CHAPTER 6. ESTIMATION

6.1. DESIRABLE PROPERTIES OF ESTIMATORS

6.1.1 Consider data **x** that comes from a *data generation process* (DGP) that has a density f(**x)**. Suppose we do not know f(·), but do know (or assume that we know) that f(·) is a member of a family of densities ***G***. The estimation problem is to use the data **x** to select a member of ***G*** which is some appropriate sense is close to the true f(·). Suppose we index the members of ***G*** by the elements of some set **Θ**, and identify f (·) with a particular index value θ_o. Then, another way of stating the estimation problem is that in the family of densities f(x,θ) parameterized by $\theta \in \mathbf{\Theta}$, we want to use the data **x** to estimate the true parameter value θ_o. The parameterization chosen for an estimation problem is not necessarily unique; i.e., there may be more than one way to parameterize the same family of densities ***G***. Sometimes this observation can be used to our advantage, by choosing parameterizations that simplify a problem. However, a parameterization can create difficulties. For example, you might set up **Θ** in such a way that more than one value of θ picks out the true density f; e.g., for some $\theta_o \neq \theta_1$, one has $f(x,\theta_o) = f(x,\theta_1)$ for all x. Then you are said to have an *identification problem*. Viewed within the context of a particular parameterization, identification problems cause real statistical difficulties and have to be dealt with. Viewed from the standpoint of the fundamental estimation problem, they are an artificial consequence of an unfortunate choice of parameterization. Another possible difficulty is that the family of densities generated by your parametric specification f(x,θ), $\theta \in \mathbf{\Theta}$, may fail to coincide with ***G***. A particularly critical question is whether the true f(·) is in fact in your parametric family. You cannot be sure that it is unless your family contains all of ***G***. Classical statistics always assumes that the true density is in the parametric family, and we will start from that assumption too. In Chapter 28, we will ask what the statistical properties and interpretation of parameter estimates are when the true f is not in the specified parametric family. A related question is whether your parametric family contains densities that are not in ***G***. This can affect the properties of statistical inference; e.g., degrees of freedom for hypothesis tests and power calculations.

In basic statistics, the parameter θ is assumed to be a scalar, or possibly a finite-dimensional vector. This will cover many important applications, but it is also possible to consider problems where θ is infinite-dimensional. It is customary to call estimation problems where θ is finite-dimensional *parametric*, and problems where θ is infinite-dimensional *semiparametric* or *nonparametric*. (It would have been more logical to call them “finite-parametric” and “infinite-parametric”, respectively, but the custom is too ingrained to change.) Several chapters in the latter half of this book, particularly Chapter 28, deal with infinite-parameter problems.

6.1.2. In most initial applications, we will think of **x** as a simple random sample of size n, $\mathbf{x} = (x_1,...,x_n)$, drawn from a population in which x has a density $f(x,\theta_o)$, so that the DGP density is $f(\mathbf{x},\theta) = f(x_1,\theta_o)\cdot...\cdot f(x_n,\theta_o)$. However, the notation $f(\mathbf{x},\theta_o)$ can also cover more complicated DGP, such as time-series data sets in which the observations are serially correlated. Suppose that θ_o is an unknown k×1 vector, but one knows that this DGP is contained in a family with densities $f(\mathbf{x},\theta)$

indexed by $\theta \in \mathbf{\Theta}$. An important leading case is k = 1, so that θ_o is a scalar. For many of the topics in this Chapter, it is useful to concentrate first on this case, and postpone dealing with the additional complications introduced by having a vector of parameters. However, we will use definitions and notation that cover the vector as well as the scalar case. Let $\boldsymbol{X}$ denote the domain of **x**, and Θ denote the domain of θ. In the case of a simple random sample where an observation x is a point in a space X, one has $\boldsymbol{X} = \boldsymbol{X}^n$. The statistical inference task is to estimate θ_o. In Chapter 5, we saw that an estimator T(**x**) of θ_o was desirable from a Bayesian point of view if T(·) minimized the expected cost of mistakes. For typical cost functions where the larger the mistake, the larger the cost, Bayes estimators will try to get "close" to the true parameter value. That is, the Bayes procedure will seek estimators whose probability densities are concentrated tightly around the true θ_o. Classical statistical procedures lack the expected cost criterion for choosing estimators, but also seek estimators whose probability densities are near the true density $f(\mathbf{x},\theta_o)$.

In this Chapter, we will denote the expectation of a function $r(\mathbf{x},\gamma)$ of **x** and a vector of “parameters” γ by **E** $r(\mathbf{x},\gamma)$, or when it is necessary to identify the parameter vector of the true DGP, by $\mathbf{E}_{x|\theta}r(\mathbf{x},\gamma) = \int_{-\infty}^{+\infty} r(\mathbf{x},\gamma)\cdot f(\mathbf{x},\theta)d\mathbf{x}$. Sometimes, the notation $\mathbf{E}_{x|\theta}r(\mathbf{x},\gamma)$ is abbreviated to $\mathbf{E}_{\theta}r(\mathbf{x},\gamma)$. This notation also applies when the parameters γ are also in $\mathbf{\Theta}$. Then $\mathbf{E}_{x|\theta}r(\mathbf{x},\theta)$ is the expectation of $r(\mathbf{x},\gamma)$ when γ is set equal to the true parameter vector θ, and $\mathbf{E}_{x|\theta}r(\mathbf{x},\gamma)$ is the expectation when r is evaluated at an argument γ that is not necessarily equal to the true parameter vector θ. The first of these expectations can be interpreted as a function of θ, and the second as a function of γ <u>and</u> θ.

6.1.3. Listed below are some of the properties that are deemed desirable for classical estimators. Classical statistics often proceeds by developing some candidate estimators, and then using some of these properties to choose among the candidates. It is often not possible to achieve all of these properties at the same time, and sometimes they can even be incompatible. Some of the properties are defined relative to a *class* of candidate estimators, a set of possible T(·) that we will denote by $\boldsymbol{T}$. The density of an estimator T(·) will be denoted $\psi(t,\theta_o)$, or when it is necessary to index the estimator, $\psi_T(t,\theta_o)$. Sometimes the parameter vector θ will consist of a subvector α that is of primary interest for the application and a subvector β that is not. Then, α is termed the *primary parameter vector*, β is termed a *nuisance parameter vector,* and the DGP $f(x,\alpha,\beta)$ depends on both the primary and nuisance parameters. In this case, the problem is often to estimate α, dealing with the nuisance parameters as expediently as possible. One approach with fairly wide applicability is to replace β in the DGP by some appropriate function $r(x,\alpha)$, obtaining a *concentrated* DGP $f(x,\alpha,r(x,\alpha))$ that is a function only of the α parameters. Some statistical analysis is needed to determine when this is feasible and can be used as a device to get estimates of α with reasonable statistical properties. A specific choice of $r(x,\alpha)$ that often works is the argument that solves the problem $\max_\beta f(x,\alpha,\beta)$. Keep in mind that choice of parameterization is to some extent under the control of the analyst. Then it may be possible to choose a parameterization that defines α and isolates nuisance parameters in a way that helps in estimation of the primary parameters α.

6.1.4. *Sufficiency*. Suppose there is a one-to-one transformation from the data $\mathbf{x}$ into variables $(\mathbf{y},\mathbf{z})$. Then the DGP density $f(\mathbf{x},\theta)$ can be described in terms of the density of $(\mathbf{y},\mathbf{z})$, which we might denote $g(\mathbf{y},\mathbf{z},\theta)$ and write as the product of the marginal density of $\mathbf{y}$ and the conditional density of $\mathbf{z}$ given $\mathbf{y}$, $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$. The relationship of the density $f(\mathbf{x},\theta)$ and the density $g(\mathbf{y},\mathbf{z},\theta)$ comes from the rules for transforming random variables; see Chapter 3.8. Let $\mathbf{x} = \mathbf{x}(\mathbf{y},\mathbf{z})$ denote the inverse of the one-to-one transformation from $\mathbf{x}$ to $\mathbf{y}$ and $\mathbf{z}$, and let $J(\mathbf{y},\mathbf{z})$ denote the *Jacobian* of this mapping; i.e., the determinant of the array of derivatives of $\mathbf{x}(\mathbf{y},\mathbf{z})$ with respect to its arguments, signed so that it is positive. Then $g(\mathbf{y},\mathbf{z},\theta) = f(\mathbf{x}(\mathbf{y},\mathbf{z}))\cdot J(\mathbf{y},\mathbf{z})$. The Jacobian $J(\mathbf{y},\mathbf{z})$ does not depend on θ, so $g(\mathbf{y},\mathbf{z},\theta)$ factors into a term depending only on $\mathbf{y}$ and θ and a term independent of θ if and only if $f(\mathbf{x}(\mathbf{y},\mathbf{z}))$ factors in the same way.

In general, both the marginal and the conditional densities depend on θ. However, if the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ is independent of θ, $g_2(\mathbf{z}|\mathbf{y},\theta) = g_2(\mathbf{z}|\mathbf{y})$, then the variables $\mathbf{y}$ are said to be *sufficient* for θ. In this case, all of the information in the sample about θ is summarized in $\mathbf{y}$, and once you know $\mathbf{y}$, knowing $\mathbf{z}$ tells you nothing more about θ. (In Chapter 5.4, we demonstrated this by showing that the posterior density for θ, given $\mathbf{y}$ and $\mathbf{z}$, depended only on $\mathbf{y}$, no matter what the prior. Sufficiency of $\mathbf{y}$ is equivalent to a factorization $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ of the density into one term depending only on $\mathbf{y}$ and θ and a second term depending only on $\mathbf{z}$ and $\mathbf{y}$, where the terms g_1 and g_2 need not be densities; i.e., if there is such a factorization, then there is always an additional normalization by a function of $\mathbf{y}$ that makes g_1 and g_2 into densities. This characterization is useful for identifying sufficient statistics. Sufficiency can also be defined with respect to a subvector of primary parameters: if $g(\mathbf{y},\mathbf{z},\alpha,\beta) = g_1(\mathbf{y},\alpha)\cdot g_2(\mathbf{z}|\mathbf{y},\beta)$, then $\mathbf{y}$ is sufficient for α. Another situation that could arise is $g(\mathbf{y},\mathbf{z},\alpha,\beta) = g_1(\mathbf{y},\alpha)\cdot g_2(\mathbf{z}|\mathbf{y},\alpha,\beta)$, so the marginal distribution of $\mathbf{y}$ does not depend on the nuisance parameters, but the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ depends on all the parameters. It may be possible in this case to circumvent estimation of the nuisance parameters by concentrating on $g_1(\mathbf{y},\alpha)$. However, $\mathbf{y}$ is not sufficient for α in this case, as $g_2(\mathbf{z}|\mathbf{y},\alpha,\beta)$ contains additional information on α, unfortunately entangled with the nuisance parameters β.

