2

Theory and Properties of Maximum Likelihood Estimators

Maximum likelihood is a unified framework for model building and statistical inference. Others approach maximum likelihood as simply a technique for model estimation that happens to have desirable statistical properties. For Bayesians, the process of specifying a likelihood is core to the statistical enterprise. In this chapter we examine the likelihood function in greater detail, exploring features and assumptions commonly invoked in estimation of likelihood-based models. We identify and derive statistical properties of the maximum likelihood estimator critical for justifying the computational strategies subsequently developed in this book. The key result in this regard is the MLE's asymptotic normality. This chapter is the most technically dense in this volume, focusing on "just the math." We defer computational examples for more extensive treatment in subsequent chapters.

2.1 THE LIKELIHOOD FUNCTION: A DEEPER DIVE

The results of statistical analysis often appear in published work as a Big Ugly Table of Numbers (BUTON). BUTONs display select quantities as reported by a particular procedure applied to specific, observed data. But most statistical models both entail assumptions and engender much more output than what typically appears in a BUTON. In building a likelihood-based model we posit a joint probability distribution for the observed data. In estimating the specified model we construct a generative analogue of the processes at work. Formally,

> **Definition 2.1** (Likelihood and MLE)**.** Suppose we observe data, x, assumed to have been generated under the probability model $f(x;\theta^*)$, where the "true" parameter θ^* is unknown. The *likelihood* of θ given x is

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

$$\mathcal{L}(\theta \mid x) = h(x)f(x;\theta)$$
$$\propto f(x;\theta),$$

for some function h. The *Maximum Likelihood Estimator (MLE) for* θ^*, denoted $\hat{\theta}$, is

$$\hat{\theta} \equiv \arg\max_{\theta} \mathcal{L}(\theta \mid x).$$

Note the presence of $h(x)$, an *unknown* scaling factor that is a function of the data but not the parameter. This factor reflects the likelihood principle.

Definition 2.2 (Likelihood Principle)**.** The *likelihood principle* says that, after data have been collected, all the information relevant for inference about θ is contained in the likelihood function for the observed data under the assumed probability model. Moreover, two likelihood functions that are proportional to one another contain the same information about θ.

Thus, the likelihood has no scale nor any direct interpretation as a probability. Rather, we interpret likelihoods relative to one another. In what follows we suppress the term $h(x)$, as it only adds to notational complexity, and it cannot be estimated in any case.

Information about θ^* comes from *both* the observed data and our assumptions about the data-generating process (DGP), as embodied in the probability model $f(x;\theta)$. The likelihood function does more than generate values for tables. It provides a mathematical structure in which to incorporate new data or make probabilistic claims about yet-to-be-observed events.

As we saw in Chapter 1, likelihoods from independent samples sharing a DGP are easily combined. If x_1 and x_2 are independent events governed by the same $f(x;\theta^*)$, then $\mathcal{L}(\theta \mid x_1, x_2) = f(x_1;\theta)f(x_2;\theta)$, and the log-likelihood is $\log \mathcal{L}(\theta \mid x_1, x_2) = \log f(x_1;\theta) + \log f(x_2;\theta)$.

2.1.1 Likelihood and Bayesian Statistics

R. A. Fisher came to believe that likelihood is a self-contained framework for statistical modeling and inference. But the likelihood also plays a fundamental role in Bayesian statistics.

Bayesian thinking builds on Thomas Bayes's Rule, which provides a principled way in which to combine our prior knowledge and beliefs with newly acquired data to generate an updated set of beliefs. Formally, if we are concerned with the value of $\boldsymbol{\theta}$, we might describe our current beliefs about that value with the prior probability distribution $\Pr(\boldsymbol{\theta})$. We then observe new data, $\mathbf{x} = (x_1, \ldots, x_n)$, and seek to describe our updated or "posterior" beliefs, $\Pr(\boldsymbol{\theta} \mid \mathbf{x})$. Bayes's Rule states

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Theorem 2.1 (Bayes's Rule).

$$\begin{aligned}\Pr(\boldsymbol{\theta} \mid \mathbf{x}) &= \frac{\Pr(\mathbf{x} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})}{\Pr(\mathbf{x})} \\ &\propto \Pr(\mathbf{x} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) \\ &\propto \mathcal{L}(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}).\end{aligned}$$

Proof The result follows directly from the fact that $\Pr(\boldsymbol{\theta}, \mathbf{x}) = \Pr(\boldsymbol{\theta} \mid \mathbf{x}) \Pr(\mathbf{x}) = \Pr(\mathbf{x} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$. □

From Bayes's Rule we see that the posterior distribution of $\boldsymbol{\theta}$ is the prior times the likelihood, scaled by the marginal probability of the data, $\mathbf{x}$. As the volume of observed data increases, the information in the data dominate our prior beliefs and the Bayesian posterior distribution comes to resemble the likelihood function. If we assume that $\Pr(\boldsymbol{\theta})$ follows a uniform distribution with support that includes $\hat{\boldsymbol{\theta}}$, then the mode of the posterior distribution will be the MLE. Under somewhat more general conditions, the posterior distribution for $\boldsymbol{\theta}$ approximates the asymptotic distribution for the MLE discussed in Section 2.2.2. Although Bayesian thinking and the analysis of fully Bayesian models is much more involved, the likelihood framework and the logic of Bayesian statistics are closely linked. In many situations, a parametric Bayesian approach and the likelihood framework will yield similar results, even if the interpretation of these estimates differs.

2.1.2 Regularity

The method of maximum likelihood summarizes the information in the data and likelihood function with $\hat{\boldsymbol{\theta}}$. To find this maximum we might simply try many values of $\boldsymbol{\theta}$ and pick the one yielding the largest value for $\mathcal{L}(\boldsymbol{\theta})$. This is tedious and carries no guarantee that the biggest value we found is the biggest that could be obtained. Calculus lets us find the extrema of functions more quickly and with greater certainty. The conditions under which we can apply all the machinery of differential calculus in the likelihood framework are referred to as "regularity conditions."

These regularity conditions can be stated in several ways. Some are quite technical. Figures 1.4 and 1.1 provide some intuition. For a likelihood to be "regular," we would like it to be smooth without any kinks or holes (i.e., continuous), at least in the neighborhood of $\hat{\boldsymbol{\theta}}$. In regular problems, we also require that the MLE not fall on some edge or boundary of the parameter space, nor does the support of the distribution or mass function for X depend on the value of $\boldsymbol{\theta}$. Likelihoods with a single maximum, rather than many small, local hills or plateaus, are easier to work with.

With regularity concerns satisfied, the log-likelihood will be at least twice differentiable around the MLE. Each of these derivatives of the log-likelihood is important enough to have a name, so we address each in turn.

