$

Therefore, the unbiased estimator $\mathbf{b}$, whose variance is $\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ by Proposition 1.1, attains the Cramer-Rao bound. We have thus proved

Proposition 1.6 (b is the Best Unbiased Estimator (BUE)): *Under Assumptions 1.1–1.5, the OLS estimator* $\mathbf{b}$ *of* $\boldsymbol{\beta}$ *is BUE in that any other unbiased (but not necessarily linear) estimator has larger conditional variance in the matrix sense.*

This result should be distinguished from the Gauss-Markov Theorem that $\mathbf{b}$ is minimum variance among those estimators that are unbiased *and* linear in $\mathbf{y}$. Proposition 1.6 says that $\mathbf{b}$ is minimum variance in a larger class of estimators that includes nonlinear unbiased estimators. This stronger statement is obtained under the normality assumption (Assumption 1.5) which is not assumed in the Gauss-Markov Theorem. Put differently, the Gauss-Markov Theorem does not exclude the possibility of some nonlinear estimator beating OLS, but this possibility is ruled out by the normality assumption.

As was already seen, the ML estimator of σ^2 is biased, so the Cramer-Rao bound does not apply. But the OLS estimator s^2 of σ^2 is unbiased. Does it achieve the bound? We have shown in a review question to the previous section that

$$\operatorname{Var}(s^2 \mid \mathbf{X}) = \frac{2\sigma^4}{n-K}$$

under the same set of assumptions as in Proposition 1.6. Therefore, s^2 does not attain the Cramer-Rao bound $2\sigma^4/n$. However, it can be shown that an unbiased estimator of σ^2 with variance lower than $2\sigma^4/(n-K)$ does not exist (see, e.g., Rao, 1973, p. 319).

The F-Test as a Likelihood Ratio Test

The **likelihood ratio test** of the null hypothesis compares L_U, the maximized likelihood without the imposition of the restriction specified in the null hypothesis, with L_R, the likelihood maximized subject to the restriction. If the likelihood ratio $\lambda \equiv L_U/L_R$ is too large, it should be a sign that the null is false. The F-test of the null hypothesis $H_0\colon \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ considered in the previous section is a likelihood ratio test because the F-ratio is a monotone transformation of the likelihood ratio λ. For the present model, L_U is given by (1.5.16) where the SSR, the sum of squared residuals minimized without the constraint H_0, is the SSR_U in (1.4.11). The restricted likelihood L_R is given by replacing this SSR by the restricted sum of squared residuals, SSR_R. So

$$L_R = \max_{\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2 \text{ s.t. } H_0} L(\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2) = \Big(\frac{2\pi}{n}\Big)^{-n/2} \cdot \exp\Big(-\frac{n}{2}\Big) \cdot (SSR_R)^{-n/2}, \tag{1.5.22}$$

and the likelihood ratio is

$$\lambda \equiv \frac{L_U}{L_R} = \left(\frac{SSR_U}{SSR_R}\right)^{-n/2}. \tag{1.5.23}$$

Comparing this with the formula (1.4.11) for the F-ratio, we see that the F-ratio is a monotone transformation of the likelihood ratio λ:

$$F = \frac{n-K}{\#\mathbf{r}}(\lambda^{2/n} - 1), \tag{1.5.24}$$

so that the two tests are the same.

Quasi-Maximum Likelihood

All these results assume the normality of the error term. Without normality, there is no guarantee that the ML estimator of $\boldsymbol{\beta}$ is OLS (Proposition 1.5) or that the OLS estimator $\mathbf{b}$ achieves the Cramer-Rao bound (Proposition 1.6). However, Proposition 1.5 does imply that $\mathbf{b}$ is a **quasi-** (or **pseudo-**) **maximum likelihood estimator**, an estimator that maximizes a misspecified likelihood function. The misspecified likelihood function we have considered is the normal likelihood. The results of Section 1.3 can then be interpreted as providing the finite-sample properties of the quasi-ML estimator when the error is incorrectly specified to be normal.

Conditional vs. Joint ML

Since a (joint) density is the product of a marginal density and a conditional density, the joint density of $(\mathbf{y}, \mathbf{X})$ can be written as

$$f(\mathbf{y}, \mathbf{X}; \boldsymbol{\zeta}) = f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}) \cdot f(\mathbf{X}; \boldsymbol{\psi}), \tag{1.5.25}$$

where $\boldsymbol{\theta}$ is the subset of the parameter vector $\boldsymbol{\zeta}$ that determines the conditional density function and $\boldsymbol{\psi}$ is the subset determining the marginal density function. For the linear regression model with normal errors, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ and $f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta})$ is given by (1.5.12).