An implication of sufficiency is that the search for a good estimator can be restricted to estimators $T(\mathbf{y})$ that depend only on sufficient statistics $\mathbf{y}$. In some problems, only the full sample $\mathbf{x}$ is a sufficient statistic, and you obtain no useful restriction from sufficiency. In others there may be many different transformations of $\mathbf{x}$ into $(\mathbf{y},\mathbf{z})$ for which $\mathbf{y}$ is sufficient. Then, among the alternative sufficient statistics, you will want to choose a $\mathbf{y}$ that is a *minimal sufficient statistic*. This will be the case if there is no further one-to-one transformation of $\mathbf{y}$ into variables $(\mathbf{u},\mathbf{v})$ such that $\mathbf{u}$ is sufficient for θ and of lower dimension than $\mathbf{y}$. Minimal sufficient statistics will be most useful when their dimension is low, and/or they isolate nuisance parameters so that the marginal distribution of $\mathbf{y}$ depends only on the primary parameters.

An example shows how sufficiency works. Suppose one has a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ. The DGP density is the product of univariate exponential densities, $f(\mathbf{x},\lambda) = (\lambda\cdot\exp(-\lambda x_1))\cdot...\cdot(\lambda\cdot\exp(-\lambda x_n)) = \lambda^n\cdot\exp(-\lambda(x_1 + ... + x_n))$. Make the one-to-one transformation $y = x_1 + ... + x_n$, $z_1 = x_1,..., z_{n-1} = x_{n-1}$, and note that the inverse transformation implies $x_n = y - z_1 - ... - z_{n-1}$. Substitute the inverse transformation into f to obtain $g(\mathbf{y},\mathbf{z}) = f(\mathbf{x}(y,\mathbf{z})) = \lambda^n\cdot e^{-\lambda y}$. Then, g factors trivially into a marginal gamma density $g_1(y,\lambda)$

$= \lambda^n y^{n-1}\cdot e^{-\lambda y}/(n-1)!$ for y, and a conditional uniform density $g_2(\mathbf{z}|y) = (n-1)!/y^{n-1}$ on the simplex $0 \le z_1 - ... - z_{n-1} \le y$. Then, y is a sufficient statistic for λ, and one need consider only estimators for λ that are functions of the univariate statistic $y = x_1 + ... + x_n$. In this case, y is a minimal sufficient statistic since obviously no further reduction in dimension is possible.

In this exponential example, there are other sufficient statistics that are not minimal. For example, any **y** such as $\mathbf{y} = (x_1 + ... + x_{n-2},x_{n-1},x_n)$ whose components can be transformed to recover the sum of the x's is sufficient. Knowing only that one can restrict the search for an estimator to functions of such a **y** is not as useful as knowing that one only needs to look at the minimal sufficient statistic.

6.1.5. *Ancillarity*. As in the discussion of sufficiency, suppose there is a one-to-one transformation from the data **x** into variables (**y**,**z**). Also suppose that the parameter vector θ is composed of a vector α of primary parameters and a vector β of nuisance parameters. Then the DGP density can be written as the product of the marginal density of **y** and the conditional density of **z** given **y**, $g_1(\mathbf{y},\alpha,\beta)\cdot g_2(\mathbf{z}|\mathbf{y},\alpha,\beta)$. Both g_1 and g_2 depend in general on α and β. However, the data **y** are *ancillary* to α if g_1 does not depend on α and g_2 does not depend on β. In this case, all the information in the data about α is contained in the conditional distribution of **z** given **y**. This implies that the search for an estimator for θ can concentrate solely on the conditional density of **z** given **y**, and that the nuisance parameters drop out of this analysis.

An example where ancillarity is useful arises in data $\mathbf{x} = (x_1,...,x_n)$ where the x_i are independent observations from an exponential density and the sample size n is random with a Poisson density $\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$ for n = 1,2,.... The DGP density is then $\lambda^n\cdot\exp(-\lambda(x_1 + ... + x_n))\cdot\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$. This density factors into the density $\lambda^n y^{n-1}\cdot e^{-\lambda y}$, with $y = x_1 +...+ x_n$, that is now the conditional density of y given n, times a marginal density that is a function of n, y, and γ, but not of λ. Then, the principle of ancillarity says that to make inferences on λ, one should condition on n and not be concerned with the nuisance parameter γ that enters only the marginal density of n.

6.1.6. *Admissibility*. An estimator T(·) for a scalar parameter θ from a class of estimators ***T*** is *admissible relative to* ***T*** if there is no second estimator T′(·) in ***T*** with the property that $\mathbf{E}_{\mathbf{x}|\theta}(T'(\mathbf{x}) - \theta)^2 \le \mathbf{E}_{\mathbf{x}|\theta}(T(\mathbf{x}) - \theta)^2$ for all θ, with inequality strict for at least one $\theta \in \mathbf{\Theta}$. This is the same as the definition of admissibility in statistical decision theory when the cost of a mistake is defined as *mean squared error* (MSE), the expected value of the square of the difference between the estimate and the true value of θ. An inadmissible estimator is undesirable because there is an identified alternative estimator that is more closely clustered around the true parameter value. One limitation of admissibility is that there will often be many admissible estimators, and this criterion does not choose between them. A second limitation is that one might in fact have a cost criterion that is inconsistent with minimizing mean squared error. Suppose, for example, you incur a cost of zero if your estimate is no greater than a distance M from the true value, and a cost of one otherwise. Then, you will prefer the estimator that gives a higher probability of being within distance M, even if it occasionally has large deviations that make its MSE large. The concept of admissibility can be extended to vectors of parameters by saying that an estimator is admissible if it is admissible for each linear combination of the parameter vector.

6.1.7. *Unbiasedness*. An estimator T(·) is *unbiased* for θ if $\mathbf{E}_{x|\theta}T(\mathbf{x}) \equiv \theta$ for all θ; i.e., $\theta \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})f(\mathbf{x},\theta)d\mathbf{x}$. An estimator with this property is "centered" at the true parameter value, and will not systematically be too high or too low. Unbiasedness is an intuitively appealing criterion that is often used in classical statistics to select estimators. However, unbiased estimators are usually inadmissible, a conflict between two reasonable criteria. An example illustrates the issue. Suppose T(·) is an unbiased estimator. Suppose θ* is an arbitrary point in **Θ** and c is a small positive constant, and define T′(·) = (1-c)T(·) + cθ* ; this is called a *Stein shrinkage* estimator. Then

$$\mathbf{E}_{\mathbf{x}|\theta}(T'(\mathbf{x}) - \theta)^2 = \mathbf{E}_{\mathbf{x}|\theta}\,[(1-c)(T(x) - \theta) - c(\theta^* - \theta)]^2 = c^2(\theta^* - \theta)^2 + (1-c)^2\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2,$$

implying that $\partial\mathbf{E}_{\mathbf{x}|\theta}(T'(\mathbf{x}) - \theta)^2/\partial c = 2c(\theta^* - \theta)^2 - 2(1-c)\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2 < 0$ for c sufficiently small. Then, for a problem where $(\theta^* - \theta)^2$ and $\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2$ are bounded for all $\theta \in \mathbf{\Theta}$, one can find c for which T′(·) has lower MSE than T(·), so that T(·) is inadmissible.

6.1.8. *Efficiency*. An estimator T(·) of a scalar parameter is *efficient relative to* an estimator T′(·) if for all θ one has $\mathbf{E}_{x|\theta}(T(\mathbf{x}) - \theta)^2 \le \mathbf{E}_{x|\theta}(T'(\mathbf{x}) - \theta)^2$. The estimator T(·) is efficient relative to a class of estimators ***T*** if it is efficient relative to T′(·) for all T′(·) in ***T***. An efficient estimator provides estimates that are most closely clustered around the true value of θ, by the MSE measure, among all the estimators in ***T***. The limitation of efficiency is that for many problems and classes of estimators ***T***, there will be no efficient estimator, in that one cannot satisfy the required inequality uniformly for all θ. Note that every efficient estimator is admissible, but not every admissible estimator is efficient. If ***T*** contains a unique efficient estimator, then all the other estimators in ***T*** must be inadmissible. The concept of efficiency extends to parameter vectors by requiring that it apply to each linear combination of the parameter vector. The following theorem establishes an important efficiency result for estimators that are functions of sufficient statistics:

Theorem 6.1. (Blackwell) If T′(·) is any estimator of θ from data **x**, and **y** is a sufficient statistic, then there exists an estimator T(·) that is a function solely of the sufficient statistic and that is efficient relative to T′(·). If T′(·) is unbiased, then so is T(·). If an unbiased estimator T(·) is uncorrelated with every unbiased estimator of zero, then T(·) has a smaller variance than any other unbiased estimator, and is the unique efficient estimator in the class of unbiased estimators.