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

2.1.3 Score

The vector of first partial derivatives of a multivariate function is called the *gradient*, denoted ∇. The score function is the gradient of the log-likelihood. Evaluated at a particular point $\boldsymbol{\theta}$, the score function will describe the steepness of the log-likelihood. If the function is continuous, then at the maximum or minimum, this vector of first partial derivatives equals the zero vector, $\mathbf{0}$.

Definition 2.3 (Score function)**.** Given log-likelihood, $\log \mathcal{L}(\boldsymbol{\theta})$, and observed data x, the *score function*, $S(\boldsymbol{\theta})$, is

$$S(\boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}),$$

and the MLE $\hat{\boldsymbol{\theta}}$ solves the score equation: $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

If we imagine that the data we observe are but one realization of all observable possible data sets given the assumed probability model, $f(x;\boldsymbol{\theta})$, then we can consider the score function as a random variable. As a matter of vocabulary, we refer to the *score statistic* when the score is considered as a random variable (with $\boldsymbol{\theta}$ fixed at the "true" value for the DGP). In Theorem 2.2 we derive the expected value of the score statistic.

Theorem 2.2. *If the likelihood function $\mathcal{L}(\boldsymbol{\theta})$ satisfies appropriate regularity conditions, then $E[S(\boldsymbol{\theta})] = \mathbf{0}$, where the expectation is taken over possible data X with support $\mathcal{X}$.*

Proof

$$\begin{aligned}
\mathrm{E}[S(\boldsymbol{\theta})] &= \int_{\mathcal{X}} f(x;\boldsymbol{\theta} \mid \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}) dx \quad \text{(score def., expectation)} \\
&= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})}{\mathcal{L}(\boldsymbol{\theta})} f(x;\boldsymbol{\theta} \mid \boldsymbol{\theta}) dx \quad \text{(Chain Rule, derivative of ln)} \\
&= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})}{f(x;\boldsymbol{\theta} \mid \boldsymbol{\theta})} f(x;\boldsymbol{\theta} \mid \boldsymbol{\theta}) dx \quad \text{(likelihood def.)} \\
&= \int_{\mathcal{X}} \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) dx \quad \text{(algebra)} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{X}} \mathcal{L}(\boldsymbol{\theta}) dx \quad \text{(regularity conditions)} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{X}} f(x;\boldsymbol{\theta} \mid \boldsymbol{\theta}) dx \quad \text{(likelihood def.)} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} 1 = 0 \quad \text{(pdf def., derivative of a constant).}
\end{aligned}$$

□

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

2.1.4 Fisher Information

Recall the likelihood surface displayed in Figure 1.1. The peak of this curve represents the most likely value of the parameter to have generated the observed data, whereas points near the peak are less likely. A likelihood function steeply curved around the MLE is one in which values of $\boldsymbol{\theta}$ close to $\hat{\boldsymbol{\theta}}$ are much less likely than $\hat{\boldsymbol{\theta}}$. The data, in the context of the likelihood function, provide a lot of information about $\boldsymbol{\theta}$. It is precisely estimated. Conversely, a likelihood function that is nearly flat around the MLE is one in which finding the maximum is hard since values near the maximum are nearly as likely as the MLE. In the limiting case of a flat likelihood, a single MLE does not exist since several values are equally consistent with the observed data.

The second derivative is closely related to the notion of curvature of a function. It is no surprise that the second derivative of the log-likelihood is critical in both estimating the MLE and describing our uncertainty. Intuitively the bigger the absolute value of the second derivative the more "curved" the function. Since the second derivative is negative at any local maximum we know that the second derivative of the likelihood will be negative at $\hat{\boldsymbol{\theta}}$ (assuming appropriate regularity conditions). Taking the negative of the second derivative gives us an index of how much information about $\boldsymbol{\theta}$ we have at the MLE. This quantity is called *Fisher information*.

Definition 2.4 (Observed Fisher Information)**.** Assuming appropriate regularity conditions for $f(x;\boldsymbol{\theta})$, the *observed Fisher Information*, $I(\boldsymbol{\theta})$, is given as

$$I(\hat{\boldsymbol{\theta}}) \equiv -\frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}}.$$

Note that observed Fisher information is a quantity, not a function. It is calculated by evaluating the second derivative of the log-likelihood at the MLE, itself calculated using the observed data.

Just as we did with the score function and score statistic we might think of the observed Fisher information as resulting from only one of many possible data sets we could observe from the DGP described by $f(x;\boldsymbol{\theta})$. If we imagine $\boldsymbol{\theta}$ as fixed and then average over possible data sets, we have *expected* Fisher information, i.e., how much information about $\boldsymbol{\theta}$ can we expect from the random variable X. We will define expected Fisher information a bit differently from observed information and then show how we can get from expected information to our expression for observed information.

Definition 2.5 (Expected Fisher Information)**.** Assuming appropriate regularity conditions for $f(x;\boldsymbol{\theta})$, the *expected Fisher Information*, $\mathcal{I}(\boldsymbol{\theta})$, is given as

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \mathrm{E}[S(\boldsymbol{\theta})S(\boldsymbol{\theta})^\intercal],$$

where the expectation is taken over possible data X with support $\mathcal{X}$.

There are several equivalent ways of writing the expected information, including as the negative expected derivative of the score:

Theorem 2.3. *Assuming appropriate regularity conditions for $f(x;\boldsymbol{\theta})$,*

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{var}[S(\boldsymbol{\theta})] \tag{2.1}$$

$$= -\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})\right], \tag{2.2}$$

where the expectation is taken over possible data X with support $\mathcal{X}$.

Proof We state the proof for the case where θ is single-valued (scalar), but it readily generalizes to the case where $\boldsymbol{\theta}$ is a vector. Noting that $\mathrm{var}(Y) = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2$, and by applying Theorem 2.2, we obtain Equation (2.1).

