ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

An Introduction to Maximum Likelihood

Semester 1, 2025

Version: April 25, 2025

Contents

1 Maximum likelihood: The Single Parameter Case

1.1 Estimation

We are concerned with explaining the observed behaviour of some variable y. In what follows we will assume that we have n observations on y and we will distinguish between these observations through use of a subscript i which will range over the values 1, 2, ..., n. If we wish to talk about our entire sample we will write y_i, $i = 1, \ldots, n$. Equally, if we wish to simply talk about an arbitrary observation then we will generically refer to y_i.

Now, the way we typically go about constructing a model stems from the trivial identity

$$y_i = f_i + u_i, \qquad i = 1, \ldots, n, \tag{1}$$

where f_i is some quantity to be discussed shortly and u_i will be dubbed a *disturbance term*, which is implicitly defined to make (1) true; that is, $u_i \equiv y_i - f_i$. For example, suppose that $f_i = 10$ then we can always decompose a set of observations according to

$$y_i = 10 + u_i, \qquad i = 1, \ldots, n.$$

where u_i is simply the difference between the observed value of y_i and the number 10. This is a silly example but it illustrates a very important point, namely that disturbance terms are conditional things which are devoid of meaning until we have defined f_i. Continuing the silly example, if we had defined f_i to be 20 instead of 10 then there would still be disturbances but now they would take values $u_i = y_i - 20$ rather than $u_i = y_i - 10$.

So what about f_i? In what follows we shall think of f_i as our model of y_i. In the example of the previous paragraph f_i was a constant, either 10 or 20, by which I mean that it did not vary with the index i. Constants tend not to be very interesting models although they are appropriate if you are trying to model the unconditional mean (expectation) of some random variable. For example, suppose that your observations are a simple random sample from a population of the form $y \sim \mathrm{N}(\mu, \sigma^2)$. Then we can write each of the observations as

$$y_i = \mu + u_i, \qquad i = 1, \ldots, n.$$

The modelling problem is then to estimate μ.

As a first year student you will have learned that an appropriate estimator for μ is[1]

$$\hat{\mu}_n = \bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

[1] Our first exposure to summation notation. If summation notation causes you any problems then you need to make sure that you are on top of it as soon as possible.

There the estimator is motivated as a least squares estimator; that is[2]

$$\hat{\mu}_n = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{2}$$

This equation contains another important distinction. We have defined the quantity $y_i - \mu$ to be a disturbance term. The quantity $e_i = y_i - \bar{y}_n$ is called a *residual*, it is the difference between an observation and an estimated or 'fitted' model. And so the least squares estimator is that value which minimizes the sum of squared residuals.

In this course we will frequently be interested in maximum likelihood estimators, or mles for short. In order to construct an mle we need to know about the joint probability distribution of our sample of observations $y_1, \ldots, y_n$. If, as we assumed earlier, our observations constitute a simple random sample from a Normal population with constant mean μ and variance σ^2,[3] then the joint density function for the sample, denoted $f(y_1, \ldots, y_n; \mu, \sigma^2)$, is the product of the density functions for each of the y_i, denoted $f_i(y_i; \mu, \sigma^2)$; that is[4]

$$f(y_1, \ldots, y_n; \mu, \sigma^2) = \prod_{i=1}^{n} f_i(y_i; \mu, \sigma^2).$$

Substituting for the Normal density function we have[5]

$$\begin{aligned} f(y_1, \ldots, y_n; \mu, \sigma^2) &= \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}. \end{aligned} \tag{3}$$

When we think of $f(y_1, \ldots, y_n; \mu, \sigma^2)$ as a function of $y_1, \ldots, y_n$ it is a density function with parameters μ and σ^2. If, instead, we think of it as a function of the parameters for a given data set then it is called a *likelihood function* and we will write it as $\mathcal{L}_n(\mu, \sigma^2; y_1, \ldots, y_n)$, which we will variously abbreviate as $\mathcal{L}_n(\mu, \sigma^2)$ or simply $\mathcal{L}$ for short if all else is clear from the context. In particular, the likelihood function here is

$$\mathcal{L} \equiv \mathcal{L}_n(\mu, \sigma^2) \equiv \mathcal{L}_n(\mu, \sigma^2; y_1, \ldots, y_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$

Maximum likelihood estimators are then those functions of the data that maximize the likelihood function. In this case, the mles are those values $\hat{\mu}_n$ and $\hat{\sigma}^2_n$ satisfying[6]

$$\underset{\mu, \sigma^2}{\operatorname{argmax}} \, \mathcal{L}_n(\mu, \sigma^2; y_1, \ldots, y_n).$$

[2] The expression $\operatorname{argmin}_x g(x)$ means find that value of x which minimizes the function $g(x)$.

[3] By a simple random sample we mean that the observations are independent of one another and identically distributed, typically indicated by saying that the observations are *iid*.

[4] The fact that the y_i are identically distributed means that we could omit the subscript on the density functions because it is a constant function, namely $f_i(y_i; \mu, \sigma^2) = f(y_i; \mu, \sigma^2)$. This is not always the case, however, and you should be aware that these functions may differ from observation to observation. As we are unlikely to encounter such complicated circumstances we will tend to omit the subscript as indicated above.

[5] More notation! The symbol $\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$.

[6] $\operatorname{argmax}_x g(x)$ is just like $\operatorname{argmin}_x g(x)$ except that the former is seeking that value of x which maximizes $g(x)$.

There are two important points that arise from this expression. First, in contrast to the least squares approach, maximum likelihood attempts to *simultaneously* provide estimators for all of the parameters of the model. Least squares does not provide a natural estimator for σ^2 and use of $s_n^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$ is motivated by its unbiasedness and because its use in t-statistics yields a t distribution, but not because of the least squares approach. Second, the maximization problem is often easier (in this case) if we take natural logarithms of both sides. This is justified because the natural logarithm is a monotonic function of its argument — meaning that if $x_1 > x_2$ then $\ln x_1 > \ln x_2$ and *vice verse* — and so values of $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ that maximize $\ln \mathcal{L}$ will also maximize $\mathcal{L}$. The quantity $\ln \mathcal{L}$ is called the *log-likelihood function* and will be used extensively throughout what follows.

There are two other quantities of which you need to be aware. First, a common approach to maximizing the likelihood function is to differentiate it with respect to the parameters and solve the resulting first-order conditions obtained when the derivative is set equal to zero. This solution may yield any of a maximum, a minimum, or a saddle-point (also called a point of inflection when there is only one parameter) which is neither a global maximum nor a global minimum but is rather a flat spot in the likelihood function. To determine what we have found we can check the second-order conditions; in the case of one parameter we want the second derivative of the log-likelihood to be negative. These terms all have special names and purposes. The derivative of the log-likelihood function is called the *efficient score* or simply the *score* function. We shall denote it by $\mathcal{S}_n(\theta)$. Where the likelihood function is a function of a single parameter, θ say, then the score is a scalar function. For example, for a given log-likelihood function $\ln \mathcal{L}_n(\theta)$ the score is simply[7]

$$\mathcal{S}_n(\theta; y) = \frac{\mathrm{d} \ln \mathcal{L}_n(\theta; y)}{\mathrm{d}\theta} = \sum_{i=1}^n \frac{\mathrm{d} \ln \mathcal{L}_n(\theta; y_i)}{\mathrm{d}\theta} = \sum_{i=1}^n \mathcal{S}_n(\theta; y_i).$$

In particular, continuing our earlier example, suppose that we know the variance σ^2 but need to estimate the mean of a Normal population. Here the log-likelihood function is

$$\ln \mathcal{L}_n(\mu) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2$$

and the score reduces to

$$\mathcal{S}_n(\mu) = \frac{d \ln \mathcal{L}_n(\mu)}{d\mu} = \frac{d}{d\mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right] = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu),$$

because the derivative of the other (constant) term with respect to μ is zero. The corresponding first-order condition of $\mathcal{S}_n(\theta) = 0$ simplifies to

$$\sum_{i=1}^n (y_i - \hat{\mu}_n) = 0.$$

Observe that μ now has a $\hat{}$ on it to indicate that it is that value of μ at which the first-order condition holds.[8] At this point you might recognize that this first-order condition

[7] In the following expression, and frequently hereafter, we use a vector (y in this instance) to denote the set of observations $y_1, \ldots, y_n$.

[8] Observe that this solution need not be unique, although in this case it is.

is exactly the same as that for the least squares problem (2), and so the the mle for μ is $\hat{\mu}_n = \bar{y}_n$.