Proof: Suppose there is a scalar parameter. Make a one-to-one transformation of the data **x** into (**y**,**z**), where **y** is the sufficient statistic, and let $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ denote the DGP density. Define T(**y**) = $\mathbf{E}_{z|y}T'(\mathbf{y},\mathbf{z})$. Write T′(**y**,**z**) - θ = T′(**y**,**z**) - T(**y**) + T(**y**) - θ. Then

$$\mathbf{E}(T'(\mathbf{y},\mathbf{z}) - \theta)^2 = \mathbf{E}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}))^2 + \mathbf{E}(T(\mathbf{y}) - \theta)^2 + 2\cdot\mathbf{E}(T(\mathbf{y}) - \theta)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}))\,.$$

But the last term satisfies

$$2\cdot\mathbf{E}(T(\mathbf{y}) - \theta)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 2\cdot\mathbf{E}_y(T(\mathbf{y}) - \theta)\cdot\mathbf{E}_{z|y}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 0\,.$$

Therefore, $\mathbf{E}(\mathrm{T}'(\mathbf{y},\mathbf{z}) - \theta)^2 \geq \mathbf{E}(\mathrm{T}(\mathbf{y}) - \theta)^2$. If $\mathrm{T}'(\mathbf{y},\mathbf{z})$ is unbiased, then $\mathbf{E}\mathrm{T}(\mathbf{y}) = \mathbf{E}_y\mathbf{E}_{z|y}\mathrm{T}'(\mathbf{y},\mathbf{z}) = \theta$, and $\mathrm{T}(\cdot)$ is also unbiased. Finally, suppose $\mathrm{T}(\cdot)$ is uncorrelated with any estimator $\mathrm{U}(\cdot)$ that is an unbiased estimator of zero, i.e., $\mathbf{E}\mathrm{U}(\mathbf{y},\mathbf{z}) = 0$ implies $\mathbf{E}\mathrm{U}(\mathbf{y},\mathbf{z})\cdot(\mathrm{T}(\mathbf{y}) - \theta) = 0$. Then, any unbiased $\mathrm{T}'(\mathbf{y},\mathbf{z})$ has $\mathrm{U}(\mathbf{y},\mathbf{z})\cdot= \mathrm{T}'(\mathbf{y},\mathbf{z}) - \mathrm{T}(\mathbf{y})$ an unbiased estimator of zero, implying

$$\begin{aligned}\mathbf{E}(\mathrm{T}'(\mathbf{x}) - \theta)^2 &= \mathbf{E}(\mathrm{T}'(\mathbf{x}) - \mathrm{T}(\mathbf{x}) + \mathrm{T}(\mathbf{x}) - \theta)^2 = \mathbf{E}(\mathrm{T}'(\mathbf{x}) - \mathrm{T}(\mathbf{x}))^2 + \mathbf{E}(\mathrm{T}(\mathbf{x}) - \theta)^2 + 2\cdot\mathbf{E}\mathrm{T}(\mathbf{x})\cdot(\mathrm{T}'(\mathbf{x}) - \mathrm{T}(\mathbf{x}))\\ &= \mathbf{E}(\mathrm{T}'(\mathbf{x}) - \mathrm{T}(\mathbf{x}))^2 + \mathbf{E}(\mathrm{T}(\mathbf{x}) - \theta)^2 > \mathbf{E}(\mathrm{T}(\mathbf{x}) - \theta)^2.\end{aligned}$$

The theorem also holds for vectors of parameters, and can be established by applying the arguments above to each linear combination of the parameter vector. □

6.1.9 (*MVUE*) If ***T*** is a class of unbiased estimators of a scalar parameter, so that $\mathbf{E}_{x|\theta}\mathrm{T}'(\mathbf{x}) \equiv \theta$ for every estimator $\mathrm{T}'(\cdot)$ in this class, then an estimator is efficient in this class if its variance is no larger than the variance of any other estimator in the class, and is termed a *minimum variance unbiased estimator* (MVUE). There are many problems for which no MVUE estimator exists. We next give a lower bound on the variance of an unbiased estimator. If a candidate satisfies this bound, then we can be sure that it is MVUE. However, the converse is not true: There may be a MVUE, its variance may still be larger than this lower bound; i.e., the lower bound may be unobtainable. Once again, the MVUE concept can be extended to parameter vectors by requiring that it apply to each linear combination of parameters.

Theorem 6.2. (Cramer-Rao Bound) Suppose a simple random sample $\mathbf{x} = (x_1,...,x_N)$ with $f(x,\theta)$ the density of an observation x. Assume that log $f(x,\theta)$ is twice continuously differentiable in a scalar parameter θ, and that this function and its derivatives are bounded in magnitude by a function that is independent of θ and has a finite integral in x. Define the *Fisher information* in an observation, $\mathbf{J} = \mathbf{E}_{x|\theta}\,[\nabla_\theta \log f(x,\theta)][\nabla_\theta \log f(x,\theta)]'$. Suppose $\mathrm{T}(\mathbf{x})$ has $\mathbf{E}_{x|\theta}\mathrm{T}(\mathbf{x}) \equiv \theta + \mu(\theta)$. Then $\mu(\theta)$ is the *bias* of the estimator. Suppose that $\mu(\theta)$ is differentiable. Then, the variance of $\mathrm{T}(\mathbf{x})$ satisfies

$$\mathbf{V}_{x|\theta}(\mathrm{T}(\mathbf{x})) \geq (\mathbf{I} + \nabla_\theta\mu(\theta))(\mathrm{n}\mathbf{J})^{-1}(\mathbf{I} + \nabla_\theta\mu(\theta))'.$$

If the estimator is unbiased, so $\mu(\theta) \equiv 0$, this bound reduces to

$$\mathbf{V}_{x|\theta}(\mathrm{T}(\mathbf{X})) \geq (\mathbf{nJ})^{-1},$$

so that *the variance of an unbiased estimator is at least as large as the inverse of the Fisher information in the sample*. This result continues to hold when θ is a vector, with $\mathbf{V}_{x|\theta}(\mathrm{T}(\mathbf{x}))$ a covariance matrix and "≥" interpreted to mean than the matrix difference is positive semidefinite.

Proof: Assume θ is a scalar. Let $\mathrm{L}(\mathbf{x},\theta) = \sum_{i=1}^{n} \log f(x_i,\theta)$, so that the DGP density is $f(\mathbf{x},\theta) = e^{\mathrm{L}(x,\theta)}$. By construction,

$$1 \equiv \int_{-\infty}^{+\infty} e^{L(x,\theta)}d\mathbf{x} \quad \text{and} \quad \theta + \mu(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot e^{L(x,\theta)}d\mathbf{x}.$$

The conditions of the Lebesgue dominated convergence theorem are met, allowing differentiation under the integral sign. Then, differentiate each integral with respect to θ to get

$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x} \quad \text{and} \quad 1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x}.$$

Combine these to get an expression for the covariance of T and $\nabla_\theta L$,

$$1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} [T(\mathbf{x}) - \theta]\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x}.$$

Apply the *Cauchy-Schwartz inequality*; see 3.5.9. In this case, the inequality can be written

$$(1 + \mu'(\theta))^2 = \left(\int_{-\infty}^{+\infty}[T(\boldsymbol{x}) - \theta]\cdot\nabla_\theta L(\boldsymbol{x},\theta)\cdot e^{L(x,\theta)}d\boldsymbol{x}\right)^2 \le [\mathbf{E}_{x|\theta}(T(\mathbf{x}) - \theta)^2]\cdot]\mathbf{E}_{x|\theta}[\nabla_\theta L(\mathbf{x},\theta)]^2].$$

Dividing both sides by the Fisher information in the sample, which is simply the variance of the sample score, $\mathbf{E}_{x|\theta}\,[\nabla_\theta L(\mathbf{x},\theta)]^2$, gives the bound.

When θ is k×1, one has $\theta + \mu(\theta) = \int_{-\infty}^{+\infty}T(\boldsymbol{x})\,e^{L(x,\theta)}d\boldsymbol{x}$. Differentiating with respect to θ gives

$\mathbf{I} + \nabla_\theta\mu(\theta) = \int_{-\infty}^{+\infty}T(\boldsymbol{x})\cdot\nabla_\theta L(\boldsymbol{x},\theta)\cdot e^{L(x,\theta)}d\boldsymbol{x} = \int_{-\infty}^{+\infty}(T(\boldsymbol{x}) - \theta - \mu(\theta))\cdot\nabla_\theta L(\boldsymbol{x},\theta)\cdot e^{L(x,\theta)}d\boldsymbol{x}$. The vector $(T(x) - \theta - \mu(\theta))'$, $\nabla_\theta L(\mathbf{x},\theta))$ has a positive semidefinite covariance matrix that can be written in partitioned form as $\begin{bmatrix} \boldsymbol{V}_{x|\theta}(T(x)) & [\boldsymbol{I}+\nabla_\theta\mu(\theta)] \\ [\boldsymbol{I}+\nabla_\theta\mu(\theta)]' & n\boldsymbol{J} \end{bmatrix}$. If one premultiplies this matrix by **W**, and postmultiplies by $\mathbf{W}'$, the result is positive semidefinite. Taking $W = \begin{bmatrix}\boldsymbol{I} & -[\boldsymbol{I}+\nabla_\theta\mu(\theta)](n\boldsymbol{J})^{-1}\end{bmatrix}$ gives the Cramer-Rao bound for the vector case. □

6.1.10. *Invariance*. In some conditions, one would expect that a change in a problem should not alter an estimate of a parameter, or should alter it in a specific way. Generically, these are called invariance properties of an estimator. For example, when estimating a parameter from data obtained by a simple random sample, the estimate should not depend on the indexing of the observations in the sample; i.e., $T(x_1,...,x_n)$ should be *invariant under permutations of the observations*. A second example is *invariance with sample scale*: if $T_n(x_1,...,x_n)$ denotes the estimator for a sample of size n, and the observations all equal a constant c, then the estimator should not change with sample size, or $T_n(c,...,c) = T_1(c)$. A sample mean, for example, has these two invariance properties.

Sometimes a parameter enters a DGP in such a way that there is a simple relationship between shifts in the parameter and the shifts one would expect to observe in the data. For example, suppose the density of an observation is of the form $f(x_i|\theta) \equiv h(x_i-\theta)$; in this case, θ is called a *location parameter*. If the true value of θ shifts up by an amount Δ, one would expect observations on average to shift up by the same amount Δ. If $T_n(x_1,...,x_n)$ is an estimator of θ_o in this problem, a reasonable property to impose on $T_n(\cdot)$ is that $T_n(x_1+\Delta,...,x_n+\Delta) = T_n(x_1,...,x_n) + \Delta$. In this case, $T_n(\cdot)$ is termed *location invariant*. For this problem, one can restrict attention to estimators with this invariance property.