To prove Equation (2.2), note that $S(\theta) = \frac{\partial}{\partial \theta} \log f(x;\theta) = \frac{f'(x;\theta)}{f(x;\theta)}$. Then by the quotient rule,

$$\begin{aligned}
\frac{\partial}{\partial \theta} S(\theta) &= \frac{f''(x;\theta)f(x;\theta) - f'(x;\theta)^2}{f(x;\theta)^2} \\
&= \frac{f''(x;\theta)}{f(x;\theta)} - \left(\frac{f'(x;\theta)}{f(x;\theta)}\right)^2 \\
&= \frac{f''(x;\theta)}{f(x;\theta)} - S(\theta)^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial}{\partial \theta} S(\theta)\right] &= \int_{\mathcal{X}} \left(\frac{f''(x;\theta)}{f(x;\theta)} - S(\theta)^2\right) f(x;\theta)\,dx \\
&= \int_{\mathcal{X}} f''(x;\theta)\,dx - \int_{\mathcal{X}} S(\theta)^2 f(x;\theta)\,dx \\
&= \int_{\mathcal{X}} f''(x;\theta)\,dx - \mathcal{I}(\theta).
\end{aligned}$$

The assumed regularity conditions let us reverse the order of differentiation and integration, implying that the first term of the last expression is $\frac{\partial^2}{\partial \theta^2} \int_{\mathcal{X}} f(x;\theta)\,dx = 0$, since $f(x;\theta)$ is a probability density function. □

When $\boldsymbol{\theta}$ is a k-dimensional vector, we use $H(\boldsymbol{\theta})$ to refer to its matrix of second partial derivatives, also known as the *Hessian* matrix.[1]

[1] The name Hessian refers to the mathematician Ludwig Otto Hesse, not the German mercenaries contracted by the British Empire.

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Definition 2.6 (Hessian matrix)**.** Given a twice-differentiable function such as a regular log-likelihood $\log \mathcal{L}(\boldsymbol{\theta})$ with k-dimensional parameter $\boldsymbol{\theta}$, the $k \times k$ *Hessian matrix* $H(\boldsymbol{\theta})$ is given as

$$\begin{aligned} H(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta}) \\ &= \frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta \partial \theta^\intercal} \\ &= \begin{bmatrix} \dfrac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1^2} & \cdots & \dfrac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \cdots & \dfrac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k^2} \end{bmatrix}. \end{aligned}$$

Expected Fisher information, written in terms of the Hessian matrix, is therefore

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \mathrm{E}\left[S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^\intercal\right] = -\mathrm{E}[H(\boldsymbol{\theta})].$$

In any particular application, the observed Fisher information is now a $k \times k$ matrix of numbers.

The distinction between expected and observed Fisher information is somewhat subtle. Observed information is closest to the actual data in any particular application. Statistical programs like $\mathcal{R}$ use and report quantities based on the observed Fisher information. In many routine applications using models from the exponential family, the observed and expected Fisher information are equivalent at the MLE. Expected information is a theoretical idea used mainly to derive some of the MLE's properties that are utilized in frequentist statistics. While these properties may be interesting, the idea of averaging across hypothetical data sets generated by a fixed parameter sits uneasily with the basic likelihood principle of treating observed data as fixed and parameters as unknown quantities.

Fisher Information from Independent Samples

Just as we can combine likelihoods from independent events or samples governed by the same DGP, we can also combine Fisher information. If we observe n independent samples from $f(x; \boldsymbol{\theta})$ then we can construct the log-likelihood as

$$\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \log \mathcal{L}(x_i; \boldsymbol{\theta}).$$

It follows that $H(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} H(\boldsymbol{\theta} \mid x_i)$. So, from the definition of $\mathcal{I}(\boldsymbol{\theta})$, we get

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

$$\mathcal{I}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \mathrm{E}\left[H(\boldsymbol{\theta} \mid x_i)\right] = n\mathcal{I}(\boldsymbol{\theta} \mid x_i).$$

In other words, the expected Fisher information from a random sample of size n is simply n times the expected information from a single observation.

2.2 PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

With knowledge of the score, Hessian, and Fisher information in hand, we can derive some of the most useful statistical properties of the likelihood function and MLE.

2.2.1 Invariance

One frequently useful property of the MLE is that, for any sample size, the MLE is *functionally invariant*:

> **Theorem 2.4.** *Let $g(\cdot)$ be a function. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $g(\hat{\boldsymbol{\theta}}) = \widehat{g(\boldsymbol{\theta})}$.*

Proof We prove this result supposing that $g(\cdot)$ is injective (this is a simplifying, though not necessary, assumption). If $g(\cdot)$ is one-to-one, then $\mathcal{L}(g^{-1}(g(\boldsymbol{\theta})) = \mathcal{L}(\boldsymbol{\theta})$. Then, since $\hat{\boldsymbol{\theta}}$ is the MLE, our previous statement implies that $\hat{\boldsymbol{\theta}} = g^{-1}(\widehat{g(\boldsymbol{\theta})})$. Therefore applying $g(\cdot)$ to both sides, we see that $g(\hat{\boldsymbol{\theta}}) = \widehat{g(\boldsymbol{\theta})}$. □

In words, a transformation of the MLE equals the MLE of that transformation. We can use the invariance property to easily re-parameterize our models for estimation and interpretation. For example, if we estimated a variance parameter, $\widehat{\sigma^2}$, we can easily calculate the MLE for the standard deviation ($\hat{\sigma} = \sqrt{\widehat{\sigma^2}}$) or precision ($\widehat{\sigma^{-2}} = 1/\widehat{\sigma^2}$).

2.2.2 Large-Sample Properties

Let $\hat{\boldsymbol{\theta}}_n$ be the MLE for the parameters $\boldsymbol{\theta}$ given sample size n and probability model $f(x; \boldsymbol{\theta})$. The other oft-cited properties of MLEs are asymptotic, i.e., they are derived as the limiting behavior of the MLE for a sequence of $\{\hat{\boldsymbol{\theta}}_n\}$ as we let $n \to \infty$. The theoretical basis for these results relies on the Law of Large Numbers and the Central Limit Theorem. We state both here by way of review and without proofs, which can be quite technical.

> **Theorem 2.5** (Weak Law of Large Numbers). *Assume an infinite sequence of independent and identically distributed (i.i.d.) random*

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

variables, X_i, with finite expected value, μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\bar{X}_n \xrightarrow{p} \mu$ as $n \to \infty$.

Theorem 2.6 (Central Limit Theorem). *Assume an infinite sequence of i.i.d. random variables, X_i, with finite expected value, μ, and finite variance, σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$.*

In case you were wondering … 2.1 ≈ Taylor Series

The Taylor series approximation to a function is a widely used tool in statistics, especially when relying on the Central Limit Theorem to derive asymptotic results. A Taylor series represents a differentiable function at a point by a polynomial approximation at that same point. In the theory of maximum likelihood we are most commonly interested in quadratic approximations.

Suppose we have a function $\log \mathcal{L}(\theta)$ and we are interested in its approximation around a point, $\hat{\theta}$. The second-order, or quadratic Taylor approximation around $\hat{\theta}$ is

$$\log \mathcal{L}(\theta) = \log \mathcal{L}(\hat{\theta}) + (\theta - \hat{\theta}) \frac{\partial}{\partial \theta} \log \mathcal{L}(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta}) \frac{\partial^2}{\partial \theta \partial \theta^\intercal} \log \mathcal{L}(\hat{\theta}) + R,$$

where R is the remainder. We typically work with a truncated Taylor series omitting the remainder, often described as an "approximation" ($\approx$).