Let $\widetilde{\boldsymbol{\zeta}} \equiv (\widetilde{\boldsymbol{\theta}}', \widetilde{\boldsymbol{\psi}}')'$ be a hypothetical value of $\boldsymbol{\zeta} = (\boldsymbol{\theta}', \boldsymbol{\psi}')'$. Then the (full or joint) likelihood function is

$$f(\mathbf{y}, \mathbf{X}; \widetilde{\boldsymbol{\zeta}}) = f(\mathbf{y} \mid \mathbf{X}; \widetilde{\boldsymbol{\theta}}) \cdot f(\mathbf{X}; \widetilde{\boldsymbol{\psi}}). \tag{1.5.26}$$

If we knew the parametric form of $f(\mathbf{X}; \widetilde{\boldsymbol{\psi}})$, then we could maximize this joint likelihood function over the entire hypothetical parameter vector $\widetilde{\boldsymbol{\zeta}}$, and the ML estimate of $\boldsymbol{\theta}$ would be the elements of the ML estimate of $\boldsymbol{\zeta}$. We cannot do this for the classical regression model because the model does not specify $f(\mathbf{X}; \widetilde{\boldsymbol{\psi}})$. However, if there is no functional relationship between $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\psi}}$ (such as a subset of $\widetilde{\boldsymbol{\psi}}$ being a function of $\widetilde{\boldsymbol{\theta}}$), then maximizing (1.5.26) with respect to $\widetilde{\boldsymbol{\zeta}}$ is achieved by separately maximizing $f(\mathbf{y} \mid \mathbf{X}; \widetilde{\boldsymbol{\theta}})$ with respect to $\widetilde{\boldsymbol{\theta}}$ and maximizing $f(\mathbf{X}; \widetilde{\boldsymbol{\psi}})$ with respect to $\widetilde{\boldsymbol{\psi}}$. Thus, in this case of no functional relationship between $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\psi}}$, the conditional ML estimate of $\boldsymbol{\theta}$ is numerically equal to the joint ML estimate of of $\boldsymbol{\theta}$.

QUESTIONS FOR REVIEW

1. (Use of regularity conditions) Assuming that taking expectations (i.e., taking integrals) and differentiation can be interchanged, prove that the expected value of the score vector given in (1.5.5), if evaluated at the true parameter value $\boldsymbol{\theta}$, is zero. **Hint:** What needs to be shown is that

$$\int \frac{\partial \log f(\mathbf{y}; \boldsymbol{\theta})}{\partial \widetilde{\boldsymbol{\theta}}} f(\mathbf{y}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{y} = \mathbf{0}.$$

Since $f(\mathbf{y}; \widetilde{\boldsymbol{\theta}})$ is a density, $\int f(\mathbf{y}, \widetilde{\boldsymbol{\theta}}) \, \mathrm{d}\mathbf{y} = 1$ for any $\widetilde{\boldsymbol{\theta}}$. Differentiate both sides with respect to $\widetilde{\boldsymbol{\theta}}$ and use the regularity conditions, which allows us to change the

order of integration and differentiation, to obtain $\int [\partial f(\mathbf{y};\boldsymbol{\theta})/\partial\widetilde{\boldsymbol{\theta}}]\,\mathrm{d}\mathbf{y} = \mathbf{0}$. Also, from basic calculus,

$$\frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\theta}}} = \frac{1}{f(\mathbf{y};\boldsymbol{\theta})}\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial\widetilde{\boldsymbol{\theta}}}.$$

2. (Concentrated log likelihood with respect to $\widetilde{\sigma}^2$) Writing $\widetilde{\sigma}^2$ as $\widetilde{\gamma}$, the log likelihood function for the classical regression model is

$$\log L(\widetilde{\boldsymbol{\beta}},\widetilde{\gamma}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widetilde{\gamma}) - \frac{1}{2\widetilde{\gamma}}(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}).$$