Another example is *scale invariance*. Suppose the density of an observation has the form $f(x_i|\theta) \equiv \theta\cdot h(\theta x_i)$. Then θ is called a *scale parameter*. If θ is reduced by a proportion λ, one would expect observations on average to be scaled up by λ. The corresponding invariance property on an estimator $T_n(\cdot)$ is that $T_n(\lambda\cdot x_1,...,\lambda\cdot x_n) = T_n(x_1,...,x_n)/\lambda$.

To illustrate the use of invariance conditions, consider the example of a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ, with the DGP density $f(\mathbf{x},\lambda) = \lambda^n\exp(-\lambda(x_1 + ... + x_n))$. Then $y = x_1 + ... + x_n$ is sufficient and we need consider only estimators $T_n(y)$. Invariance with respect to scale implies $T_n(y) = T_n(1)/y$. Invariance with sample scale requires that if $x_1 = ... = x_n = 1$, so that $y = n$, then $T_n(n) = T_1(1)$. Combining these conditions, $T_1(1) = T_n(1)/n$ and hence $T_n(y) = T_1(1)/\bar{y}$, so that an estimator that is a function of the sufficient statistic and has these invariance properties must be inversely proportional to the sample mean.

6.1.11. The next group of properties refer to the limiting behavior of estimators in a sequence of larger and larger samples, and are sometimes called *asymptotic properties*. The rationale for employing these properties is that when one is working with a large sample, then properties that hold in the limit will also hold, approximately, for this sample. The reason for considering such properties at all, rather than concentrating on the sample you actually have, is that one can use these approximate properties to choose among estimators in situations where the exact finite sample property cannot be imposed or is analytically intractable to work out.

Application of asymptotic properties raises several conceptual and technical issues. The first question is what it would mean to increase sample size indefinitely, and whether various methods that might be used to define this limit correspond to approximations that are likely to be relevant to a specific problem. There is no ambiguity when one is drawing simple random samples from an infinite population. However, if one samples from a finite population, a finite sequence of samples of increasing size will terminate in a complete census of the population. While one could imagine sampling with replacement and drawing samples that are larger than the population, it is not obvious why estimators that have some reasonable properties in this limit are necessarily appropriate for the finite population. Put another way, it is not obvious that this limit provides a good approximation to the finite sample.

The issue of the appropriate asymptotic limit is particularly acute for time series. One can imagine extending observations indefinitely through time. This may provide approximations that are appropriate in some situations for some purposes, but not for others. For example, if one is trying to estimate the timing of a particular event, a local feature of the time series, it is questionable that extending the time series indefinitely into the past and future leads to a good approximation to the statistical properties of the estimator of the timing of an event. Other ways of thinking of increasing sample sizes for time series, such as sampling from more and more "parallel" universes, or sampling at shorter and shorter intervals, have their own idiosyncrasies that make them questionable as useful approximations.

A second major issue is how the sequence of estimators associated with various sample sizes is defined. A conceptualization introduced in Chapter 5 defines an estimator to be a functional of the empirical CDF of the data, $T(F_n)$. Then, it is natural to think of $T(F(\cdot,\theta))$ as the limit of this sequence of estimators, and the Glivenko-Cantelli theorem stated in Chapter 5.1 establishes an approximation property that the estimator $T(F_n)$ converges almost surely to $T(F(\cdot,\theta))$ if the latter exists. This suggests that defining estimators as “continuous” functions of the CDF leads to a situation in which the asymptotic limit will have reasonable approximation properties in large samples. However, tt is important to avoid reliance on asymptotic arguments when it is clear that the asymptotic approximation is irrelevant to the behavior of the estimator in the range of sample sizes actually encountered. Consider an estimation procedure which says "Ignore the data and estimate θ_o to be zero in all samples of size less than 10 billion, and for larger samples employ some computationally complex but statistically sound estimator". This procedure may technically have good asymptotic properties, but this approximation obviously tells you nothing about the behavior of the estimator in economic sample sizes of a few thousand observations.

6.1.12. *Consistency*. A sequence of estimators $T_n(\mathbf{x}) = T_n(x_1,...,x_n)$ for samples of size n are *consistent* for θ_o if the probability that they are more than a distance $\varepsilon > 0$ from θ_o goes to zero as n increases; i.e., $\lim_{n\to\infty} P(|T_n(x_1,...,x_n) - \theta_o| > \varepsilon) = 0$. In the terminology of Chapter 4, this is *weak convergence* or *convergence in probability*, written $T_n(x_1,...,x_n) \to_p \theta_o$. One can also talk about *strong consistency*, which holds when $\lim_{n\to\infty} P(\sup_{m\geq n}|T_m(x_1,...,x_{n'}) - \theta_o| > \varepsilon) = 0$, and corresponds to almost sure convergence, $T_n(x_1,...,x_n) \to_{as} \theta_o$.

6.1.13. *Asymptotic Normality*. A sequence of estimators $T_n(\cdot)$ for samples of size n are *consistent asymptotically normal* (CAN) for θ if there exists a sequence r_n of scaling constants such that $r_n \to +\infty$ and $r_n\cdot(T_n(\mathbf{x}_n) - \theta)$ converges in distribution to a normally distributed random variable with some mean $\mu = \mu(\theta)$ and variance $\sigma^2 = \sigma(\theta)^2$. If $\Psi_n(t)$ is the CDF of $T_n(\mathbf{x}_n)$, then $Q_n = r_n\cdot(T_n(\mathbf{x}_n) - \theta)$ has the CDF $P(Q_n \leq q) = \Psi_n(\theta + q/r_n)$. From Chapter 4, one will have convergence in distribution to a normal, $r_n(T_n(\mathbf{x}_n) - \theta) \to_d Z$ with $Z \sim N(\mu,\sigma^2)$, if and only if for each q, the CDF of Q_n satisfies $\lim_{n\to\infty} |\Psi_n(\theta + q/r_n) - \Phi((q-\mu)/\sigma)| = 0$. This is the conventional definition of convergence in distribution, with the continuity of the normal CDF Φ permitting us to state the condition without excepting jump points in the limit distribution. In this setup, $\Psi_n(t)$ is converging in distribution to $\mathbf{1}(t\geq\theta)$, the CDF of the constant random variable equal to θ. However, r_n is blowing up at just the right rate so that $\Psi_n(\theta + q/r_n)$ has a non-degenerate asymptotic distribution, whose shape is determined by the local shape of Ψ_n in shrinking neighborhoods of θ. The mean μ is termed the *asymptotic bias*, and σ^2 is termed the *asymptotic variance*. If $\mu = 0$, the estimator is said to be *asymptotically unbiased*. An unbiased estimator will be asymptotically unbiased, but the reverse is not necessarily true. Often, when a sequence of estimators is said to be asymptotically normal, asymptotic unbiasedness is taken to be part of the definition unless stated explicitly to the contrary.

The scaling term r_n can be taken to be $n^{1/2}$ in almost all finite-parameter problems, and unless it is stated otherwise, you can assume that this is the scaling that is being used. When it is important to make this distinction clear, one can speak of *Root-n consistent asymptotically normal* (RCAN) sequences of estimators.

Convergence in distribution to a normal is a condition that holds pointwise for each true parameter θ. One could strengthen the property by requiring that this convergence be uniform in θ; i.e., by requiring for each $\varepsilon > 0$ and q that there be a sample size $n(\varepsilon,q)$ beyond which $\sup_\theta |\Psi(\theta_o + q/r_n) - \Phi((q-\mu(\theta_o))/\sigma(\theta_o))| < \varepsilon$. If this form of convergence holds, and in addition $\mu(\theta)$ and $\sigma(\theta)^2$ are continuous functions of θ, then the estimator is said to be *consistent uniformly asymptotically normal* (CUAN).

6.1.14. *Asymptotic Efficiency*. Consider a family ***T*** of sequences of estimators $T_n(\cdot)$ that are CUAN for a parameter θ and have asymptotic bias $\mu(\theta) \equiv 0$. An estimator $T^*(\cdot)$ is *asymptotically efficient* relative to class ***T*** if its asymptotic variance is no larger than that of any other member of the family. The reason for restricting attention to the CUAN class is that in the absence of uniformity, there exist "super-efficient" estimators, constructed in the following way: Suppose $T_n(\cdot)$ is an asymptotically efficient estimator in the CUAN class. For an arbitrary θ^*, define $T_n^*(\cdot)$ to equal $T_n(\cdot)$ if $n^{1/2}|T_n(\mathbf{x}) - \theta^*| \geq 1$, and equal to θ^* otherwise. This estimator will have the same asymptotic variance as $T_n(\cdot)$ for fixed $\theta \neq \theta^*$, and an asymptotic variance of zero for $\theta = \theta^*$. Thus, it is more efficient. On the other hand, it has a nasty asymptotic bias for parameter vectors that are "local" to θ^*, so that it is not CUAN, and would be an unattractive estimator to use in practice. Once these non-uniform superefficient estimators are excluded by restricting attention to the CUAN class, one has the result that under reasonable regularity conditions, an asymptotic version of the Cramer-Rao bound for unbiased estimators holds for CUAN estimators.

6.1.15. *Asymptotic sufficiency*. In some problems, sufficiency does not provide a useful reduction of dimension in finite samples, but a weaker "asymptotic" form of sufficiency will provide useful restrictions. This could arise if the DGP density can be written $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$ for a low-dimensional statistic $\mathbf{y}$, but both g_1 and g_2 depend on θ so $\mathbf{y}$ is not sufficient. However, $g_2(\mathbf{z}|\mathbf{y},\theta)$ may converge in distribution to a density that does not depend on θ. Then, there is a large sample rationale for concentrating on estimators that depend only on $\mathbf{y}$.

6.2. GENERAL ESTIMATION CRITERIA

6.2.1. It is useful to have some general methods of generating estimators that as a consequence of their construction will have some desirable statistical properties. Such estimators may prove adequate in themselves, or may form a starting point for refinements that improve statistical properties. We introduce several such methods:

6.2.2. *Analogy Estimators*. Suppose one is interested in a feature of a target population that can be described as a functional of its CDF $F(\cdot)$, such as its mean, median, or variance, and write this feature as $\theta = \mu(F)$. An analogy estimator exploits the similarity of a population and of a simple random sample drawn from this population, and forms the estimator $T(\mathbf{x}) = \mu(F_n)$, where μ is the functional that produces the target population feature and F_n is the empirical distribution function. For example, a sample mean will be an analogy estimator for a population mean.