One final point to note about the score function is that, for a correctly specified model, it has zero expectation. In our example we see that

$$\mathrm{E}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathrm{E}\left[y_i\right]-\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu-\mu) = 0.$$

It is reasonably easy to show that this result holds more generally (see for example Appendix B.1). If you are interested in learning more of the theory of maximum likelihood methods then good places to look include the elegant text by Silvey (1970) and also the book by Cramer (1986).

The second-order condition in the single parameter case is that the second derivative of the log-likelihood function, denoted by $\mathcal{H}_n(\theta)$, be negative. Continuing our example, we have

$$\mathcal{H}_n(\mu) = \frac{d^2 \ln \mathcal{L}_n(\mu)}{d\mu^2} = \frac{d}{d\mu}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\right] = -\frac{n}{\sigma^2} < 0,$$

because $n > 0$ is simply the number of observations in the sample and $\sigma^2 > 0$ by assumption.

The second derivative of $\ln \mathcal{L}_n(\mu)$ possesses a number of important properties. By definition

$$\mathcal{I}_n(\theta) = -\mathrm{E}\left[\mathcal{H}_n(\theta)\right] = -\sum_{i=1}^{n}\mathrm{E}\left[\frac{\mathrm{d}^2 \ln \mathcal{L}_n(\theta; y_i)}{\mathrm{d}\theta^2}\right] = \sum_{i=1}^{n}\imath_{\theta,i},$$

where $\imath_{\theta,i} = -\mathrm{E}_{y_i}\left[\frac{\mathrm{d}^2 \ln \mathcal{L}_n(\theta;y_i)}{\mathrm{d}\theta^2}\right]$. Whereas $\mathcal{I}_n(\theta)$ is a measure of the information in a sample of size n, $\imath_{\theta,i}$ is a measure of the information in the i-th observation. In particular, both these quantities are referred to as Fisher's information. In the event that our data are identically distributed, $\imath_{\theta,1} = \cdots = \imath_{\theta,n} = \imath_\theta$, say.[9]

As a first observation, given that

$$\mathrm{E}\left[\mathcal{S}_n(\theta; y_i)\right] = 0,$$

we see that $\mathcal{I}_n(\theta) = \mathrm{Var}\left[\mathcal{S}_n(\theta)\right]$. Second, letting $\tilde{\theta}_n$ denote any unbiased estimator for θ, it can be shown that

$$\mathrm{Var}\left[\tilde{\theta}_n\right] \geq (\mathcal{I}_n(\theta))^{-1}. \tag{4}$$

This result is known as the Cramér-Rao inequality and it provides a lower bound for the variance of any unbiased estimator. The quantity $(\mathcal{I}_n(\theta))^{-1}$ is called the Cramér-Rao lower bound (CRLB).[10] Third, it can also be shown that there exists an unbiased estimator which attains the CRLB if and only if

$$\mathcal{S}_n(\theta) = \mathcal{I}_n(\theta)(\tilde{\theta}_n - \theta).$$

[9] Instead of abstracting away from an index for the observation, as has been done here with $\imath_\theta$, an alternative is to use the symbol for the information in an arbitrarily chosen observation, say the first, which would then see something like $\imath_{\theta,1}$ used to denote the information in a single observation.

[10] At a superficial level the CRLB is something of a maximum likelihood equivalent to the Gauss-Markov theorem that you have met in the context of least squares estimators.

Continuing our example, $\mathrm{E}[\bar{y}_n] = \mu$ and so $\bar{y}_n$ is unbiased for μ. Similarly, the variance of $\bar{y}_n$ is σ^2/n.[11] But is there a more efficient unbiased estimator for μ? As an unbiased estimator we know that its variance is bounded from below by the CRLB. We have already established that

$$\mathcal{H}_n(\theta) = -\frac{n}{\sigma^2}.$$

Clearly, $(\mathcal{I}_n(\theta))^{-1} = (-\mathrm{E}[-n/\sigma^2])^{-1} = \sigma^2/n$, and so the variance of $\bar{y}_n$ attains the CRLB meaning that, even if there is another unbiased estimator for μ that is equally efficient, there is no unbiased estimator that is more efficient. As an aside, for this example, we see that the score is

$$\mathcal{S}_n(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = \frac{n}{\sigma^2}(\bar{y}_n - \mu) = \mathcal{I}_n(\mu) \times (\hat{\mu}_n - \mu) \tag{5}$$

as required to attain the CRLB.

It is not the case that mles always attain the CRLB, nor for that matter are they even always unbiased. However, if certain regularity conditions are met (typically that certain derivatives exist and behave in standard ways), mles have certain desirable properties that make them the benchmark against which all other estimators are compared. A subset of these properties include:

1. It can be shown that they are consistent, usually written as either

   $$\operatorname{plim}\hat{\theta} = \theta \quad \text{or} \quad \hat{\theta} \underset{p}{\to} \theta.$$

   This means that as the sample size grows the sampling distribution of $\hat{\theta}$ collapses down to a single point at θ and so, in the limiting case of an infinitely big sample, the probability of observing a value for $\hat{\theta}$ other than θ is zero.

2. Asymptotically they attain the CRLB, i.e.,

   $$\lim_{n\to\infty} \mathrm{Var}\left[\sqrt{n}(\hat{\theta} - \theta)\right] = \lim_{n\to\infty}(n^{-1}\mathcal{I}_n(\theta))^{-1} = \lim_{n\to\infty} \imath_\theta^{-1},$$

   where $\imath_\theta = n^{-1}\mathcal{I}_n(\theta)$. Our previous property essentially said that as the sample size gets big the variance of $\hat{\theta}$ tends to zero. For this reason we have to scale the variable to get something that doesn't have a limiting variance of zero. We subtract θ so that our scaling doesn't cause the mean to become infinitely big in the limit. That is, $\mathrm{E}\left[\hat{\theta} - \theta)\right] = 0$ and no amount of scaling will change that, whereas, if $\mathrm{E}\left[\hat{\theta} - \theta)\right] = c$ then $\sqrt{n}\,\mathrm{E}\left[\hat{\theta} - \theta)\right] = \sqrt{n}\,c \to \infty$ as $n \to \infty$, for any non-zero c.

3. For sufficiently large samples their distribution is approximately Normal:

   $$\hat{\theta} \underset{a}{\sim} N(\theta, (\mathcal{I}_n(\theta))^{-1}).$$

   This is just like the central limit theorem that you should have seen back in first year (recall that for sufficiently big samples sample means have sampling distributions that are approximately Normal, even if the population isn't).

[11] In case you have forgotten how to do it, these results are proved in Appendix A.

4. They possess a useful invariance property. In particular, if $\hat{\theta}$ is the mle for θ then $h(\hat{\theta})$ is the mle for $h(\theta)$. This result is particularly useful in non-linear models where it might be quite hard to estimate a parameter of interest but is relatively easy to estimate functions of it.

Remark 1 ($\imath_\theta$ and $\mathcal{I}_n(\theta)$). There are two related notions of information here. $\mathcal{I}_n(\theta)$ is the total information in the sample and $\imath_\theta$ is the average information per observation in the sample. In an iid environment the average information per observation in the sample is the same as the information in any single observation.

1.2 Testing

1.2.1 Fundamental Concepts

Hypothesis testing is one of those topics for which it is easy to lose sight of the forest for the trees, so let us begin by recalling some fundamental concepts. First, the aim of the exercise is to choose between two competing hypotheses — the null hypothesis, denoted H_0, and the alternative hypothesis, denoted H_1, respectively — on the basis of sample data. The process by which this choice is made, namely the decision rule, is then an hypothesis test. There are two possible choices one can make; namely, either choose to accept H_0 or choose not to accept H_0.[12] Equally there are two possible states of nature; either H_0 is true and H_1 is false or H_1 is true and H_0 is false. This leads to four possible pairs of choice and state of nature which are presented in Table 1.

Table 1: Hypothesis Testing Outcomes

	States of Nature	
Choice	H_0 true	H_1 true
Accept H_0	Correct Choice	Type II Error
Reject H_0	Type I Error	Correct Choice

Clearly, if you either (i) accept H_0 when it is true, or (ii) reject H_0 when it is not true, then you have made a correct decision. If you reject H_0 when it is true then you have made an error, a so-called *Type I error*. Similarly, if you accept H_0 when H_0 is not true then again you have made an error, a *Type II error* in this case.