Consistency of the MLE

Suppose that the data generating process for x is $f(x, \boldsymbol{\theta}^*)$, so that $\boldsymbol{\theta}^*$ is the "true" value of $\boldsymbol{\theta}$ that generates the data. Then the MLE $\hat{\boldsymbol{\theta}}$ is *consistent*; that is, given infinite data, the MLE collapses to the "true" parameter value $\boldsymbol{\theta}^*$. We present one statement of the theorem and give a sketch of the proof for scalar θ to provide the key intuition.

Theorem 2.7 (Consistency of the MLE). *Let $\boldsymbol{\theta}^*$ be "true" value for $\boldsymbol{\theta}$. Assuming appropriate regularity conditions for $f(\mathbf{x}; \boldsymbol{\theta}^*)$, and that $\mathbf{x} = (x_1, \ldots, x_n)$ forms an i.i.d. sample from $f(x; \boldsymbol{\theta}^*)$, and further that $\mathcal{L}(\boldsymbol{\theta})$ is globally concave, then $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$ as $n \to \infty$.*

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Proof We sketch the proof for the case of scalar θ, but it readily generalizes to the multiparameter case. First note that the sample likelihood function converges to the expected log-likelihood by the Law of Large Numbers:

$$\frac{1}{n}\log \mathcal{L}(\theta \mid \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\log \mathcal{L}(\theta \mid x_i)$$

$$\xrightarrow{p} \mathrm{E}\left[\log \mathcal{L}(\theta \mid \mathbf{x})\right].$$

We then show that θ^* is the maximizer of the expected log-likelihood by showing that θ^* satisfies the first-order conditions (FOC) for a maximum. The FOC are

$$\frac{\partial}{\partial\theta}\mathrm{E}\left[\log \mathcal{L}(\theta^* \mid \mathbf{x})\right] = \int_{\mathcal{X}} \frac{f'(x;\theta)}{f(x;\theta)} f(x;\theta^*)dx = 0,$$

which follows from the regularity conditions and the definition of the log-likelihood. The FOC are satisfied when $\theta = \theta^*$ by Theorem 2.2. □

Asymptotic Distribution of the MLE

The MLE is *asymptotically normally distributed*. This property justifies conventional hypothesis testing and z-scores for MLE results. As we will show shortly, the square root of the main diagonal of $-H(\hat{\boldsymbol{\theta}})^{-1}$ gives us the standard errors so frequently seen in BUTON. But the likelihood approach contains much more than just point estimates and standard errors. The normality result allows us to use modern computing technology to combine the $\hat{\boldsymbol{\theta}}$, the Hessian matrix, and the (multivariate) normal distribution to describe our models in much more interesting and informative ways. The normality result provides the framework for simulating nearly any quantity of interest from a likelihood-based model, including predictions of yet-to-be-observed data. The MLE's limiting distribution has a direct connection to the Bayesian approach of interpreting models via the posterior distribution.

We state Slutsky's Theorem without proof.

Theorem 2.8 (Slutsky's Theorem). *Let A_n, B_n, and X_n be sequences of random variables such that $A_n \xrightarrow{p} a$, $B_n \xrightarrow{p} b$, and $X_n \xrightarrow{d} X$ then $A_n + B_n X_n \xrightarrow{d} a + bX$.*

Theorem 2.9 (Distribution of the MLE). *Let $\mathbf{x} = (x_1, \ldots, x_n)$ form an i.i.d. sample from $f(x;\boldsymbol{\theta}^*)$ with corresponding likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})$, which satisfies the regularity conditions for a consistent MLE. Then $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$.*

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Proof The proof begins with a Taylor series expansion of the FOC around $\hat{\boldsymbol{\theta}}_n$, where $\boldsymbol{\theta}_0$ is some point between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$:

$$0 = S(\hat{\boldsymbol{\theta}}_n) \approx S(\boldsymbol{\theta}^*) + H(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*).$$

It follows that

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) &= -H(\boldsymbol{\theta}_0)^{-1} S(\boldsymbol{\theta}^*) \\ \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) &= -\left(\frac{1}{n} H(\boldsymbol{\theta}_0)\right)^{-1} \frac{1}{\sqrt{n}} S(\boldsymbol{\theta}^*).\end{aligned}$$

By the Law of Large Numbers we have

$$H(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^{n} H(\boldsymbol{\theta}_0 \mid x_i) \xrightarrow{p} \mathrm{E}\left[H(\boldsymbol{\theta}^*)\right] = -\mathcal{I}(\boldsymbol{\theta}^*)$$

and, by the Central Limit Theorem, we have

$$\frac{1}{\sqrt{n}} S(\boldsymbol{\theta}^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(\boldsymbol{\theta}^* \mid x_i) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\boldsymbol{\theta}^*)\right).$$

Therefore the result follows from Theorem 2.8. □

In actual applications we need to estimate the asymptotic variance of the MLE. Since the MLE is consistent, we can do this by substituting observed Fisher information, $I(\hat{\boldsymbol{\theta}})^{-1}$, for the expected Fisher information, $\mathcal{I}(\boldsymbol{\theta}^*)^{-1}$. This result implies that $\hat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}(\boldsymbol{\theta}^*, I(\hat{\boldsymbol{\theta}})^{-1})$, or, in words, the sampling distribution of the MLE is approximately normal with the mean centered at the true value of $\boldsymbol{\theta}$ and covariance matrix given by the inverse of the Fisher information, reinforcing the intuition that more information should lead to more precise estimates.

Recalling that $I(\hat{\boldsymbol{\theta}}) = -H(\hat{\boldsymbol{\theta}})$, we can also see that the variance of the MLE is given by the negative inverse of the Hessian. The diagonal elements of this covariance matrix give the variance of each element of $\boldsymbol{\theta}$. If we take the square root of the diagonal elements, we obtain our estimates of the standard errors for $\hat{\boldsymbol{\theta}}$.

Other Properties

We mention two other properties of the MLE only in passing. First, the asymptotic distribution of the MLE implies that the MLE is *efficient* in that it reaches the Cramér-Rao lower bound. The MLE achieves this *asymptotically*. In other words, the MLE attains the smallest possible variance among all consistent estimators. Equivalently, the MLE achieves (asymptotically) the smallest mean-squared error of all consistent estimators.

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

In case you were wondering … 2.2 Cramér-Rao lower bound

The Cramér-Rao Lower Bound (CRLB) states that if $T(X)$ is an unbiased estimator for θ, then $\text{var}(T(X)) \geq \mathcal{I}(\theta)^{-1}$. Any unbiased estimator achieving the CRLB achieves the minimum possible variance and therefore attains maximal efficiency.