6.2.3. *Moment Estimators*. Population moments will depend on the parameter index in the underlying DGP. This is true for ordinary moments such as means, variances, and covariances, as well as more complicated moments involving data transformations, such as quantiles. Let m(x) denote a function of an observation and $\mathbf{E}_{x|\theta}m(x) = \gamma(\theta)$ denote the population moment formed by taking the expectation of m(x). In a sample $\mathbf{x} = (x_1,...,x_n)$, the idea of a moments estimator is to form a sample moment $n^{-1}\sum_{i=1}^{n} m(x_i) \equiv \mathbf{E}_n m(x)$, and then to use the analogy of the population and sample moments to form the approximation $\mathbf{E}_n m(x) \approx \mathbf{E}_{x|\theta} = \gamma(\theta)$. The sample average of a function m(x) of an observation can also be interpreted as its expectation with respect to the empirical distribution of the sample; we use the notation $\mathbf{E}_n m(x)$ to denote this empirical expectation. The moment estimator $T(\mathbf{x})$ solves $\mathbf{E}_n m(x) = \gamma(T(\mathbf{x}))$. When the number of moment conditions equals the number of parameters, an exact solution is normally obtainable, and $T(\mathbf{x})$ is termed a *classical method of moments estimator*. When the number of moment conditions exceeds the number of parameters, it is not possible in general to find $T(\mathbf{x})$ that sets them all to zero at once. In this case, one may form a number of linear combinations of the moments equal to the number of parameters to be estimated, and find $T(\mathbf{x})$ that sets these linear combinations to zero. The linear combinations in turn may be derived starting from some metric that provides a measure of the distance of the moments from zero, with $T(\mathbf{x})$ interpreted as a minimand of this metric. This is called *generalized method of moments* estimation.

6.2.4. *Maximum likelihood estimators*. Consider the DGP density $f(\mathbf{x},\theta)$ for a given sample as a function of θ. The maximum likelihood estimator of the unknown true value θ is the statistic $T(\mathbf{x})$ that maximizes $f(\mathbf{x},\theta)$. The intuition behind this estimator is that if we guess a value for θ that is far away from the true θ_o, then the probability law for this θ would be very unlikely to produce the data that are actually observed, whereas if we guess a value for θ that is near the true θ_o, then the probability law for this θ would be likely to produce the observed data. Then, the $T(\mathbf{x})$ which maximized this likelihood, as measured by the probability law itself, should be close to the true θ. The maximum likelihood estimator plays a central role in classical statistics, and can be motivated solely in terms of its desirable classical statistical properties in large samples.

When the data are a sample of n independent observations, each with density $f(x,\theta)$, then the likelihood of the sample is $f(\mathbf{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$. It is often convenient to work with the logarithm of the density, $l(x,\theta) \equiv \text{Log } f(x,\theta)$. Then, the *Log Likelihood* of the sample is $L(\mathbf{x},\theta) \equiv \text{Log } f(\mathbf{x},\theta) =$

$\sum_{i=1}^{n}$ $l(x_i,\theta)$. The *maximum likelihood estimator* is the function t = T(**x**) of the data that when substituted for θ maximizes f(**x**,θ), or equivalently L(**x**,θ).

The gradient of the log likelihood of an observation with respect to θ is denoted $s(\mathbf{x},\theta) \equiv \nabla_\theta l(x,\theta)$, and termed the *score*. The maximum likelihood estimator is a zero of the sample expectation of the score, $\mathbf{E}_n s(\mathbf{x},T(\mathbf{x}))$. Then, the maximum likelihood estimator is a special case of a moments estimator.

Maximum likelihood estimators will under quite general regularity conditions be consistent and asymptotically normal. Under uniformity conditions that rule out some odd non-uniform "super-efficient" alternatives, they are also asymptotically efficient. They often have good finite-sample properties, or can be easily modified so that they do. However, their finite-sample properties have to be determined on a case-by-case basis. In multiple parameter problems, particularly when there are primary parameters α and nuisance parameters β, the maximum likelihood principle can sometimes be used to handle the nuisance parameters. Specifically, maximum likelihood estimation for all parameters will find the parameter values that solve $\max_{\alpha,\beta} L(\mathbf{x},\alpha,\beta)$. But one could get the same solution by first maximizing in the nuisance parameters β, obtaining a solution β = r(x,α), and substituting this back into the likelihood function to obtain L(x,α,r(x,α)). This is called the concentrated likelihood function, and it can now be maximized in α alone. The reason this can be an advantage is that one may be able to obtain r(x,α) "formally" without having to compute it.

6.3. ESTIMATION IN NORMALLY DISTRIBUTED POPULATIONS

6.3.1. Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which observations are normally distributed with mean μ and variance σ^2. Let $\varphi(v) = (2\pi)^{-1/2}\exp(-v^2/2)$ denote the standard normal density. Then the density of observation x_i is $\varphi((x_i - \mu)/\sigma)/\sigma$. The log likelihood of the sample is

$$L(\mathbf{x},\mu,\sigma^2) = -\ \frac{n}{2}\ \cdot\text{Log}(2\pi) -\ \frac{n}{2}\ \cdot\text{Log}\ \sigma^2 -\ \frac{1}{2}\ \sum_{i=1}^{n}\ (x_i - \mu)^2/\sigma^2.$$

We will find estimates μ_e and σ_e^2 for the parameters μ and σ^2 using the maximum likelihood method, and establish some of the statistical properties of these estimators.

6.3.2. The first-order-conditions for maximizing $L(\mathbf{x},\mu,\sigma^2)$ in μ and σ^2 are

$$0 = \sum_{i=1}^{n}\ (x_i-\mu)/\sigma^2 \Longrightarrow \mu_e = \bar{x} \equiv\ n^{-1}\sum_{i=1}^{n}\ x_i,$$

$$0 = -n/2\sigma^2 + \sum_{i=1}^{n}\ (x_i-\mu)^2/2\sigma^4 \Longrightarrow \sigma_e^{\ 2} =\ n^{-1}\sum_{i=1}^{n}\ (x_i-\bar{x})^2.$$

The maximum likelihood estimator of μ is then the sample mean, and the maximum likelihood estimator of σ^2 is the sample variance. Define $s^2 = \sigma_e{}^2{\cdot}n/(n\text{-}1) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i\text{-}\bar{x})^2$, the sample variance with a sample size correction. The following result summarizes the properties of these estimators:

Theorem 6.3. If **x** = $(x_1,...,x_n)$ is a simple random sample from a population in which observations are normally distributed with mean μ and variance σ^2, then

(1) $(\bar{x},s^2)$ are joint minimal sufficient statistics for (μ,σ^2)**.**
(2) $\bar{x}$ is an unbiased estimator for μ, and s^2 an unbiased estimator for σ^2**.**
(3) $\bar{x}$ is a Minimum Variance Unbiased Estimator (MVUE) for μ**;** s^2 is MVUE for σ^2**.**
(4) $\bar{x}$ is Normally distributed with mean μ and variance σ^2/n.
(5) $(n\text{-}1)s^2/\sigma^2$ has a Chi-square distribution with n-1 degrees of freedom.
(6) $\bar{x}$ and s^2 are statistically independent.
(7) $n^{1/2}(\bar{x} - \mu)/s$ has a Student's-T distribution with n-1 degrees of freedom.
(8) $(\bar{x} - \mu)^2/s^2$ has an F-distribution with 1 and n-1 degrees of freedom.

Proof: (1) Factor the log likelihood function as

$$L(\mathbf{x},\mu,\sigma^2) = -\ \frac{n}{2}\cdot\mathrm{Log}(2\pi) - \frac{n}{2}\cdot\mathrm{Log}\ \sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2/\sigma^2$$

$$= -\ \frac{n}{2}\cdot\mathrm{Log}(2\pi) - \frac{n}{2}\cdot\mathrm{Log}\ \sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n} (x_i - \bar{x})^2/\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n} (\bar{x}\text{-}\mu)^2/\sigma^2$$

$$= -\ \frac{n}{2}\cdot\mathrm{Log}(2\pi) - \frac{n}{2}\cdot\mathrm{Log}\ \sigma^2 - \frac{1}{2}\cdot\frac{(n-1)s^2}{\sigma^2} - \frac{n}{2}(\bar{x}\text{-}\mu)^2/\sigma^2\ .$$

This implies that $\bar{x}$ and s^2 are jointly sufficient for μ and σ^2. Because the dimension of $(\bar{x},s^2)$ is the same as the dimension of (μ,σ^2), they are obviously minimal sufficient statistics.

(2) The expectation of $\bar{x}$ is $\mathbf{E}\,\bar{x} = n^{-1}\sum_{i=1}^{n} \mathbf{E}x_i = \mu$, since the expectation of each observation is μ. Hence $\bar{x}$ is unbiased. To establish the expectation of s^2, first form the n×n matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$, where $\mathbf{I}_n$ is the n×n identity matrix and $\mathbf{1}_n$ is a n×1 vector of ones. The matrix **M** is idempotent (check) and its trace satisfies $\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}(\mathbf{1}_n\mathbf{1}_n'/n) = n - \mathrm{tr}(\mathbf{1}_n'\mathbf{1}_n/n) = n - 1$. The result then follows from Theorem 3.11 (viii). For a direct demonstration, let $Z' = (x_1\text{-}\mu,...,x_n\text{-}\mu)$ denote the vector of deviations of observations from the population mean. This vector contains independent identically distributed normal random variables with mean zero and variance σ^2, so that $\mathbf{E}ZZ' = \sigma^2\mathbf{I}_n$. Further, $Z'\mathbf{M} = (x_1 - \bar{x},...,x_n - \bar{x})$ and $s^2 = Z'\mathbf{M}{\cdot}\mathbf{M}Z/(n\text{-}1) = Z'\mathbf{M}Z/(n\text{-}1)$. Therefore, $\mathbf{E}s^2$

$= \mathbf{E}(Z'\mathbf{M}Z)/(n\text{-}1) = \mathbf{E}\ \mathrm{tr}(Z'\mathbf{M}Z)/(n\text{-}1) = \mathbf{E}\ \mathrm{tr}(\mathbf{M}ZZ')/(n\text{-}1) = \mathrm{tr}(\mathbf{M}\cdot\mathbf{E}(ZZ'))/(n\text{-}1) = \sigma^2\cdot\mathrm{tr}(\mathbf{M})/(n\text{-}1) = \sigma^2$. Hence, s^2 is unbiased.