The probability of making a Type I error, that is the probability of rejecting H_0 given that H_0 is true, is called either the *size of the test* or the *level of significance of the test* (or, more simply, the *level of the test*) and is typically denoted by the symbol α. The probability of making a Type II error is typically denoted by the symbol β. In an ideal world we would never make errors of judgement and so we would like both $\alpha = 0$ and $\beta = 0$. Sadly there is no interesting problem where this happens.

Before moving on, let us say a little more about the nature of null and alternative hypotheses. In a classical setting, competing hypotheses specify mutually exclusive and exhaustive subsets of the parameter space. This means that no parameter value can simultaneously satisfy both the null and the alternative hypotheses and that every element of the parameter space must belong to one or the other of the competing hypotheses. Consider a parameter space Θ. If the set defining an hypothesis contains only a single

[12] Strictly, we never accept or reject hypotheses. Rather, we either conclude that the sample data supports one particular hypothesis against the other or that it does not.

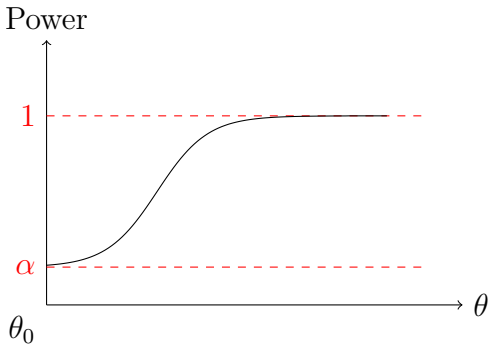

Figure 1: Power Function of a Test of a Simple Null Hypothesis Against a One-Sided Alternative

element, e.g. $\theta = \theta_0 \in \Theta$, then it is said to be a *simple hypothesis*. If the set contains more than one element, e.g. $\theta \in \Theta \setminus \theta_0$,[13] then it is said to be a *composite hypothesis*. While it is clear what is meant by accepting a simple hypothesis, accepting a composite hypothesis is a far more nuanced concept. What exactly are you accepting in such a case? You are certainly not accepting a proposition that the true parameter is a specific value. Rather accepting a composite hypothesis merely says that the data are favouring a proposition that the true parameter value is a member of a particular subset of the parameter space without providing any indication of where in that subset it might be. Why is this important? Null hypotheses are typically simple hypotheses whereas alternative hypotheses are often composite hypotheses. Testing procedures are very asymmetric about what they can say about the competing hypotheses in the case of composite hypotheses. The asymmetry in treatment of a simple null hypothesis against a composite alternative hypotheses is a pervasive characteristic of classical hypothesis testing.

Now is as good a time as any to introduce the concept of the *power of a test*. In short, if the probability of incorrectly accepting a false null hypothesis — that is, making a Type II error — is denoted by β, then the power of a test is given by the quantity $1-\beta$. That is, the power of a test is the probability of correctly rejecting a false null hypothesis. Equivalently, it is the probability of correctly accepting the alternative hypothesis when it is true. As it is the probability of making a correct decision we clearly would like our test procedures to have good power. As it is a probabilty, power is bounded to lie in the interval $[0, 1]$. Clearly the greater the power of a test the better. Recall that it was previously asserted that alternative hypotheses are often composite hypotheses. In such cases there is a (possibly infinite) set of possible true parameter values, even though only one parameter value can be the true one, and the actual power of the test will likely vary depending upon what the true parameter value actually is. The locus of all such powers is known as the *power function* of the test. Consider the problem of testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta > \theta_0$, where $\Theta = \{\theta : \theta \geq \theta_0\}$. We might expect the power function of any reasonable test of these hypotheses to have a power function like that depicted in Figure 1.

A couple of features of Figure 1 merit comment. First, when the null hypothesis is

[13]The notation $\theta \in \Theta \setminus \theta_0$ means "θ belongs to that subset of Θ containing all of the elements of Θ except for θ_0, which is specifically excluded". One reason for adopting this sort of notation is that it makes clear that the null hypothesis can be thought of as imposing a restriction on the parameter space, a restriction that reduces its dimension. In the case of a simple hypothesis it may be reducing a set of infinite dimension to a singleton.

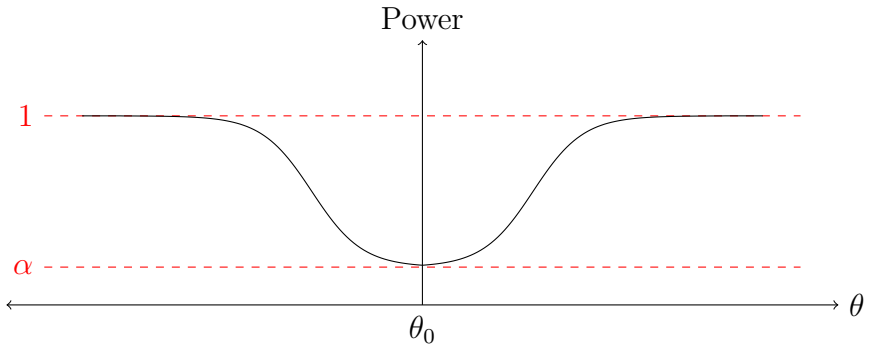

Figure 2: Power Function of a Test of a Simple Null Hypothesis Against a Two-Sided Alternative

true, to reject the null hypothesis is to make a Type I error, the probability of which is α. Hence, at $\theta = \theta_0$, the power function takes the value α. Second, being a probability, the power can never exceed unity and approaches the line Power = 1 from below. The further the true value of θ departs from θ_0 the closer to unity we would expect the power of the test to get. We might reasonably expect this to happen monotonically. If the power function of a test ever dips below the line Power = α then we say that the test is a *biased test*. Otherwise the test is said to be *unbiased*. If we have multiple tests of a given pair of hypotheses then we might expect their power functions to cross at different values of θ, as different tests tend to perform better (have greater power) in different parts of the parameter space. If one of the power functions is always at least as great as any other power function, and greater than any other power function for at least one value of θ, then we say that that test is the *uniformly most powerful* (UMP) test in θ. Typically there is no UMP test, although there may exist a UMP tests within classes of tests defined by the addition of various additional requirements that the tests must satisfy. (This is very much a topic for another time.)

As a second example of a power function, consider Figure 2 which presents a stylised power function for a test of $H_0 : \theta = \theta_0$ against the two-sided alternative $H_1 : \theta \neq \theta_0$. Note the classical urn shape for the power function that one might expect well-behaved tests to exhibit in such cases.[14] All the comments that were made above for the one-sided case apply equally here.

With the preceding concepts defined, a classical approach to hypothesis testing proceeds as follows. Define a test statistic, T say, which is a function of sample data; that is, $T \equiv T(y_1, \ldots, y_n)$. Now, because the sample data is random, T is a random variable. An hypothesis test takes all the possible values that T can take and it allocates them to one, and only one, of two sets, so that the two sets are disjoint. One set will contain all of those values of T which, if observed, would be taken as evidence in favour of H_0. This set is an *acceptance set*, often called the *acceptance region*, of the test. The other set contains all of those possible values of T that would be taken as evidence against H_0 if they were observed. This set is typically called either the *rejection region* or the *critical region* of the test.[15] Thus, the probability of an observed value for T belonging to the critical region when H_0 is true is equal to α, whereas the probability of this event

[14] An interesting counter-example is discussed in Poskitt and Skeels (2008).

[15] Observe the asymmetric nature of classical testing, where elements of the critical region need not form evidence in favour of H_1, all that is required is that elements of the critical region be deemed evidence against H_0 being true.

occurring if H_1 is true is the *power of the test*. Clearly we can make the size of a test small by decreasing the size of the rejection region, thereby making the acceptance region bigger, but this will also reduce the power of the test which we would prefer to be as close to unity as possible. Hence, there is a trade-off between the size and power of test, and much of the practice of hypothesis testing is concerned with managing this trade-off.

Now, the important part of a test is not the statistic itself but rather the partitioning of its sample space into acceptance and rejection regions; that is, the test is the decision rule. For example, suppose that we define a new statistic $S = s(T)$, for some function $s(\cdot)$, and we define a partitioning of the sample space of S into acceptance and rejection regions. If, for every possible sample $y_1, \ldots, y_n$, S and T yield identical outcomes in terms of acceptance and rejection of H_0 then the two partitionings (decision rules) constitute the same test.