There is *no* guarantee that the MLE is unbiased. For example, the MLE for the normal variance is biased. In fact, the MLE is frequently biased in small samples. But bias is a second-order concern here for three reasons. First, bias disappears as n increases (Theorem 2.7). In most applications, we are well away from the point where small-sample bias will be material. Second, bias in the MLE can be estimated and accounted for, if needed, using the cross-validation methods discussed in Chapter 5. Third, even if $T(X)$ is unbiased, any nonlinear transformation of $T(X)$ will be biased; unbiasedness is not robust to re-parameterization.

2.3 DIAGNOSTICS FOR MAXIMUM LIKELIHOOD

How can we evaluate the results of a model estimated by maximum likelihood? Since we cannot recover the constant $h(x)$ the value of the (log) likelihood at the MLE is, by itself, not meaningful. What we can do, however, is compare likelihoods for different values of $\boldsymbol{\theta}$ given the same underlying observed data and likelihood function. The value of the likelihood at the MLE enters into several of the most commonly discussed model diagnostic and comparison tools: the *likelihood ratio*, the *score* test, the *Wald* statistic, the the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The first three of these all have asymptotic distributions that underpin conventional frequentist hypothesis testing.

2.3.1 Likelihood Ratios

The most straightforward way to compare likelihoods is simply to divide one by the other, canceling out the unknown scaling factor, $h(\mathbf{x})$.

Definition 2.7 (Likelihood ratio)**.** Given a probability model $f(x;\boldsymbol{\theta})$, observed data x, and two hypothesized values of $\boldsymbol{\theta}$, namely $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}_G$, we define the likelihood ratio as

$$\log\left(\frac{\mathcal{L}(\boldsymbol{\theta}_R \mid x)}{\mathcal{L}(\boldsymbol{\theta}_G \mid x)}\right) = \log \mathcal{L}(\boldsymbol{\theta}_R \mid x) - \log \mathcal{L}(\boldsymbol{\theta}_G \mid x).$$

Since the value of the log-likelihood for the sample is the sum of the likelihoods for each observation, the likelihood ratio must be calculated based

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

on the *exact same* observed data for both. The only thing that differs between the likelihood functions in the numerator and denominator of the ratio is the candidate values of $\boldsymbol{\theta}$. With this likelihood ratio we can assess the relative strength of evidence in favor of particular $\boldsymbol{\theta}$ values.

The likelihood ratio is used to compare "nested" models, i.e., two models in which one is a restricted or special case of the other. Recall the simple regression of CO_2 on per capita GDP from Section 1.4.1. In that example $\boldsymbol{\theta}_R = (\beta_0, \beta_1, \sigma^2)$. Suppose we fit a second model, wherein we include population density as a covariate with associated regression coefficient β_2. In the second model, we have $\boldsymbol{\theta}_G = (\beta_0, \beta_1, \beta_2, \sigma^2)$. In the first model, we implicitly constrained the β_2 coefficient to be 0. The first model is thus a special case of the more general second one. If the restriction is appropriate, i.e., $\boldsymbol{\theta}^*$ lies in the more restricted parameter space, then the (log) likelihoods of the two models should be approximately equal at their respective MLEs. The likelihood ratio should be approximately one or, equivalently, the difference in the log-likelihoods should be about 0. For reasons that will be clear shortly, the conventional definition of the *likelihood ratio statistic* is

$$\mathrm{LR}(\boldsymbol{\theta}_R, \boldsymbol{\theta}_G \mid \mathbf{x}) = -2 \log \frac{\mathcal{L}(\boldsymbol{\theta}_R \mid \mathbf{x})}{\mathcal{L}(\boldsymbol{\theta}_G \mid \mathbf{x})}.$$

Some presentations reverse the numerator and denominator of the likelihood ratio, leading to a likelihood ratio statistic without the negative sign.

Distribution of the Likelihood Ratio Statistic

In the CO_2 emissions example, we obtain a log-likelihood of -395.2 for the restricted model, i.e., the model excluding population density, whereas we obtain -392.4 for the more general model. Our likelihood ratio is -2.8 and our likelihood ratio statistic is 5.7. Is this a lot? It turns out that under the assumption that the parameter space for the restricted model is correct (along with regularity conditions), the likelihood ratio follows a known asymptotic distribution.

In case you were wondering … 2.3 χ^2 distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(0, 1)$. Let $Z = \sum_{i=1}^{n} X_i^2$. We say that

$$Z \sim \chi_n^2.$$

This distribution has n "degrees of freedom," with $\mathrm{E}[Z] = n$ and $\mathrm{var}(Z) = 2n$.

Theorem 2.10. *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *be an i.i.d. sample from some distribution* $f(x; \boldsymbol{\theta}^*)$ *with corresponding likelihood* $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})$ *satisfying*

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

standard regularity conditions. Let $r = dim(\boldsymbol{\theta}_R) \leq dim(\boldsymbol{\theta}_G) = g$*; that is, the elements of* $\boldsymbol{\theta}_R$ *are a subset of those of* $\boldsymbol{\theta}_G$*. Under the hypothesis that* $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_R$*,*

$$LR(\hat{\boldsymbol{\theta}}_R, \hat{\boldsymbol{\theta}}_G \mid \mathbf{x}) \xrightarrow{d} \chi^2_{g-r}.$$

Proof We sketch the proof for scalar θ. We suppress the subscript n on the MLEs.

Begin with a second-order Taylor expansion of the log-likelihood around the MLE:

$$\begin{aligned}\log \mathcal{L}(\theta_r) &\approx \log \mathcal{L}(\hat{\theta}) + (\hat{\theta} - \theta_r)S(\hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta_r)^2 \\ &\approx \log \mathcal{L}(\hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta_r)^2 \qquad (\text{by } S(\hat{\theta}) = 0).\end{aligned}$$

This implies that the likelihood ratio statistic is

$$\begin{aligned}\text{LR} &\approx -2\left[\log \mathcal{L}(\hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta_r)^2 - \log \mathcal{L}(\hat{\theta})\right] \\ &\approx I(\hat{\theta})(\hat{\theta} - \theta_r)^2 \\ &= I(\hat{\theta})(\hat{\theta} - \theta)^2.\end{aligned}$$

From Theorem 2.9 we know that $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, I(\hat{\theta}^*)^{-1}\right)$, so $\sqrt{n}I(\hat{\theta})^{\frac{1}{2}}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0,1)$. Therefore $nI(\hat{\theta})(\hat{\theta} - \theta^*)^2 \xrightarrow{d} \chi^2_1$. □

This theorem states that the likelihood ratio statistic for two nested models is distributed χ^2 under the "null" hypothesis, which states that the more restricted model is the "correct" model. The degrees of freedom of the null distribution equals the number of parameter restrictions imposed as we go from the general to restricted model. We can therefore calculate the probability of observing a likelihood ratio statistic at least as great as the one we observe for our data set and proposed models. In the CO_2 emissions example, we calculate this p-value as 0.02.