(3) The MVUE property of $\bar{x}$ and s^2 is most easily proved by application of the Blackwell theorem. We already know that these estimators are unbiased. Any other unbiased estimator of μ then has the property that the difference of this estimator and $\bar{x}$, which we will denote by $h(\mathbf{x})$, must satisfy $\mathbf{E}h(\mathbf{x}) \equiv 0$. Alternately, $h(\mathbf{x})$ could be the difference of s^2 and any other unbiased estimator of σ^2. We have condition (a) that $0 \equiv \mathbf{E}h(\mathbf{x}) \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\exp(L(\mathbf{x},\mu,\sigma^2))d\mathbf{x}$. Striking terms that can be taken outside the integral gives condition (b) that $0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\exp(- \sum_{i=1}^{n} (x_i- \mu)^2/2\sigma^2)d\mathbf{x}$.

Differentiate (b) with respect to σ^2 and strike out terms that can be taken outside the integral to obtain condition (c) that $0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot \sum_{i=1}^{n} (x_i- \mu)^2\cdot\exp(- \sum_{i=1}^{n} (x_i- \mu)^2/2\sigma^2)d\mathbf{x}$. Differentiate (b) with respect to μ, again strike out terms that can be taken outside, and use (b) to obtain condition (d) that $0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot \sum_{i=1}^{n} x_i\cdot\exp(- \sum_{i=1}^{n} (x_i-\mu)^2/2\sigma^2)d\mathbf{x}$, which implies that $\mathbf{E}h(\mathbf{x})\cdot\bar{x} \equiv 0$. Differentiate (d) with respect to μ, once again strike out terms and eliminate terms that are zero by property (b) to obtain the condition (e) that $0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot \sum_{i=1}^{n} x_i\cdot\exp(- \sum_{i=1}^{n} (x_i- \mu)^2/2\sigma^2)d\mathbf{x}$, which implies that $\mathbf{E}h(\mathbf{x})\cdot\bar{x}^2 = 0$, and hence by (b) and (d), $\mathbf{E}h(\mathbf{x})\cdot(\bar{x}\text{-}\mu)^2 = 0$. But (c) can be written $0 = \mathbf{E}h(\mathbf{x})\cdot[((n\text{-}1)s^2 + n\ (\bar{x}\text{-}\mu)^2]$, and this combined with the last result implies $\mathbf{E}h(\mathbf{x})\cdot s^2 = 0$. Then, the estimators $\bar{x}$ and s^2 are uncorrelated with any unbiased estimator of zero. The Blackwell theorem then establishes that they are the unique minimum variance estimators among all unbiased estimators.

(4) Next consider the distribution of $\bar{x}$. We use the fact that linear transformations of multivariate normal random vectors are again multivariate normal: If $Z \sim N(\boldsymbol{\mu},\boldsymbol{\Omega})$ and $W = \mathbf{C}Z$, then $W \sim N(\mathbf{C}\boldsymbol{\mu},\mathbf{C}\boldsymbol{\Omega}\mathbf{C}')$. This result holds even if Z and W are of different dimensions, or $\mathbf{C}$ is of less than full rank. (If the rank of $\mathbf{C}\boldsymbol{\Omega}\mathbf{C}'$ is less than full, then the random variable has all its density concentrated on a subspace.) Now $\bar{x} = \mathbf{C}\mathbf{x}$ when $\mathbf{C} = (1/n,...,1/n)$. We have $\mathbf{x}$ multivariate normal with mean $\mathbf{1}_n\mu$ and covariance matrix $\sigma^2\mathbf{I}_n$, where $\mathbf{1}_n$ is a n×1 vector of ones and $\mathbf{I}_n$ is the n×n identity matrix. Therefore, $\bar{x} \sim N(\mu C\mathbf{1}_n,\sigma^2CC') = N(\mu,\sigma^2/n)$.

(5) Next consider the distribution of s^2. Consider the quadratic form $(\mathbf{x}/\sigma)'\mathbf{M}(\mathbf{x}/\sigma)$, where $\mathbf{M}$ is the idempotent matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ from (2). The vector $(\mathbf{x}/\sigma)$ is independent standard normal, so that Theorem 3.11(iii) gives the result.

(6) The matrices $\mathbf{C} = (1/n,...,1/n) = \mathbf{1}_n'$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ have $\mathbf{CM} = \mathbf{0}$. Then Theorem 3.11(vii) gives the result that $\mathbf{C}(\mathbf{x}/\sigma) = \bar{x}/\sigma$ and $(\mathbf{x}/\sigma)'\mathbf{M}(\mathbf{x}/\sigma) = (n\text{-}1)s^2/\sigma^2$ are independent.

For (7), Use Theorem 3.9(ii), and for (8), use Theorem 3.9(iii). □

4. LARGE SAMPLE PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

This section provides a brief and informal introduction to the statistical properties of maximum likelihood estimators and similar estimation methods in large samples. Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which the density of an observation is $f(x,\theta_o)$. The DGP density or likelihood of the sample is then $f(\mathbf{x},\theta) = f(x_1,\theta)\cdot...\cdot f(x_n,\theta)$, with θ_o the true value of θ. The log likelihood of an observation is $l(x,\theta) = \log f(x,\theta_o)$, and the log likelihood of the sample is $L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} l(x_n,\theta)$. The maximum likelihood estimator $T_n(\mathbf{x})$ is a value of θ which maximizes $L_n(\mathbf{x},\theta)$. The first-order condition for this maximum is that the *sample score*,

$$\nabla_\theta L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} \nabla_\theta l(x_i,\theta),$$

equal zero at $\theta = T_n(\mathbf{x})$. The second order condition is that the *sample hessian*$\nabla_{\theta\theta}L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} \nabla_{\theta\theta}l(x_i,\theta)$, be negative at $\theta = T(\mathbf{x})$. When the parameter θ is more than one-dimensional, the second-order condition is that the sample hessian is a negative definite matrix.

Under very mild regularity conditions, the expectation of the score of an observation is zero at the true parameter vector. Start from the identity $\int_{-\infty}^{+\infty} \exp(l(x,\theta))\cdot dx \equiv 1$ and differentiate with respect to θ under the integral sign to obtain the condition $\int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)\cdot\exp(l(x,\theta))\cdot dx \equiv 0$. (Regularity conditions are needed to assure that one can indeed differentiate under the integral; this will be supplied by assuming a dominance condition so that the Lebesgue dominated convergence theorem can be applied; see Theorem 3.1 and the discussion following.) Then, at the true parameter θ,one has $\mathbf{E}_{x|\theta}\nabla_\theta l(x,\theta) = 0$, the condition that the *population score* is zero when $\theta = \theta_o$. Another regularity condition requires that $E_{\mathbf{x}|\theta_o} \nabla_\theta l(x,\theta) = 0$ <u>only</u> if $\theta = \theta_o$; this has the interpretation of an *identification condition*. The maximum likelihood estimator can be interpreted as an analogy estimator that chooses $T_n(\mathbf{x})$ to satisfy a sample condition (that the sample score be zero) that is analogous to the population score condition. One could sharpen the statement of this analogy by writing the population score as an explicit function of the population DGP, $\mu(\theta,F(\cdot,\theta_o)) \equiv E_{\mathbf{x}|\theta_o} \nabla_\theta l(x,\theta)$, and writing the sample score as $\mu(\theta,F_n) \equiv \mathbf{E}_n\nabla_\theta l(x,\theta)$, where "$\mathbf{E}_n$" stands for empirical expectation, or sample average. The mapping $\mu(\theta,\cdot)$ is linear in its second argument, and this is enough to assure that it is continuous (in an appropriate sense) in this argument. Then one has almost sure convergence of $\mu(\theta,F_n)$ to $\mu(\theta,F(\cdot,\theta_o))$ for each θ, from the Glivenko-Cantelli theorem.

A few additional regularity conditions are enough to ensure that this convergence is uniform in θ, and that a solution $T_n(\mathbf{x})$ that sets the sample score to zero converges almost surely to the value θ_o that sets the population score to zero.

The basic large sample properties of maximum likelihood estimators are that, subject to suitable regularity conditions, T_n converges in probability to the true parameter vector θ_o, and $n^{1/2}(T_n - \theta_o)$ converges in distribution to a normal random variable with mean zero and a variance which achieves the Cramer- Rao bound for an unbiased estimator. These results imply that in large samples, T_n will become a more and more precise estimate of the true parameter. Further, the convergence in distribution to a Normal permits one to use the properties of a Normal population to construct approximate hypothesis tests and confidence bounds, and get approximations for significance levels and power whose accuracy increases with sample size. The achievement of the Cramer-Rao lower bound on variance indicates that in large samples there are no alternative estimators which are uniformly more precise, so MLE is the "best" one can do.

We next list a series of regularity conditions under which the results stated above can be shown to hold. Only the single parameter case will be presented. However, the conditions and results have direct generalizations to the multiple parameter case. This list is chosen so the conditions are easy to interpret and to check in applications. Note that these are conditions on the population DGP, not on a specific sample. Hence, "checking" means verifying that your model of the DGP and your assumptions on distributions of random variables are logically consistent with the regularity conditions. They cannot be verified empirically by looking at the data, but it is often possible to set up and carry out empirical tests that may allow you to conclude that some of the regularity conditions fail. There are alternative forms for the regularity conditions, as well as weaker conditions, which give the same or similar limiting results. The regularity conditions are quite generic, and will be satisfied in many economic applications. However, it is a serious mistake to assume without checking that the DGP you assume for your problem is consistent with these conditions. While in most cases the mantra "I assume the appropriate regularity conditions" will work out, you can be acutely embarrassed if your DGP happens to be one of the exceptions that is logically inconsistent with the regularity conditions, particularly if it results in estimators that fail to have desirable statistical properties. Here are the conditions:

A.1. There is a single parameter θ which is permitted to vary in a closed bounded subset $\mathbf{\Theta}$. The true value θ_o is in the interior of $\mathbf{\Theta}$.