To make this example more concrete, consider the problem of testing $H_0 : \mu = 3$ against a two-sided alternative of the form $H_0 : \mu \neq 3$ on the basis of a simple random sample from a Normal population with unit variance. A sensible statistic on which to base a test is the usual z statistic, $z = \sqrt{n}\,(\bar{y}_n - 3)$ when $\sigma^2 = 1$, and the test is to reject H_0 for large values of $|z|$, i.e. reject H_0 for $|z| > z_{\alpha/2}$, where $z_{\alpha/2}$ denotes the critical value which cuts off an upper-tail probability of $\alpha/2$ in a standard Normal distribution. Alternatively, we might equally use $q = z^2$ as the test statistic, and then the decision rule is to reject for H_0 for $q > \chi^2_{1,\alpha}$, where $\chi^2_{1,\alpha}$ denotes the critical value which cuts off an upper-tail probability of α in a chi-squared distribution with one degree of freedom. In particular, for $\alpha = 0.05$, $z_{\alpha/2} = 1.96$ and, not surprisingly, $\chi^2_{1,\alpha} = 3.8416 = (1.96)^2$. So different statistics result in different decision rules but the tests are the same. Any sample which results in z rejecting H_0 will also result in q rejecting H_0, and vice versa.

The previous example may give you the impression that any transformation of T might be acceptable, however, this is not the case. As we shall see acceptable transformations are determined by H_1. To see this consider a minor modification to the previous example where now $H_1 : \mu > 3$. An appropriate test statistic is still z but now the decision rule is to reject H_0 for $z > z_\alpha$. If we attempt to use q as the test statistic there is no decision rule that we can choose that will yield the same test. The reason for this is that the transformation from z to q loses information on the sign of z which is necessary in deciding whether a particular sample provides evidence in favour of H_0 or not. For instance, whereas previously both $z = -2.5$ and $z = 2.5$ would have been interpreted as evidence against H_0 at the 5% level (say), in this example only $z = 2.5$ is evidence against H_0. However, if q is used as the statistic we are unable to distinguish between a sample that yields $z = -2.5$ and one that yields $z = 2.5$, both will yield $q = (2.5)^2 = 6.25$ and so both will be interpreted as providing evidence against H_0. This illustrates that the two tests can provide conflicting results for a given sample and so they must be different tests.

In what follows we will introduce the *holy trinity* of tests: the likelihood ratio test, the Lagrange Multiplier (LM) test and the Wald test. These three tests adopt different approaches to testing a given null hypothesis against an alternative. The basic ideas underlying these three approaches form the basis of much of the theory of hypothesis testing.

1.2.2 The Likelihood Ratio (LR) Test

We shall begin by thinking of the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ on the basis of a simple random sample $y_1, \ldots, y_n$, from a population which can be described by a probability distribution $f(y_1, \ldots, y_n; \theta)$, where θ is an unknown parameter. Continuing an earlier example, let us suppose that our sample comes from a Normal population with unknown mean μ and known variance σ^2. Further suppose that we wish to test $H_0 : \mu = 10$ against $H_1 : \mu = 20$. This is an example of testing a simple null hypothesis against a simple alternative.

Recall that a likelihood function is simply a probability density function interpreted as a function of the parameters, θ say, rather than as a function of the data. For a given data set you can evaluate the likelihood function at any parameter value that you wish to. If you have to choose between two possible parameter values, one basis for choice would be to choose that parameter for which the observed sample has the highest probability of occurrence. This is pretty much what a Likelihood Ratio test does.[16] Let us make these ideas more precise.

If H_0 is true then the likelihood function is $\mathcal{L}_n(\theta_0; y_1, \ldots, y_n)$. Similarly, if H_1 is true then the likelihood function is $\mathcal{L}_n(\theta_1; y_1, \ldots, y_n)$. The Likelihood Ratio test is then

$$\begin{aligned} &\text{Reject } H_0 \text{ if } \lambda \leq \kappa_\alpha, \\ &\text{Accept } H_0 \text{ if } \lambda > \kappa_\alpha, \end{aligned} \tag{6}$$

where

$$\lambda \equiv \lambda(y_1, \ldots, y_n) = \frac{\mathcal{L}_n(\theta_0; y_1, \ldots, y_n)}{\mathcal{L}_n(\theta_1; y_1, \ldots, y_n)},$$

and κ_α is some constant — that is, not a function of sample data — chosen so that the size of the test is α, i.e. $P(\lambda \leq \kappa_\alpha | H_0 \text{ true}) = \alpha$.

Note that λ is not defined if $\mathcal{L}_n(\theta_1; y_1, \ldots, y_n) = 0$. But this probability being zero means that there is no chance of the alternative hypothesis being true, in which case there is no testing problem, making discussion of this case uninteresting. So we need not concern ourselves with this moving forward. As both the numerator and denominator are probabilities, i.e. non-negative, with the denominator greater than zero, it follows that $\lambda > 0$. Intuitively, however, the hypothesis that the data supports is going to have the larger probability. This suggests that we might expect $\lambda > 1$ if the data supports the null hypothesis and $\lambda < 1$ if the data supports the alternative, with $\lambda = 1$ unhelpful, although also a zero probability event (why?). Such reasoning is helpful but not necessarily all that precise, so let's make these ideas more concrete.

Consider again the problem described at the start of this section, although we shall replace 10 by μ_0 and 20 by μ_1, so that we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$. We see that

$$\begin{aligned} \lambda &= \frac{(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^n (y_i - \mu_0)^2/(2\sigma^2)\}}{(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^n (y_i - \mu_1)^2/(2\sigma^2)\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \left[(y_i - \mu_0)^2 - (y_i - \mu_1)^2\right]\right\}. \end{aligned}$$

[16] Note the contrast between estimation and testing. Estimation involves choosing the mles to maximize the probability of observing the given sample data that you have; the value of the maximized probability is not really important. Testing, however, focuses on the probability of observing the sample data *given parameter values*.

Next, we can write

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\mu_1)^2 &= \sum_{i=1}^{n}[(y_i-\mu_0)-(\mu_1-\mu_0)]^2\\ &= \sum_{i=1}^{n}(y_i-\mu_0)^2 - 2(\mu_1-\mu_0)\sum_{i=1}^{n}(y_i-\mu_0) + n(\mu_1-\mu_0)^2\\ &= \sum_{i=1}^{n}(y_i-\mu_0)^2 - 2n(\mu_1-\mu_0)(\bar{y}_n-\mu_0) + n(\mu_1-\mu_0)^2, \qquad (7)\end{aligned}$$

so that

$$\lambda = \exp\left\{-\frac{n}{2\sigma^2}\left[2(\mu_1-\mu_0)(\bar{y}_n-\mu_0)-(\mu_1-\mu_0)^2\right]\right\} \geq 0.$$

The problem here is to find a constant κ_α which yields a test of the correct size. Frankly, in its current form, this problem is pretty nasty. Our life will be made easier if we apply some transformations. To begin,

$$\begin{aligned}P(\lambda \leq \kappa) &= P\left(-2\ln\lambda \geq -2\ln\kappa\right)\\ &= P\left(\frac{n}{\sigma^2}\left[2(\mu_1-\mu_0)(\bar{y}_n-\mu_0)-(\mu_1-\mu_0)^2\right] \geq -2\ln\kappa\right)\\ &= \begin{cases} P\left(z \geq \kappa'\right), & \text{if } \mu_1 > \mu_0,\\ P\left(z \leq \kappa'\right), & \text{if } \mu_1 < \mu_0,\end{cases}\end{aligned}$$

where $z = \sqrt{n}\,(\bar{y}_n-\mu_0)/\sigma$ and

$$\kappa' = \frac{n(\mu_1-\mu_0)^2 - 2\sigma^2\ln\kappa}{2\sigma\sqrt{n}\,(\mu_1-\mu_0)}. \qquad (8)$$

These probability statements are true under both the null and alternative hypotheses.

In order to choose a critical value we need to find that value of κ, κ_α say, that cuts of a size α rejection region. That is, κ_α satisfies

$$P(\lambda \leq \kappa_\alpha \mid H_0 \text{ true}) = \alpha.$$

We don't actually want κ_α so much as we want κ'_α but we note that

$$\kappa'_\alpha = \frac{n(\mu_1-\mu_0)^2 - 2\sigma^2\ln\kappa_\alpha}{2\sigma\sqrt{n}\,(\mu_1-\mu_0)}.$$

We see that we have two situations to consider, one where $\mu_1-\mu_0 > 0$ and the other where $\mu_1-\mu_0 < 0$. Inspection of (8) can tell us a little more about what to expect. Consider first the case where $\mu_1-\mu_0 > 0$. Clearly, the denominator of (8) is positive in this case, as is the first term in the numerator. In thinking about the behaviour of the test statistic earlier we suggested that evidence against the null hypothesis was likely to generate values $0 < \lambda < 1$ and so we might also expect $0 < \kappa_\alpha < 1$. Consequently, we might expect $\ln\kappa_\alpha < 0 \implies \kappa'_\alpha > 0$. That is, if $\mu_1-\mu_0 > 0$ then our test should supports the alternative hypothesis against the null if $z \geq \kappa'_\alpha > 0$. Another way of saying this is that sufficiently large values of z constitute evidence against the null hypothesis in favour of this alternative.