Statistical packages will often report a *model* χ^2. This refers to the χ^2 value for the likelihood ratio comparing the estimated model to the null model – a model with all coefficients except the intercept constrained to equal 0. This statistic answers the usually boring question: "Is the model you fit noticeably better than doing nothing at all?"

A related quantity is model deviance, which compares the model we built to a "perfect" or "saturated" model (i.e., a model with a parameter for every observation). Deviance is given by

$$D(\mathbf{x}) = -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}_b|\mathbf{x}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_s|\mathbf{x})\right],$$

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

where $\hat{\boldsymbol{\theta}}_b$ is the MLE of $\boldsymbol{\theta}_b$, representing the model we built, and $\hat{\boldsymbol{\theta}}_s$ is the MLE of the saturated model. Since this is simply a likelihood ratio, it too follows a χ^2_{n-k} distribution, where n denotes the number of observations in our sample and k denotes the number of parameters in our model.

Statistical packages routinely report "null deviance" and "residual deviance." The former is the deviance of the null model, and the latter is the deviance of the built model. Letting $\log \mathcal{L}(\hat{\boldsymbol{\theta}}_0)$ be the maximized log-likelihood of the null model and noting that

$$\begin{aligned}\text{Residual deviance} - \text{null deviance} &= -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}_b|\mathbf{x}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_s|\mathbf{x})\right] \\ &\quad - \left(-2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\mathbf{x}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_s|\mathbf{x})\right]\right) \\ &= -2\left[\log \mathcal{L}(\hat{\boldsymbol{\theta}}_b|\mathbf{x}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\mathbf{x})\right],\end{aligned}$$

it is easy to see that the difference between null and residual deviances is simply another way of calculating the "model χ^2."

2.3.2 Score and Wald Tests

Score Tests

Score tests appear occasionally, also going under the names *Lagrange multiplier*, *Rao score test*, and the *locally most-powerful test*. The score test relies on the score function to consider the slope of the log-likelihood once we constrain a particular parameter to have some hypothesized value. If the constraint is appropriate, then the slope of the log-likelihood near the constraint should be about 0 since, at the maximum of the likelihood function, the derivative with respect to the parameters equals 0. More formally, the score test compares the slope of the log-likelihood at the hypothesized value, $\boldsymbol{\theta}_0$, against the curvature of the likelihood as described by Fisher information.

Definition 2.8 (Score test)**.** Given a log-likelihood, $\log \mathcal{L}(\theta)$, and a hypothesized value, θ_b, the score test for scalar θ is given as

$$\frac{S(\theta_b)}{\sqrt{\mathcal{I}(\theta_b)}} \dot{\sim} \mathcal{N}(0,1) \qquad \text{if } \theta^* = \theta_b,$$

$$\frac{S(\theta_b)^2}{\mathcal{I}(\theta_b)} \dot{\sim} \chi^2_1 \qquad \text{if } \theta^* = \theta_b.$$

For k-dimensional parameter $\boldsymbol{\theta}$ in which $\boldsymbol{\theta}_b$ imposes $p \leq k$ constraints and holds the other $k-p$ at their MLE, e.g., $\boldsymbol{\theta}_b = (\theta_{1,0}, \ldots, \theta_{p,0}, \hat{\theta}_{p+1}, \ldots, \hat{\theta}_k)$, the score statistic is

$$S(\boldsymbol{\theta}_b)^\intercal \mathcal{I}(\boldsymbol{\theta}_b)^{-1} S(\boldsymbol{\theta}_b) \dot{\sim} \chi^2_p \qquad \text{if } \boldsymbol{\theta}^* = \boldsymbol{\theta}_b.$$

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Wald Tests

We also have the Wald test, which is simply the squared difference of the estimated parameters from some constrained value (typically zero), weighted by the curvature of the log-likelihood.

Definition 2.9 (Wald test)**.** Given a log-likelihood, $\log \mathcal{L}(\theta)$ and hypothesized θ_h, the Wald test for scalar θ is given as

$$W = \frac{(\hat{\theta} - \theta_h)^2}{I(\hat{\theta})}.$$

If $\theta^* = \theta_h$ then $W \dot{\sim} \chi^2_1$.

For k-dimensional parameter $\boldsymbol{\theta}$, in which hypothesized $\boldsymbol{\theta}_h$ imposes $p \leq k$ constraints, the Wald statistic is

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_h)^\intercal I(\hat{\boldsymbol{\theta}})^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_h).$$

Under the maintained hypothesis, $W \dot{\sim} \chi^2_p$.

The Wald test is a generalized version of the standard t-test, which imposes only one restriction. For scalar θ, we have $\hat{\theta} \sim \mathcal{N}(\theta, I(\hat{\theta})^{-1})$, which implies a t-ratio of $\frac{\hat{\theta} - \theta_0}{I(\hat{\theta})^{1/2}}$. The t-ratio is asymptotically standard normal, so the square of this ratio, the Wald statistic, is asymptotically χ^2_1.

Comparing the Tests

Figure 2.1 displays a simplified geometric interpretation of the three test statistics. In this example θ_0 represents the hypothesized or restricted value, sometimes called the null. From here we can see that the likelihood ratio (LR) represents the difference in the value of the log-likelihood function evaluated at its unrestricted maximum and evaluated at θ_0. The Lagrange multiplier (LM) test is the slope of the log-likelihood evaluated at θ_0, while the Wald (W) statistic is the difference between $\hat{\theta}$ and θ_0.

The likelihood ratio, score, and Wald tests all seek to describe how the likelihood changes as we impose restrictions on a relatively narrow class of models. All rely on standard regularity assumptions in order to derive limiting distributions. All three tests are equivalent given a large enough sample.

The relative strengths and weaknesses of these three tools were more important when computing power was scarce and model estimation tedious and computationally expensive. For example, the likelihood ratio test enables the construction of likelihood-based confidence intervals. But the likelihood ratio requires fitting (at least) two different models, whereas the score test and Wald tests have the advantage of requiring the estimation of only one. Score tests are widely used in some time series estimation applications. But in an era of fast computing and cheap data storage, this advantage is not what it once was. Moreover, it is rare that we have one and only one model under consideration.