   In the two-step maximization procedure described in the text, we first maximized this function with respect to $\widetilde{\boldsymbol{\beta}}$. Instead, first maximize with respect to $\widetilde{\gamma}$ given $\widetilde{\boldsymbol{\beta}}$. Show that the concentrated log likelihood (concentrated with respect to $\widetilde{\gamma} \equiv \widetilde{\sigma}^2$) is

$$-\frac{n}{2}[1+\log(2\pi)] - \frac{n}{2}\log\left(\frac{(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})}{n}\right).$$

3. (Information matrix equality for classical regression model) Verify (1.5.10) for the linear regression model. **Hint:** If $\varepsilon_i \sim N(0,\sigma^2)$, then $\mathrm{E}(\varepsilon_i^3) = 0$ and $\mathrm{E}(\varepsilon_i^4) = 3\sigma^4$.

4. (Likelihood equations for classical regression model) We used the two-step procedure to derive the ML estimate for the classical regression model. An alternative way to find the ML estimator is to solve for the first-order conditions that set (1.5.18) equal to zero (the first-order conditions for the log likelihood is called the **likelihood equations**). Verify that the ML estimator given in Proposition 1.5 solves the likelihood equations.

5. (Maximizing joint log likelihood) Consider maximizing (the log of) the joint likelihood (1.5.26) for the classical regression model, where $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}', \widetilde{\sigma}^2)'$ and $\log f(\mathbf{y} \mid \mathbf{X};\widetilde{\boldsymbol{\theta}})$ is given by (1.5.13). You would parameterize the marginal likelihood $f(\mathbf{X};\widetilde{\boldsymbol{\psi}})$ and take the log of (1.5.26) to obtain the objective function to be maximized over $\boldsymbol{\zeta} \equiv (\boldsymbol{\theta}', \boldsymbol{\psi}')'$. What is the ML estimator of $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$? [Answer: It should be the same as that in Proposition 1.5.] Derive the Cramer-Rao bound for $\boldsymbol{\beta}$. **Hint:** By the information matrix equality,

$$\mathbf{I}(\boldsymbol{\zeta}) = -\,\mathrm{E}\left[\frac{\partial^2 \log L(\boldsymbol{\zeta})}{\partial\widetilde{\boldsymbol{\zeta}}\,\partial\widetilde{\boldsymbol{\zeta}}'}\right].$$

   Also, $\partial^2 \log L(\boldsymbol{\zeta})/(\partial\widetilde{\boldsymbol{\theta}}\,\partial\widetilde{\boldsymbol{\psi}}') = \mathbf{0}$.

References

Amemiya, T., 1985, *Advanced Econometrics*, Cambridge: Harvard University Press.

Averch, H., and L. Johnson, 1962, "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52, 1052–1069.

Christensen, L., and W. Greene, 1976, "Economies of Scale in US Electric Power Generation," *Journal of Political Economy*, 84, 655–676.

Christensen, L., D. Jorgenson, and L. Lau, 1973, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, 55, 28–45.

Davidson, R., and J. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford: Oxford University Press.

DeLong, B., and L. Summers, 1991, "Equipment Investment and Growth," *Quarterly Journal of Economics*, 99, 28–45.

Engle, R., D. Hendry, and J.-F. Richards, 1983, "Exogeneity," *Econometrica*, 51, 277–304.

Federal Power Commission, 1956, *Statistics of Electric Utilities in the United States, 1955, Class A and B Privately Owned Companies*, Washington, D.C.

Jorgenson, D., 1963, "Capital Theory and Investment Behavior," *American Economic Review*, 53, 247–259.

Koopmans, T., and W. Hood, 1953, "The Estimation of Simultaneous Linear Economic Relationships," in W. Hood, and T. Koopmans (eds.), *Studies in Econometric Method*, New Haven: Yale University Press.

Krasker, W., E. Kuh, and R. Welsch, 1983, "Estimation for Dirty Data and Flawed Models," Chapter 11 in Z. Griliches, and M. Intriligator (eds.), *Handbook of Econometrics*, Volume 1, Amsterdam: North-Holland.

Nerlove, M., 1963, "Returns to Scale in Electricity Supply," in C. Christ (ed.), *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford: Stanford University Press.

Rao, C. R., 1973, *Linear Statistical Inference and Its Applications* (2d ed.), New York: Wiley.

Scheffe, H., 1959, *The Analysis of Variance*, New York: Wiley.

Wolak, F., 1994, "An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction," *Annales D'Economie et de Statistique*, 34, 13–69.