Conversely, what if $\mu_1 < \mu_0 < 0$? Our reasoning about κ_α remains completely unchanged as so the numerator of (8) is positive. However, now the denominator of (8) is negative and so $\kappa'_\alpha < 0$That is, we cannot accept the null against the alternative if z is sufficiently negative.[17] The only remaining issue is what constitutes large values of z? If H_0 is true then $z \sim N(0,1)$. Therefore, we can obtain appropriate value for κ'_α using critical values from the standard Normal distribution, denoted z_α. On rearrangement of (8) we see that

$$\kappa_\alpha = \exp\left\{\left[2\sqrt{n}\,\sigma(\mu_1 - \mu_0)z_\alpha - n(\mu_1 - \mu_0)^2)\right]/2\sigma^2\right\}.$$

The important point to notice about this value for κ_α is that, although a little bit complicated, it only depends upon known constants.

The case of testing a simple null hypothesis against a simple alternative is not one that occurs frequently, although it does illustrate most aspects of the Likelihood Ratio test that you need to know. Further, it allows us to talk about a remarkable result known as the Neyman-Pearson lemma, which is the strongest optimality result in the statistical theory of hypothesis testing.[18] In short it says that, when testing a simple null against a simple alternative, the Likelihood Ratio test is uniformly most powerful amongst all tests of a fixed size α. That is, if you restrict attention to tests that have a size of α then the Likelihood Ratio test will have at least as much power as any of the other tests that you are considering *for all* of the possible values that the parameter of interest can take under H_1. In this snse the test is said to be *point optimal*.[19] It may not seem like much of a claim but it is actually a pretty strong statement, there are very few situations where such a claim can be made. This result is the theoretical basis for the popularity of the Likelihood Ratio test in more general situations.

Let us now think about a more general situation. Most of the time we will be interested in testing a simple null hypothesis against a composite alternative, where there is more than one possible value that θ might take if H_1 is true. For example, we might be interested in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, but one-sided alternatives, such as $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$, are also common. The Likelihood Ratio test, as defined, can't help us here because, when H_0 is not true, H_1 doesn't actually specify the true value of θ, it merely gives a set of possible values. The usual response to this problem is to settle on a particular value for θ that in some (useful) way characterizes H_1, which then allows us to appeal to the results of the previous case.[20]

[17] Hopefully you recognize this test as the usual z statistic used test for testing hypotheses on the mean of a random variable whose distribution is Normal against a one-sided alternative that you will have been taught about in first year, and probably at secondary school as well. Note that a one-sided test is appropriate here because in this problem we know both μ_0 and μ_1 as they are specified by the hypotheses.

[18] The original paper by Neyman and Pearson (1933) is beautifully written and not as technically intimidating as are many papers; the interested are certainly encouraged to read it. For those of you doing mathematical economics, their approach is allegedly the first use of the calculus of variations in a non-standard problem.

[19] By considering a collection of point optimal tests, where the parameter value varies, one is able to may out a *power envelope*, which defines the maximum power attainable by any test of a given size. Clearly, one would like to use a test whose power function lies as close as possible to the power envelope over the greatest range of parameter values. For more on point optimal tests the interested reader is referred to King (1987).

[20] There are various approaches that one might take here. Cox and Hinkley (1974) provide a wonderful discussion of these issues.

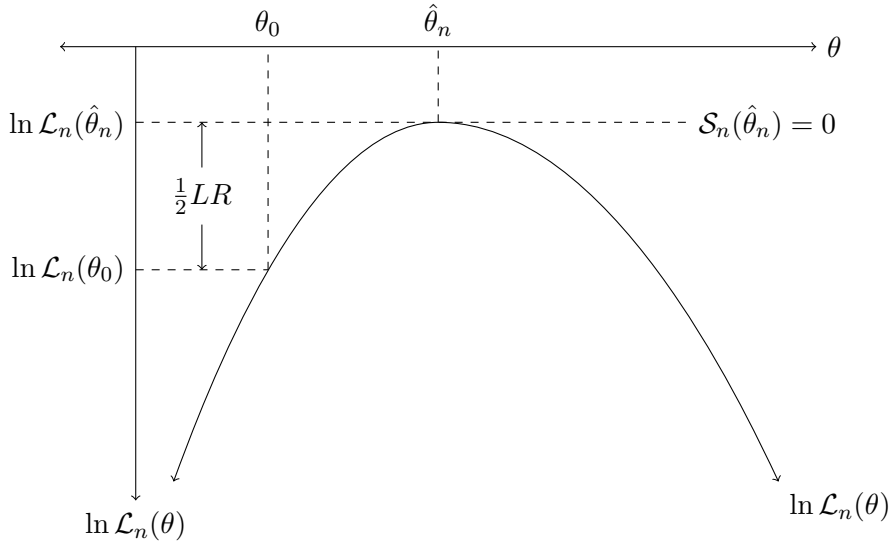

Figure 3: The Likelihood Ratio Test

Before proceeding it is useful to step back and think about what we are trying to do. If we don't know a parameter then our usual response is to estimate it. If we want to test an hypothesis about a parameter then an obvious thing to do is simply to compare our estimate with the parameter value hypothesized under H_0. If they are close to one another then this could be interpreted as sample evidence in support of the parameter restriction embodied in H_0. Conversely, if our estimate and the hypothesized value are far apart then this would constitute a lack of sample evidence in favour of H_0.[21] Rather than comparing parameter values directly, what the likelihood ratio test does is to compare the likelihood functions from the null and alternative models, respectively.

It might be helpful at this stage to consider Figure 3, which plots values of log-likelihood functions (on the vertical axis) against parameter values θ (on the horizontal axis). We see that the mle $\hat{\theta}_n$ is the value of θ at which $\ln \mathcal{L}_n(\theta)$ is maximized, with maximum value $\ln \mathcal{L}_n(\hat{\theta}_n)$. Now suppose that we hypothesize $\theta = \theta_0$. It must be the case that $\ln \mathcal{L}_n(\theta_0) \leq \ln \mathcal{L}_n(\hat{\theta}_n)$ because $\ln \mathcal{L}_n(\hat{\theta}_n)$ is an unrestricted maximum of $\ln \mathcal{L}_n(\theta)$. The basis for the Likelihood Ratio test is then the vertical distance $\ln \mathcal{L}_n(\hat{\theta}_n) - \ln \mathcal{L}_n(\theta_0)$ and our rejection rule will be to reject H_0 when this distance is too large. The only remaining issue is determination of an appropriate critical value. The exact distribution theory will differ from problem to problem, however, one characteristic of the Likelihood Ratio test that makes it important is that it can be shown, under certain regularity conditions, that

$$LR = -2 \ln \hat{\lambda} \underset{a}{\overset{H_0}{\sim}} \chi^2_1, \tag{9}$$

where

$$\hat{\lambda} = \frac{\mathcal{L}_n(\theta_0; y_1, \ldots, y_n)}{\mathcal{L}_n(\hat{\theta}_n; y_1, \ldots, y_n)} \tag{10}$$

and the notation $\underset{a}{\overset{H_0}{\sim}}$ should be read as 'under H_0 is asymptotically (or approximately) distributed as', meaning when H_0 is true, $-2 \ln \hat{\lambda}$ is approximately (which is what the

[21] This is the idea underlying the Wald test but more on that later.

'a' signifies) distributed as a chi-squared random variable with one degree of freedom.[22] And so critical values can be obtained from standard probability tables, which makes the LR test easy to use. Note that LR is the statistic that most econometricians would be thinking of when they spoke of 'the' Likelihood Ratio test. When using λ as a test statistic the decision rule is to reject H_0 for small λ. In contrast, when using LR as a test statistic the decision rule is to reject H_0 for large values of the statistic; namely, reject H_0 for $LR \geq \chi^2_{1,\alpha}$, where $\chi^2_{1,\alpha}$ is that value which cuts of an upper-tail probability of α in a χ^2_1 distribution.