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

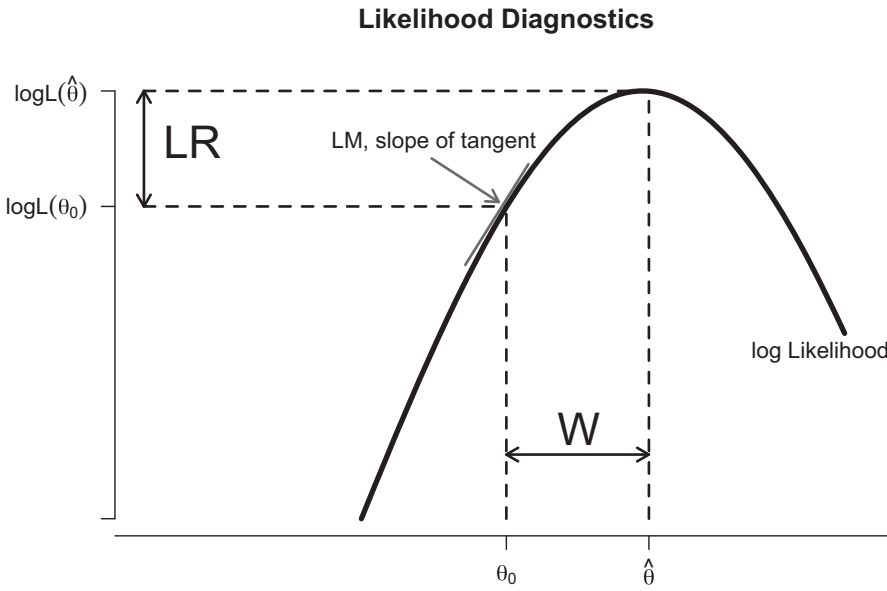

FIGURE 2.1 Geometrical interpretation of the likelihood ratio (LR), Lagrange mulitplier/score (LM), and the Wald (W) test statistics.

Typically we are interested in several working models that we must evaluate and choose among.

In small samples and particular situations, there can be other differences between the tests. The Wald test, in particular, calculates $I(\hat{\boldsymbol{\theta}})$ using the actual, observed MLE rather than assuming that $\boldsymbol{\theta}_0$ represents the correct model. In other words, there are two approximations involved with the Wald test, one for the estimate of the MLE and a second for the approximation of the variance. This can lead to some strange results in small samples. A second drawback for the Wald test is that it can yield different answers under re-parameterizations of $\boldsymbol{\theta}$, something that should not affect the MLE.

All three of these diagnostic tools are firmly rooted in the null hypothesis testing framework in which the analyst proposes a specific value for $\boldsymbol{\theta}$ and then considers the probability of witnessing a value at least as great as those if the null were correct. While this method of reasoning has its virtues it also has drawbacks, not least that the null hypothesis is arbitrary, overly specific, and becomes increasingly likely to be rejected as $n \to \infty$. In evaluating model fit and adequacy we have better tools, some of which also build on the likelihood function.

2.3.3 Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)

The maximized likelihood, $\hat{\mathcal{L}}$, serves as the basis for two more recently developed diagnostic quantities: AIC and BIC.

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

Definition 2.10 (AIC)**.** The Akaike Information Criterion, AIC, is given as

$$\text{AIC} = -2 \log \hat{\mathcal{L}} + 2k,$$

where k is the number of estimated parameters in the fitted model.

The Bayesian Information Criterion, known as BIC or, occasionally, the Schwarz Bayesian Criterion, is another model diagnostic. At the theoretical level, the BIC is an approximation to the Bayes factor – the posterior odds of two models, each conditional on the observed data.

Definition 2.11 (BIC)**.** The BIC is defined as

$$\text{BIC} = -2 \log \hat{\mathcal{L}} + k \log n.$$

For both of these information criteria, smaller values are "better." Just as in the likelihood values on which they are based, neither the AIC nor BIC have meaningful scale. They are useful in comparison only. Since both the AIC and BIC are based on likelihoods, they can be differenced across models *holding observed data constant*. Both the AIC and the BIC penalize models for greater complexity, unlike the LR and other such tests. Holding the maximized likelihood fixed, both criteria prefer simpler models with fewer parameters. Inspecting the two formulas, it is easy to see that that BIC imposes a higher penalty than the AIC so long as $n \geq 8$ or, in other words, almost all the time. Neither of these information criteria has a limiting distribution, so there is no need to specify particular null models or hypotheses. Rather, these criteria are better thought of in the context of selecting among a series of working models.

2.4 WHAT IF THE LIKELIHOOD IS WRONG?

Many of the MLE's desirable properties are predicated on the untestable assumption that our probability model correctly captured some objective, "true" DGP. If the model is wrong, then parameter estimates can be (but are not necessarily) biased and inconsistent. So what, then, is the MLE estimating?

Suppose we specify a probability model $f(x; \boldsymbol{\theta})$, but some other $g(x)$ describes the true DGP. The MLE is then $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_g[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})]$, where the expectation is taken over the unknown distribution, $g(x)$. If the regularity conditions invoked for $f(x)$ also apply to $g(x)$, we have $\mathrm{E}_g[S(\hat{\boldsymbol{\theta}})] = 0$. In other words, the MLE is doing the best it can given the constraints of the probability model assumed.

How good is this approximation? Or, equivalently, how costly is a particular distributional assumption? This is not a question that can be answered in a vacuum, disconnected from plausible competing models and actual data. Rather, this challenge highlights how important it is that we state and critically evaluate modeling assumptions *in a specific application*. In the likelihood framework

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

this is achieved by comparing models using the diagnostics just described, particularly AIC and BIC, which do not rely on formal null hypotheses. An alternative, complementary, and arguably more flexible strategy involves using working models to generate predicted values, the correctness of which can be used to adjudicate among competing sets of assumptions. We take up this strategy in detail in Chapter 5.

2.4.1 Robust Variance and Standard Errors

Under certain circumstances we can develop corrections for some misspecification problems. *If* $\hat{\boldsymbol{\theta}}$ is still consistent for $\boldsymbol{\theta}^*$ under the misspecified model, then we can arrive at the correct variance estimates (asymptotically). Such a variance estimate is said to be "robust" to this particular form of model misspecification.

Formally:

Theorem 2.11. *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *form an i.i.d. sample from unknown* $g(x)$. *Let* $\hat{\boldsymbol{\theta}}$ *be a consistent estimator of* $\boldsymbol{\theta}^*$ *under the assumed probability model,* $f(x; \boldsymbol{\theta}^*)$. *Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1} \boldsymbol{\Omega} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}\right).$$

where

$$\begin{aligned} \boldsymbol{\Omega} &= E_g\left[S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^\intercal\right]\big|_{\boldsymbol{\theta}^*} \\ \mathcal{I}(\boldsymbol{\theta}^*) &= -E_g\left[H(\boldsymbol{\theta})\right]\big|_{\boldsymbol{\theta}^*}. \end{aligned}$$

We say that $\hat{\boldsymbol{\theta}}_n \dot{\sim} \mathcal{N}\left(\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1} \boldsymbol{\Omega} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}\right)$.