A.2. The sample observations are realizations of independently identically distributed random variables $x_1,...,x_n$, with a common density $f(x,\theta_o)$.

A.3. The density $f(x,\theta)$ is continuous in θ, and three times continuously differentiable in θ, for each x, and is "well behaved" (e.g., measurable or piecewise continuous or continuous) in x for each θ.

A.4. There exists a bound $\beta(x)$ on the density and its derivatives which is uniform in θ and satisfies $|l(x,\theta)| \le \beta(x)$, $(\nabla_\theta l(x,\theta))^2 \le \beta(x)$, $|\nabla_{\theta\theta} l(x,\theta)| \le \beta(x)$, $|\nabla_{\theta\theta\theta} l(x,\theta)| \le \beta(x)$, and $\int_{-\infty}^{+\infty} \beta(x)^2 f(x|\theta_o)dx < +\infty$. (Then, $\beta(x)$ is a dominating, square-integrable function.)

A.5 The function $\lambda(\theta) = \mathbf{E}_{x|\theta}\, l(x,\theta)$ has $\lambda(\theta) < \lambda(\theta_o)$ and $\nabla_\theta\lambda(\theta) \neq 0$ for $\theta \neq \theta_o$ and $J = -\nabla_{\theta\theta}\lambda(\theta_o) > 0$.

The expression J in A.5 is termed the *Fisher information* in an observation. The first two assumptions mostly set the problem. The restriction of the parameter to a closed bounded set guarantees that a MLE exists, and can be relaxed by adding conditions elsewhere. Requiring θ_o interior to Θ guarantees that the first-order condition $\mathbf{E}_n\nabla_\theta l(x,T_n(\cdot)) = 0$ for a maximum holds for large n, rather than an inequality condition for a maximum at a boundary. This really matters because MLE at boundaries can have different asymptotic distributions and rates of convergence than the standard $n^{1/2}$ rate of convergence to the normal. The continuity conditions A.3 are satisfied for most economic problems, and in some weak form are critical to the asymptotic distribution results. Condition A.4 gives bounds that permit exchange of the order of differentiation and integration in forming expectations with respect to the population density. Condition A.5 is an identification requirement which implies there cannot be a parameter vector other than θ_o that on average always explains the data as well as θ_o.

The next result establishes that under these regularity conditions, a MLE is consistent and asymptotically normal (CAN):

Theorem 6.4. If A.1-A.5 hold, then a maximum likelihood estimator T_n satisfies

(1) T_n is consistent for θ_o.

(2) T_n is asymptotically normal: $n^{1/2}(T_n(\mathbf{x}) - \theta_o) \rightarrow_d Z_o \sim N(0,J^{-1})$, with J equal to the Fisher information in an observation, $J = E_{x|\theta_o}\ \nabla_\theta l(x,\theta_o)^2$.

(3) $\mathbf{E}_n[\nabla_\theta l(x,T_n)]^2 \rightarrow_p J$ and $-\mathbf{E}_n\nabla_{\theta\theta} l(x,T_n) \rightarrow_p J$.

(4) Suppose T_n' is any sequence of estimators that solve equations of the form $\mathbf{E}_n g(x,\theta) = 0$, where g is twice continually differentiable and satisfies $E_{x|\theta_o}\ g(x,\theta) = 0$ if and only if $\theta = \theta_o$; uniform bounds $|g(x,\theta)| \leq \beta(x)$, $|\nabla_\theta g(y,\theta)^2| \leq \beta(x)$, $|\nabla_{\theta\theta} g(x,\theta)| \leq \beta(x)$, where $\mathbf{E}\beta(x)^2 < +\infty$; and $R = -\mathbf{E}\nabla_\theta g(y,\theta_o) \neq 0$. Let $S = \mathbf{E}g(x,\theta_o)^2$. Then $T_n' \rightarrow_p \theta_o$ and $n^{1/2}(T_n' - \theta^*) \rightarrow_d Z_1 \sim N(0,V)$, where $V = R^{-1}SR'^{-1}$. Further, $V \geq J^{-1}$, so that the MLE T_n is efficient relative to T_n'. Further, Z_0 and Z_1 have the covariance property $\mathrm{cov}(Z_0,Z_1 - Z_0) = 0$.

Result (2) in this theorem implies that to a good approximation in large samples, the estimator T_n is normal with mean θ_o and variance $(nJ)^{-1}$, where J is the Fisher information in an observation. Since this variance is the Cramer-Rao bound for an unbiased estimator, this also suggests that one is not going to be able to find other estimators that are also unbiased in this approximation sense and which have lower variance. Result 3 gives two ways of estimating the asymptotic variance J^{-1} consistently, where we use the fact that J^{-1} is a continuous function of J for $J \neq 0$, so that it can be estimated consistently by the inverse of a consistent estimator of J.. Result (4) establishes that MLE is efficient relative to a broad class of estimators called *M-estimators*.

Proof: An intuitive demonstration of the Theorem will be given rather than formal proofs. Consider first the consistency result. The reasoning is as follows. Consider the expected likelihood of an observation,

$$\lambda(\theta) \equiv E_{x|\theta_o}\, l(x,\theta) = \int_{-\infty}^{+\infty} l(x,\theta)f(x,\theta_o)dx.$$

We will argue that $\lambda(\theta)$ has a unique maximum at θ_o. Then we will argue that any function which is uniformly very close to $\lambda(\theta)$ must have its maximum near θ_o. Finally, we argue by applying a uniform law of large numbers that the likelihood function is with probability approaching one uniformly very close to λ for n sufficiently large. Together, these results will imply that with probability approaching one, T_n is close to θ_o for n large.

Assumption A.4 ensures that $\lambda(\theta)$ is continuous, and that one can reverse the order of differentiation and integration to obtain continuous derivatives

$$\nabla_\theta\lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)f(x,\theta_o)dx \equiv E_{x|\theta_o}\, \nabla_\theta l(x,\theta)$$

$$\nabla_{\theta\theta}\lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_{\theta\theta} l(x,\theta)f(x,\theta_o)dx \equiv E_{x|\theta_o}\, \nabla_{\theta\theta} l(x,\theta)$$

Starting from the identity

$$1 \equiv \int_{-\infty}^{+\infty} f(x,\theta)dx \equiv \int_{-\infty}^{+\infty} e^{l(x,\theta)}dx,$$

one obtains by differentiation

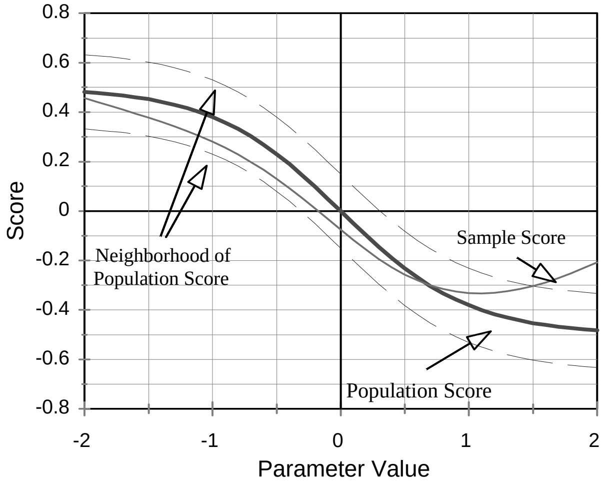
$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)e^{l(x,\theta)}dx$$

$$0 \equiv \int_{-\infty}^{+\infty} [\nabla_{\theta\theta} l(x,\theta) + \nabla_\theta l(x,\theta)^2]e^{l(x,\theta)}dx$$

Evaluated at θ_o, these imply $0 = \nabla_\theta\lambda(\theta_o)$ and $-\nabla_{\theta\theta}\lambda(\theta_o) = E_{x|\theta_o}\, \nabla_\theta l(x,\theta)^2 = J$.

Assumption A.5 requires further that $J \neq 0$, and that θ_o is the only root of $\nabla_\theta\lambda(\theta)$. Hence, $\lambda(\theta)$ has a unique maximum at θ_o, and at no other θ satisfies a first-order condition or boundary condition for a local maximum.

We argue next that any function which is close enough to $\nabla_\theta\lambda(\theta)$ will have at least one root near θ_o and no roots far away from θ_o. The figure below graphs $\nabla_\theta\lambda(\theta)$, along with a "sleeve" which is a vertical distance δ from $\nabla_\theta\lambda$. Any function trapped in the sleeve must have at least one root between $\theta_o - \varepsilon_1$ and $\theta_o + \varepsilon_2$, where $[\theta_o-\varepsilon_1,\theta_o+\varepsilon_2]$ is the interval where the sleeve intersects the axis, and must have no roots outside this interval. Furthermore, the uniqueness of the root θ_o of $\nabla_\theta\lambda(\theta)$ plus the condition $\nabla_{\theta\theta}\lambda(\theta_o) < 0$ imply that as δ shrinks toward zero, so do ε_1 and ε_2. In the graph, the sample score intersects the axis within the sleeve, but for parameter values near two is outside the sleeve. The last step in the consistency argument is to show that with probability approaching one the sample score will be entirely contained within the sleeve; i.e., that $L_n(\mathbf{x},\theta)$ is with probability approaching one contained in a δ-sleeve around $\lambda(\theta)$. For fixed θ, $L_n(\mathbf{x},\theta) = l(x_i,\theta)$ is a sample average of i.i.d. random variables $l(x,\theta)$ with mean $\lambda(\theta)$. Then Kolmogorov's SLLN implies $L_n(\mathbf{x},\theta) \to_{as} \lambda(\theta)$. This is not quite enough, because there is a question of whether $L_n(\mathbf{x},\theta)$ could converge non-uniformly to $\lambda(\theta)$, so that for any n there are some values of θ where $L_n(\mathbf{x},\theta)$ is outside the sleeve. However, assumptions A.1, A.3, and A.4 imply $\max_{\theta\in\Theta} |L_n(\mathbf{x},\theta) - \lambda(\theta)| \to_{as} 0$. This follows in particular because the differentiability of $f(x,\theta)$ in θ from A.3 and the bound on $\nabla_\theta l(x,\theta)$ from A.4 imply that $l(\cdot,\theta)$ is almost surely continuous on the compact set Θ, so that the uniform SLLN in Chapter 4.5 applies. This establishes that $T_n \to_{as} \theta$.