Let us extend our previous example to the case of testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. We proceed as before but we replace μ_1 by the unrestricted mle $\hat{\mu}_n = \bar{y}_n$. First,

$$\begin{aligned}
\hat{\lambda} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^n (y_i - \mu_0)^2/(2\sigma^2)\}}{(2\pi\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^n (y_i - \bar{y}_n)^2/(2\sigma^2)\}} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[(y_i - \mu_0)^2 - (y_i - \bar{y}_n)^2\right]\right\} \\
&= \exp\left\{-\frac{n(\bar{y}_n - \mu_0)^2}{2\sigma^2}\right\},
\end{aligned}$$

where we have used the result[23]

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y}_n)^2 &= \sum_{i=1}^n ((y_i - \mu_0) - (\bar{y}_n - \mu_0))^2 \\
&= \sum_{i=1}^n (y_i - \mu_0)^2 - 2\sum_{i=1}^n (y_i - \mu_0)(\bar{y}_n - \mu_0) + \sum_{i=1}^n (\bar{y}_n - \mu_0)^2 \\
&= \sum_{i=1}^n (y_i - \mu_0)^2 - 2(\bar{y}_n - \mu_0)\sum_{i=1}^n (y_i - \mu_0) + n(\bar{y}_n - \mu_0)^2 \\
&= \sum_{i=1}^n (y_i - \mu_0)^2 - 2(\bar{y}_n - \mu_0)n(\bar{y}_n - \mu_0) + n(\bar{y}_n - \mu_0)^2, \\
&= \sum_{i=1}^n (y_i - \mu_0)^2 - n(\bar{y}_n - \mu_0)^2,
\end{aligned}$$

which implies that

$$\sum_{i=1}^n (y_i - \mu_0)^2 - \sum_{i=1}^n (y_i - \bar{y}_n)^2 = n(\bar{y}_n - \mu_0)^2.$$

Thus,

$$P(LR \geq \kappa'_\alpha) = P\left(\frac{n(\bar{y}_n - \mu_0)^2}{\sigma^2} \geq \kappa_\alpha\right). \tag{11}$$

[22] The approximation is a large-sample asymptotic one, which means that it becomes increasingly accurate as the sample size increases. For this reason you also see people use the symbol $\underset{a}{\sim}$, or sometimes $\overset{a}{\sim}$, to denote 'asymptotically distributed as'.

[23] We used essentially the same device to obtain (7). Repeated use of a result should be a flag to add it to your kit of tricks.

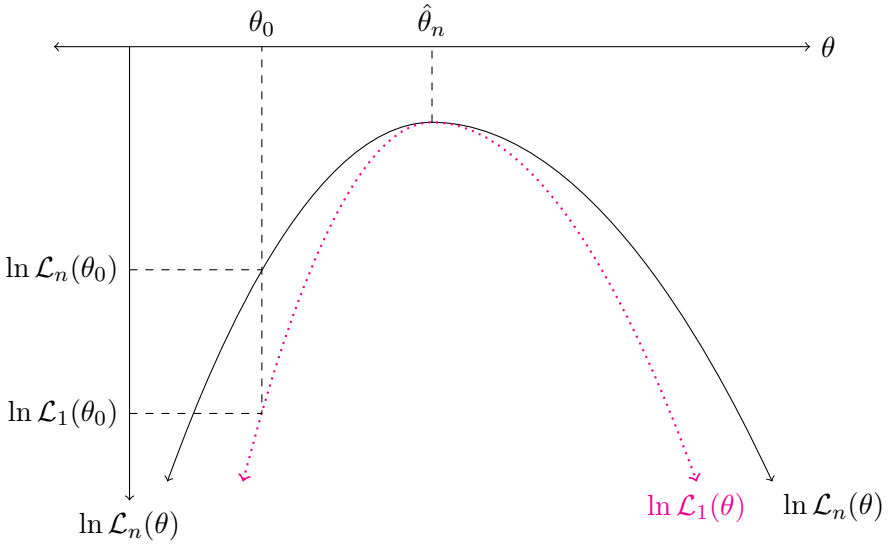

Figure 4: The Wald Test

If H_0 is true then it follows that $LR \sim \chi^2_1$ and so we should choose $\kappa_\alpha = \chi^2_{1,\alpha}$. It is interesting to note that in our example the asymptotic distribution of LR corresponds to the exact distribution, this is not true in general.

1.2.3 The Wald Test

As discussed above, it is clear that an equivalent sort of idea to that underlying the Likelihood Ratio test would be to look at the distance between the hypothesized value θ_0 and the mle $\hat{\theta}_n$; that is, base a test on the quantity $\hat{\theta}_n - \theta_0$, or $(\hat{\theta}_n - \theta_0)^2$ if sign is unimportant, as it would be if H_1 is two-sided. This is the basic idea underlying a Wald test, although we need to transform this random variable to obtain a test statistic with a standard distribution so that we can obtain appropriate critical values in an easy way.

Some intuition into the structure of the Wald test can be gained from Figure 4. Consider the two likelihood functions denoted $\mathcal{L}_n(\theta)$ and $\mathcal{L}_1(\theta)$. They both have the same maximum value occurring at $\hat{\theta}_n$ although $\mathcal{L}_n(\theta_0) > \mathcal{L}_1(\theta_0)$ for all other θ. If one was performing a Likelihood Ratio test it is easy to see how one might accept H_0 on the basis of $\mathcal{L}_n(\theta_0)$ yet reject H_0 on the basis of $\mathcal{L}_1(\theta_0)$. However, in both cases the distance $\hat{\theta}_n - \theta_0$ is the same, which suggests that $\hat{\theta}_n - \theta_0$ is inadequate as a test statistic. The important difference between the two likelihoods is that $\mathcal{L}_n(\theta)$ is flatter than is $\mathcal{L}_1(\theta)$ near their turning points, and so a given change in θ has less impact on the value of $\mathcal{L}_n(\theta)$ than it does on $\mathcal{L}_1(\theta)$. This suggests that we need to take account of the curvature of the likelihood function when trying to determine whether or not $\hat{\theta}_n - \theta_0$ is large. How does one measure curvature? One common measure is $-\mathcal{H}_n(\theta)$, which gives the rate of change of $\mathcal{S}_n(\theta)$ which, in turn, is the slope of tangent to the likelihood function. If the likelihood function is very curved then the second derivative will be large in magnitude whereas it will be closer to zero for flatter functions.[24] This suggests a statistic of the form $-\mathcal{H}_n(\theta)(\hat{\theta}_n - \theta_0)^2$, which scales $(\hat{\theta}_n - \theta_0)^2$ in accordance with the curvature of the likelihood. As it stands this statistic is not feasible, as it requires knowledge of θ in order to calculate $\mathcal{H}_n(\theta)$, but it can be made feasible if we replace θ by $\hat{\theta}_n$. Finally, it is typically

[24] If you can't see this then plot out some parabolas of the form $y = ax^2 + bx + c$ and examine how their shape changes as you vary a.

the case that we replace $-\mathcal{H}_n(\hat{\theta}_n)$ by its expectation $\mathcal{I}_n(\hat{\theta}_n)$ which yields the usual form of the Wald statistic

$$W = \mathcal{I}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \overset{H_0}{\underset{a}{\sim}} \chi^2_1.$$

Observe that the Wald statistic has the same asymptotic distribution as does the Likelihood Ratio test. This does not mean that the two tests are the same thing! For example, you could take two draws from a Normal distribution and get one extremely large negative value and one extremely large positive value. Just because two random variables have the same probability distribution does not mean that they need be close together.

Let us continue our example of testing a restriction on the mean of a random variable whose distribution is Normal, with known variance. Our statistic is of the form

$$W = \mathcal{I}_n(\hat{\mu}_n)(\hat{\mu}_n - \mu_0)^2.$$

We have already shown on Page 6 that $\mathcal{I}_n(\hat{\mu}_n) = n/\sigma^2$ and so does not depend on $\hat{\mu}_n$.[25] Furthermore $\hat{\mu}_n = \bar{y}_n$. Hence

$$W = \frac{n(\bar{y}_n - \mu_0)^2}{\sigma^2}, \tag{12}$$

and our decision rule is to reject H_0 for large values of W. Note that, in this special case, W is identical to the LR test of (11) and so all of the earlier results carry over directly.

We note in passing that things like t-tests and F-tests are very much in the spirit of Wald tests. (If you are are paying attention it should be clear that these tests can also be related to likelihood ratio tests.) We also note that W in (12) is the statistic that you would use for a two-sided alternative. If testing against a one-sided alternative then you would work with the square root of this statistic, but you may need to adjust the sign of the statistic, depend on your alternative hypothesis.