Theorems 2.11 and 2.9 look nearly identical and their proofs proceed in similar fashions (i.e., Taylor expansion around the MLE). The key difference is that in Theorem 2.11 the expectation is taken with respect to all possible data, which is governedby unknown distribution or mass function $g(\mathbf{x})$; in Theorem 2.9 we assumed that we had the distributional assumptions correct. If our assumption about the DGP is correct and $g(\mathbf{x}) = f(\mathbf{x})$, then the term $\mathcal{I}(\boldsymbol{\theta}^*)^{-1} \boldsymbol{\Omega} \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$ collapses to the conventional inverse Fisher information, $\mathcal{I}(\boldsymbol{\theta}^*)^{-1}$.

In actual estimation we replace the expectations by their corresponding sample analogues, i.e.,

$$\begin{aligned} \hat{\boldsymbol{\Omega}} &= \frac{1}{n} \sum_{i=1}^{n} S(\hat{\boldsymbol{\theta}} \mid x_i) S(\hat{\boldsymbol{\theta}} \mid x_i)^\intercal \\ I(\hat{\boldsymbol{\theta}}) &= -\frac{1}{n} \sum_{i=1}^{n} H_i(\hat{\boldsymbol{\theta}}) \;\; = \;\; -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} S(\hat{\boldsymbol{\theta}} \mid x_i) \end{aligned}$$

The variance estimator $I(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\Omega}} I(\hat{\boldsymbol{\theta}})^{-1}$ is referred to as a *sandwich variance estimator*, since $\hat{\boldsymbol{\Omega}}$ is sandwiched between the inverse information matrices. The

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

sandwich estimator yields MLE variance estimates that are (asymptotically) correct, even when the specification for the likelihood is incorrect. This seems, at first glance, to be a powerful result (due to Huber (1967)). Indeed, many papers report robust standard errors of various types, relying, at least implicitly, on Huber's result. Social scientists seem justifiably worried that their data contain dependencies that might affect inference. But they also seem willing to believe that their models for the conditional mean are correct. As Freedman (2006) points out, if the model itself is misspecified, then the parameter estimates are likely biased and this problem will not go away with more data. If the model is correct, we have no need for variance corrections in the first place. Having more "correct" standard errors around a biased estimate may be of dubious value.

Robust Standard Errors: Specific

The variance of the OLS estimator is expressed as:

$$\text{var}(\hat{\boldsymbol{\beta}}^{\text{OLS}}) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal)\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

By assuming homoskedasticity we can substitute $\sigma^2\mathbf{I}$ for $\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal)$ and simplify this expression considerably. But what if that assumption is wrong? If the error variance is nonconstant but all our other assumptions are satisfied, then we will continue to have a consistent estimator for $\boldsymbol{\beta}$, but we will end up with incorrect standard errors and possibly erroneous inference. White (1980) proposed using the OLS residuals to estimate $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal)$, specifically by proposing $\mathbf{V}$ where $\text{diag}(\mathbf{V}) = \hat{\varepsilon}_i^2$. Thus the Huber-White heteroskedastic consistent covariance estimator is given by

$$\text{var}_{\text{hc}}(\hat{\boldsymbol{\beta}}^{\text{OLS}}) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{V}\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

It is possible to motivate and derive robust standard errors for the linear regression model without appealing to likelihood theory. But, from a likelihood perspective, it is straightforward to verify that $\mathbf{X}^\intercal\mathbf{X}$ is the Fisher information for the homoskedastic linear-normal regression model. Heteroskedasticity-corrected standard errors are a form of robust sandwich estimators.

2.5 CONCLUSION

We derived the important features of MLEs that we will rely upon in subsequent chapters. First, we highlighted that parametric models require the assumption of a specific probability model. These distributional assumptions must be stated and evaluated. In the subsequent chapters we will present out-of-sample heuristics from which we can make pragmatic judgments about modeling assumptions. Once a model is specified, we showed how the MLE, under general conditions, is asymptotically consistent and normally distributed. The covariance of the asymptotic distribution is given by the negative inverse of

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

the log-likelihood's Hessian matrix, evaluated at the MLE. We also discussed existing tools for evaluating and diagnosing in-sample model fit. These tools include the likelihood ratio, score, and Wald statistics that rely on asymptotic reasoning and null hypothesis testing. But we also have the AIC and BIC that are specifically geared to comparing competing working models without resorting to arbitrary null hypotheses.

The results presented in this chapter assume that the probability model satisfies basic regularity conditions. Although the models we discuss in the remainder of this volume are "regular" likelihoods, those sculpted for specific analysis problems may not be. Or if they are they may still present a challenging likelihood surface for numerical optimization programs. When using such tools we must be careful to establish that the likelihood function is "regular" near the MLE. Graphical tools are usually best here.

2.6 FURTHER READING

Past Work

Aldrich (1997) provides a history of Fisher's development and justifications of maximum likelihood. See Pawitan (2013, ch. 1) and Stigler (2007) for excellent intellectual histories of the development of the theory and method of maximum likelihood in the mid-20th century.

Hirotugu Akaike (1981) notes that "AIC is an acronym for 'an information criterion'," despite it often being denoted Akaike's Information Criterion (Akaike, 1976, 42).

Advanced Study

Edwards (1992) argues for a fully developed likelihood paradigm for all scientific inference, focussing on the complete likelihood function rather than just the MLE. LeCam (1990) collects and summarizes several examples of how the likelihood approach can go awry by way of establishing "Principle 0: Don't trust any principle." See Mayo (2014) with the associated discussion for a critical treatment of the likelihood principle. More general and rigorous proofs of several results in this chapter can be found in standard advanced texts such as Greene (2011) and Cox and Barndorff-Nielsen (1994). General and rigorous treatments of basic probability concepts, the Law of Large Numbers, the Central Limit Theorem, and other elementary topics can be found in Resnick (2014).

The sandwich variance estimator has been extended in numerous ways, particularly to data that are "clustered" in space or time. See, for example, Cameron et al. (2008); Cameron and Miller (2015). King and Roberts (2014) argue that the divergence between robust and conventional standard errors can

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press

be used as a diagnostic tool for model misspecification, but see also Aronow (2016).

Software Notes

Several $\mathcal{R}$ libraries implement various robust and clustered standard error estimators for a variety of different models and data structures. These include `clusterSEs` (Esarey, 2017), `pcse` (Bailey and Katz, 2011), `plm` (Croissant and Millo, 2008), `rms` (Harrell, Jr., 2017), and `sandwich` (Zeileis, 2004).

https://doi.org/10.1017/9781316888544.005 Published online by Cambridge University Press