We next demonstrate the asymptotic normality of T_n. A Taylor′s expansion about θ of the first-order condition for maximization of the log likelihood function gives

(1) $0 = \nabla_\theta L_n(T_n) = \nabla_\theta L_n(\theta) + \nabla_{\theta\theta}L_n(\theta)\cdot(T_n-\theta) + \nabla_{\theta\theta\theta}L_n(T_{an})\cdot(T_n-\theta)^2/2$,

where T_{an} is some point between T_n and θ. Define the quantities

$$B_n = n^{-1}\sum_{i=1}^{n} \nabla_\theta l(y_i,\theta),\ C_n = n^{-1}\sum_{i=1}^{n} \nabla_{\theta\theta} l(y_i,\theta),\ D_n = n^{-1}\sum_{i=1}^{n} \nabla_{\theta\theta\theta} l(y_i,T_{an})$$

Multiply equation (1) by $n^{1/2}/(1+n^{1/2}|T_n-\theta|)$ and let $Z_n = n^{1/2}(T_n-\theta)/(l + n^{1/2}|T_n-\theta|)$. Then, one gets

$$0 = n^{1/2}B_n/(1+n^{1/2}|T_n-\theta|) + C_n Z_n + D_n Z_n(T_n-\theta)/2.$$

We make a limiting argument on each of the terms. <u>First</u>, the $\nabla_\theta l(y_i,\theta_o)$ are i.i.d. random variables with $\mathbf{E}\nabla_\theta l(y_i,\theta_o) = \nabla_\theta\lambda(\theta_o) = 0$ and $\mathbf{E}[\nabla_\theta l(y_i,\theta_o)]^2 = - \nabla_{\theta\theta}\lambda(\theta_o) = J$. Hence the Lindeberg-Levy CLT implies $B_n \to_d W_o \sim N(0,J)$. <u>Second</u>, $\nabla_{\theta\theta} l(Y_i,\theta_o)$ are i.i.d. random variables with $\mathbf{E}\nabla_{\theta\theta} l(Y_i,\theta_o) = -J$.

Hence the Khinchine WLLN implies $C_n \to_p -J$. Third, $|D_n| \leq n^{-1}\sum_{i=1}^{n} |\nabla_{\theta\theta\theta} l(y_i,T_{an})| \leq n^{-1}\sum_{i=1}^{n} \beta(y_i) \to_p \mathbf{E}\beta(Y) < +\infty$, by A.4 and Khinchine′s WLLN, so that $|D_n|$ is stochastically bounded. Furthermore, $|Z_n| \leq 1$, implying $Z_n = O_p(1)$. Since T_n is consistent, $(T_n - \theta_o) = o_p(1)$. Therefore, by rule 6 in Figure 4.3, $D_nZ_n(T_n- \theta_o)/2 = o_p(1)$.

Given $J/2 > \varepsilon > 0$, these arguments establish we can find n_o such that for $n > n_o$ with probability at least $1-\varepsilon$, we have $|D_nZ_n(T_n- \theta_o)/2| < \varepsilon$, $|C_n+J| < \varepsilon$ and $|B_n| < M$ for a large constant M (since $B_n \to_d W_o \Rightarrow B_n$ implies $O_p(1)$). In this event, $|C_n| > J-\varepsilon$, $|B_n + C_n n^{1/2}(T_n-\theta_o)| < \varepsilon(1 + n^{1/2}\cdot|T_n-\theta_o|)$, and $|B_n| \leq M$ imply $|C_n|n^{1/2}|T_n-\theta_o| - |B_n| \leq |B_n+C_nn^{1/2}|T_n-\theta_o)| < \varepsilon\cdot(1 - n^{1/2}\cdot|T_n-\theta_o|)$. This implies the inequality $(J - 2\varepsilon)n^{1/2}\cdot|T_n-\theta_o| < M + \varepsilon$. Therefore $n^{1/2}(T_n-\theta_o) = O_p(1)$; i.e., it is stochastically bounded. Therefore, by rule 6 in Figure 3.3, multiplying (2) by $1 + n^{1/2}\cdot|T_n-\theta_o|$ yields $0 = B_n + C_nn^{1/2}\cdot|T_n-\theta_o| + o_p(1)$. But $C_n \to_p -J < 0$ implies $C_n^{-1} \to_p -J^{-1}$. By rule 6, $(C_n+J^{-1})B_n = o_p(1)$ and $n^{1/2}(T_n-\theta_o) = J^{-1}B_n + o_p(1)$. The limit rules in Figure 3.1 then imply $J^{-1}B_n \to_d Z_o \sim N(0,J^{-1})$, $n^{1/2}\cdot|T_n-\theta_o| - J^{-1}B_n \to_p 0$, and hence $n^{1/2}\cdot|T_n-\theta_o| \to_d Z_o$.

The third result in the theorem is that J is estimated consistently by

$$(3) \quad J_n = n^{-1}\sum_{i=1}^{n} \nabla_\theta l(y_i,T_n)^2.$$

To show this, make a Taylor′s expansion of this expression around θ_o,

$$(4) \quad J_n = n^{-1}\sum_{i=1}^{n} l_\theta(y_i,\theta_o)^2 + 2\ n^{-1}\sum_{i=1}^{n} \nabla_\theta l(y_i,T_{an})\cdot\nabla_{\theta\theta} l(y_i,T_{an})(T_n-\theta_o).$$

We have already shown that the first term in (4) converges in probability to J. The second term is the product of $(T_n - \theta_o) \to_p 0$ and an expression which is bounded by $n^{-1}\sum_{i=1}^{n} 2\beta(y_i)^2 \to_p 2\mathbf{E}_Y\beta(Y)^2 < +\infty$, by Khinchine′s WLLN. Hence the second term is $o_p(1)$ and $J_n \to_p J$.

The final result in the theorem establishes that the MLE is efficient relative to any M-estimator T_n' satisfying $n^{-1}\sum_{i=1}^{n} g(y_i,T_n') = 0$, where g meets a series of regularity conditions. The first conclusion in this result is that T_n' is consistent and $n^{1/2}(T_n'-\theta_o)$ is asymptotically normal. This is actually of considerable independent interest, since many of the alternatives to MLE that are used in econometrics for reasons of computational convenience or robustness are M-estimators. Ordinary least squares is a leading example of an estimator in this class. The argument for the properties of T_n' are exactly the same as for the MLE case above, with g replacing $\nabla_\theta l$. The only difference is that R and S are not necessarily equal, whereas for $g = \nabla_\theta l$ in the MLE case, we had $R = S = J$. To make the efficiency argument, consider together the Taylor′s expansions used to get the asymptotic distributions of T_n and T_n',

$$0 = \nabla_\theta l(y_i,T_n) = n^{-1}\sum_{i=1}^{n} \nabla_\theta l(y_i,\theta_o) + n^{-1}\sum_{i=1}^{n} \nabla_{\theta\theta} l(y_i,\theta_o)\, n^{1/2}(T_n-\theta_o) + o_p(1)$$

$$0 = g(y_i,T_n') = n^{-1}\sum_{i=1}^{n} g(y_i,\theta_o) + n^{-1}\sum_{i=1}^{n} g_\theta(Y_i,\theta_o) n^{1/2}(T_n'-\theta_o) + o_p(1)$$

Solving these two equations gives

$$n^{1/2}(T_n-\theta_o) = J^{-1}W_n + o_p(1)$$

$$n^{1/2}(T_n'-\theta_o) = R^{-1}U_n + o_p(1)$$

with $W_n = n^{-1/2}\sum_{i=1}^{n} \nabla_\theta l(y_i,\theta_o)$ and $U_n = n^{-1/2}\sum_{i=1}^{n} g(y_i,\theta_o)$. Consider any weighted average of these equations,

$$n^{1/2}((1-\gamma)T_n + \gamma T_n' - \theta_o) = J^{-1}(1-\gamma)W_n + R^{-1}\gamma U_n + o_p(1)\,.$$

The Lindeberg-Levy CLT implies that this expression is asymptotically normal with mean zero and variance

$$\Omega = J^{-2}(1-\gamma)^2\mathbf{E}\nabla_\theta l(Y|\theta_o)^2 + R^{-2}\gamma^2\mathbf{E}g(Y,\theta_o)^2 + 2J^{-1}R^{-1}(1-\gamma)\gamma\mathbf{E}l_\theta(Y|\theta_o)g(Y,\theta_o)$$

.
The condition $0 \equiv \int g(y,\theta)f(y|\theta)dy \equiv \int g(y,\theta)e^{l(y|\theta)}dy$, implies, differentiating under the integral sign,

$$0 \equiv \int \nabla_\theta g(y,\theta)e^{l(y,\theta)}dy + \int \nabla_\theta l(y,\theta)g(y,\theta)e^{l(y,\theta)}dy\,.$$

Evaluated at θ_o, this implies $0 \equiv -R + \mathbf{E}\nabla_\theta l(Y|\theta_o)g(Y,\theta_o)$. Hence,

$$\Omega = J^{-1}(1-\gamma)^2 + R^{-2}S\,\gamma^2 + 2(1-\gamma)\gamma\, J^{-1}R^{-1}R = J^{-1} + [R^{-2}S - J^{-1}]\gamma^2.$$

Since $\Omega \geq 0$ for any γ, this requires $V = R^{-2}S \geq J^{-1}$, and hence $\Omega \geq J^{-1}$. Further, note that

$$\Omega = \text{var}(Z_o+\gamma(Z_1-Z_o)) = \text{var}(Z_o) + \gamma^2\,\text{var}(Z_1-Z_o) + 2\gamma\,\text{cov}(Z_o,Z_1-Z_o)\,,$$

and $\text{var}(Z_o) = J^{-1}$, implying

$$2\gamma\,\text{cov}(Z_o, Z_1 - Z_o) \geq -\gamma^2\,\text{var}(Z_1 - Z_o)\,.$$

Taking γ small positive or negative implies $\text{cov}(Z_o, Z_1 - Z_o) = 0$. □

THIS PAGE LEFT BLANK FOR FUTURE MATERIAL