1.2.4 The Lagrange Multiplier (or Efficient Score) Test

This test has two names because it has (at least) two equivalent forms, the efficient score test and the Lagrange Multiplier test, which were proposed independently by Rao (1948) and Silvey (1959), respectively.[26] The efficient score version of the test is the easiest to visualize. Consider next Figure 5. If we look at the solid black curve $\ln \mathcal{L}_n(\theta)$ we see that at $\hat{\theta}_n$ the slope of the tangent to the likelihood function (depicted by the dashed line labelled $\mathcal{S}_n(\hat{\theta}_n) = 0$) is zero. But tangents to the likelihood function at any point θ have a slope equal to the score function evaluated at that value of θ, which is why $\mathcal{S}_n(\hat{\theta}_n) = 0$. This suggests that we can test $H_0 : \theta = \theta_0$ by comparing $\mathcal{S}_n(\theta_0)$ to zero, which is the fundamental idea behind the efficient score test although there is still more work to be done. Note that in Figure 5, the tangent to $\ln \mathcal{L}_n(\theta)$ at θ_0 is depicted by the blue dashed line labelled AA'.

If we look at the red dotted curve labelled $\ln \mathcal{L}_2(\theta)$ we see that it too is tangent to AA' at θ_0. If one was to simply consider the value of the score at θ_0 then it would be the case that such a test would yield the same outcome regardless of whether the log-likelihood function is $\ln \mathcal{L}_n(\theta)$ or $\ln \mathcal{L}_2(\theta)$. But there is no reason why this should be so because θ_0 is clearly much closer to $\hat{\theta}_{n,2}$, the maximized value of $\ln \mathcal{L}_2(\theta)$, that it is to $\hat{\theta}_n$, and so one might expect that the outcome of the test should depend upon the characteristics of the

[25] In general $\mathcal{I}_n(\hat{\theta}_n)$ will depend on $\hat{\theta}_n$.

[26] There is also a version due to Breusch and Pagan (1979) and there may be others as well.

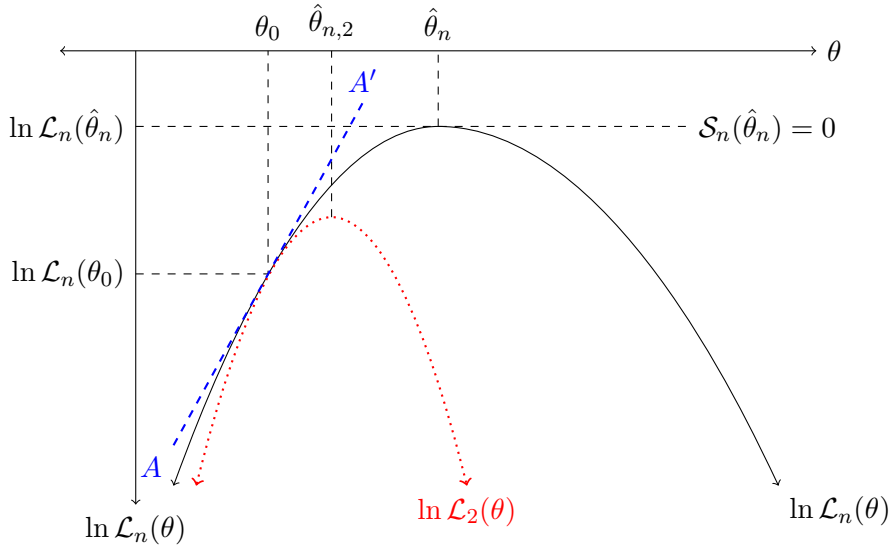

Figure 5: The Lagrange Multiplier Test

log-likelihood function. If we think in terms of the curvature of the log-likelihood then we see that the curvature is greatest at the turning point; that is, $\mathcal{H}_n(\theta)$ will be largest at $\hat{\theta}_n$. If H_0 is true or close to true then $-\mathcal{H}_n(\theta)$ at $\theta = \theta_0$ should be large or, equivalently, $(-\mathcal{H}_n(\theta))^{-1}$ should be small near θ_0 if H_0 is true. As we saw with the Wald test, the so-called *efficient form* of the test involves replacing $-\mathcal{H}_n(\theta_0)$ by $\mathcal{I}_n(\theta_0)$, which yields the following result:

$$LM = (\mathcal{I}_n(\theta_0))^{-1}(\mathcal{S}_n(\theta_0))^2 \underset{a}{\overset{H_0}{\sim}} \chi^2_1. \tag{13}$$

There are four important comments that need to be made. First, and most importantly, be aware that, in contrast to the Wald test which uses unrestricted mles, the LM test evaluates everything under the null hypothesis, so θ is replaced by θ_0 rather than by $\hat{\theta}_n$. Second, the LM test is an optimal test in that it is said to be *locally best*. What this means is that among tests of a given size, that test with the steepest power function local to the origin is the LM test. Now I am not going to claim that this is necessarily a compelling property because, close to the null, it often doesn't make a great deal of difference whether you believe the null to be true or not. (That said, it can, as in the case of unit root models, where there is a cosmic shift in behaviours between a stationary model and a non-stationary model.) Nevertheless, any optimality property is arguably better than none at all. Third, we see that LM has the same asymptotic distribution as do LR and W. As observed in the discussion of W, this does not mean that they take the same values, merely that they all have approximately the same sampling distribution. Any given sample, however, may result in values from anywhere in the support of that distribution, which is the unbounded interval $[0, \infty)$. So the statistics may take very different values and may yield different outcomes to the testing problem. Finally, I have dubbed the statistic LM, although the development given is that for the efficient score test. This reflects econometric terminology, however, in many other areas the test would be known as the score test rather than the Lagrange Multiplier test, which is not unreasonable given the historical precedence of the former.

So where does the Lagrange Multiplier name come from? Set up the Lagrangian $\Phi(\theta)$

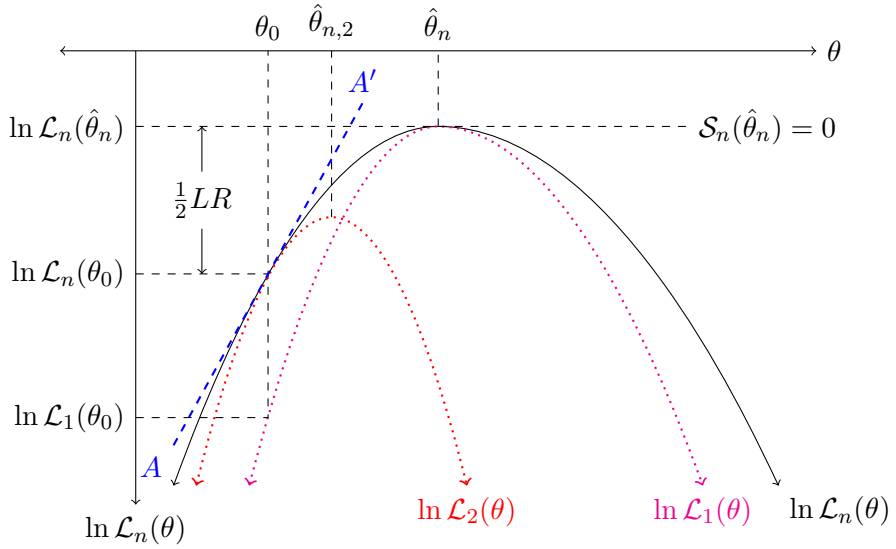

Figure 6: A Trinity of Tests

associated with imposing the null hypothesis during estimation. Thus,

$$\Phi(\theta) = \ln \mathcal{L}_n(\theta) - \phi(\theta - \theta_0).$$

where ϕ denotes the Lagrange Multiplier. The first-order conditions are

$$\frac{\partial \Phi(\theta)}{\partial \theta} = \mathcal{S}_n(\theta) - \phi = 0 \Rightarrow \mathcal{S}_n(\theta) = \phi$$

and

$$\frac{\partial \Phi(\theta)}{\partial \phi} = \theta - \theta_0 = 0 \Rightarrow \theta = \theta_0.$$

Consequently, large values of the score $\mathcal{S}_n(\theta)$ correspond to large values of ϕ and they both measure the cost of imposing the restriction.

We will conclude this section by completing our on-going example. Adapting (5), if H_0 is true then $\mathcal{S}_n(\mu_0) = n(\bar{y}_n - \mu_0)/\sigma^2$. From the discussion immediately preceding (5) we have that $(\mathcal{I}_n(\mu_0))^{-1} = \sigma^2/n$. It follows from (13), therefore, that $LM = n(\bar{y}_n - \mu_0)^2/\sigma^2$. This statistic is identical to both the LR and W statistics and so will yield an identical value for the test statistic. Clearly, under our assumptions $LM \sim \chi^2_1$.

The fact that, in this example, all three statistics are the same is an artifact of our example, specifically because we have treated σ^2 as known, and is not true in general. So whereas all three tests will necessarily yield the same outcome for any given data set, that is not true in general.

As a final observation, it is quite common to see Figures 3 – 5 combined into a single diagram. This is done in Figure 6.[27]

1.2.5 A Concrete Example

Suppose that $X \sim \mathcal{B}(\pi)$, where $\mathcal{B}(\pi)$ denotes a Bernoulli distribution such that $\mathrm{P}(X = 1) = \pi$, $\pi \in [0, 1]$. Further suppose that we wish to test $H_0 : \pi = \pi_0$, $\pi_0 \in (0, 1)$, against a two-sided alternative on the basis of a simple random sample of size n.

[27] Figures 3 – 5 are due to Buse (1982).

Let X_i denote the ith observation on X. Then the log-likelihood is[28]

$$\begin{aligned}\ln \mathcal{L}_n(\pi) &= \sum_{i=1}^{n}[X_i \ln \pi + (1 - X_i)\ln(1-\pi)], \quad 0 < \pi < 1, \\ &= n\overline{X}_n \ln \pi + n(1-\overline{X}_n)\ln(1-\pi), \quad \text{where } \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \end{aligned} \tag{14}$$

The standard trinity of test statistics is as follows:

$$LR = 2[\ln \mathcal{L}_n(\hat{\pi}) - \ln \mathcal{L}_n(\pi_0)] \tag{15}$$

$$LM = \left(\frac{d \ln \mathcal{L}_n(\pi_0)}{d\pi}\right)^2 / I(\pi_0) \tag{16}$$

$$W = (\hat{\pi} - \pi_0)^2 I(\hat{\pi}) \tag{17}$$

where Fisher's information is defined to be

$$I(\pi^*) = -\mathrm{E}\left[\frac{d^2 \ln \mathcal{L}_n(\pi)}{d\pi^2}\right]_{\pi=\pi^*}.$$

where π^* denotes the population value of π.

To begin, the MLE is the solution to the first-order condition $dL(\pi)/d\pi = 0$, where

$$\frac{d \ln \mathcal{L}_n(\pi)}{d\pi} = \frac{n\overline{X}_n}{\pi} - \frac{n(1-\overline{X}_n)}{1-\pi} = \frac{n\overline{X}_n(1-\pi) - n(1-\overline{X}_n)\pi}{\pi(1-\pi)} = \frac{n(\overline{X}_n - \pi)}{\pi(1-\pi)}.$$

Clearly, $\hat{\pi} = \overline{X}_n$ is the solution to the first-order condition. We see immediately that the LR test statistic is[29]

$$LR = 2n\left[\overline{X}_n \ln\left(\frac{\overline{X}_n}{\pi_0}\right) + (1-\overline{X}_n)\ln\left(\frac{1-\overline{X}_n}{1-\pi_0}\right)\right]. \tag{18}$$

[28] If $\pi \in \{0, 1\}$ then the joint density of X is degenerate at either 0 or 1, as appropriate. Consequently, any sample with at least one 0 and one 1 excludes these cases. Any sample comprised of only zeros or ones is useless for statistical inference as we have no way of knowing whether it tells us the value of π with certainty or whether it is just a completely uninformative sample.

[29] As an aside, asymptotically we know that the sampling distribution of LR under H_0 is χ^2_1. At a minimum, this requires that LR be non-negative for all possible samples, or equivalently, for all possible $\overline{X}_n$. (Strictly, this should read for all possible samples except for a set of measure zero, being either a sample of all zeroes or a sample of all ones, whereupon either $\overline{X}_n = 0$ or $\overline{X}_n = 1$, respectively.) To see that this is satisfied, consider the behaviour of LR as $\overline{X}_n$ varies, conditional upon π_0. Observe that

$$\frac{dLR}{d\overline{X}_n} = 2n\left[\ln\left(\frac{\overline{X}_n}{\pi_0}\right) - \ln\left(\frac{1-\overline{X}_n}{1-\pi_0}\right)\right],$$

and so the first-order condition suggests that, in the interval $0 < \overline{X}_n < 1$, there is but one stationary point, occurring at $\overline{X}_n = \pi_0$. Checking the second order condition, we see that

$$\frac{d^2 LR}{d\overline{X}_n^2} = 2n\left[\frac{1}{\overline{X}_n} + \frac{1}{1-\overline{X}_n}\right] > 0, \qquad \text{for all } 0 < \overline{X}_n < 1 \text{ and for all } \pi_0,$$

implying that our stationary point is a minimum. Finally, as $LR|_{\overline{X}_n=\pi_0} = 0$, we see that $LR \geq 0$ for all $\overline{X}_n$, as required.

Both the Lagrange Multiplier (LM) and Wald (W) tests require knowledge of Fisher's information and so before considering either test we will examine that. Our starting point is

$$-H(\pi) = -\frac{d^2 \ln \mathcal{L}_n(\pi)}{d\pi^2} = \frac{n\overline{X}_n}{\pi^2} + \frac{n(1-\overline{X}_n)}{(1-\pi)^2}. \tag{19}$$

Hence, the true value of[30]

$$I(\pi^*) = \frac{n\mathrm{E}[\overline{X}_n]}{(\pi^*)^2} + \frac{n(1-\mathrm{E}[\overline{X}_n])}{(1-\pi^*)^2} = \frac{n}{\pi^*(1-\pi^*)}, \quad \text{as } \mathrm{E}[\overline{X}_n] = \pi^*. \tag{20}$$

Of course, in practice, π^* is unknown and so the question is how to proceed. One obvious approach is to plug-in values for π^* in equation (20). Specifically,

$$I(\pi_0) = \frac{n}{\pi_0(1-\pi_0)} \qquad (LM \text{ approach}) \tag{21}$$

$$I(\overline{X}_n) = \frac{n}{\overline{X}_n(1-\overline{X}_n)} \qquad (\text{Wald approach}) \tag{22}$$

An alternative approach is to work with equation (19). Observe that, if H_0 is true, so that $\pi_0 = \pi^*$, then $n^{-1}H(\pi_0)$ will converge to $\lim_{n\to\infty n} n^{-1}I(\pi^*)$,[31] because $\overline{X}_n$ converges to π^*. However, for any finite sample,[32]

$$\frac{n\overline{X}_n}{\pi_0^2} + \frac{n(1-\overline{X}_n)}{(1-\pi_0)^2} \neq \frac{n}{\pi_0(1-\pi_0)} \quad a.s.$$

Contrast that result with the Wald approach where $-H(\overline{X}_n)$ is identically equal to $I(\overline{X}_n)$. Moreover, because $\overline{X}_n$ converges to π^*, $n^{-1}I(\overline{X}_n)$ converges to $\lim_{n\to\infty n} n^{-1}I(\pi^*)$, regardless of whether or not $\pi_0 = \pi^*$.

[30] Equation (20) clearly illustrates that Fisher's information is an increasing function of sample size, the intuition being that each additional observation brings with it some more information that can be used for inference. When this relationship breaks down, i.e. when additional observations don't increase the informational content of the sample, then our usual techniques of inference break down. This problem can manifest itself in many different ways. For example, in talking about asymptotically uncooperative regressors, Schmidt (1976, p. 87) uses the time series example of a regressor of the form $x_t = \lambda^t$, where $|\lambda| < 1$. Essentially the same problem manifests itself when trying to estimate ratios where the denominator may be zero, e.g. Hirschberg and Lye (2005), and the weak instrument problem, e.g. Stock, Wright, and Yogo (2002), Dufour (2003), Hahn and Hausman (2003), Andrews and Stock (2007), Poskitt and Skeels (2013).

[31] Note the need to be careful with scaling to ensure that none of the quantities being discussed diverge as $n \to \infty(n)$.

[32] The notation *a.s.* stands for *almost surely* which means that it is true for every sample except for a set of measure zero. What that means in this case is that there may be a some finite set of sample where the equality

$$\frac{n\overline{X}_n}{\pi_0^2} + \frac{n(1-\overline{X}_n)}{(1-\pi_0)^2} = \frac{n}{\pi_0(1-\pi_0)}$$

actually holds, but that the probability of observing such a sample, from the infinitude of possible samples (when n is infinitely large) is zero.

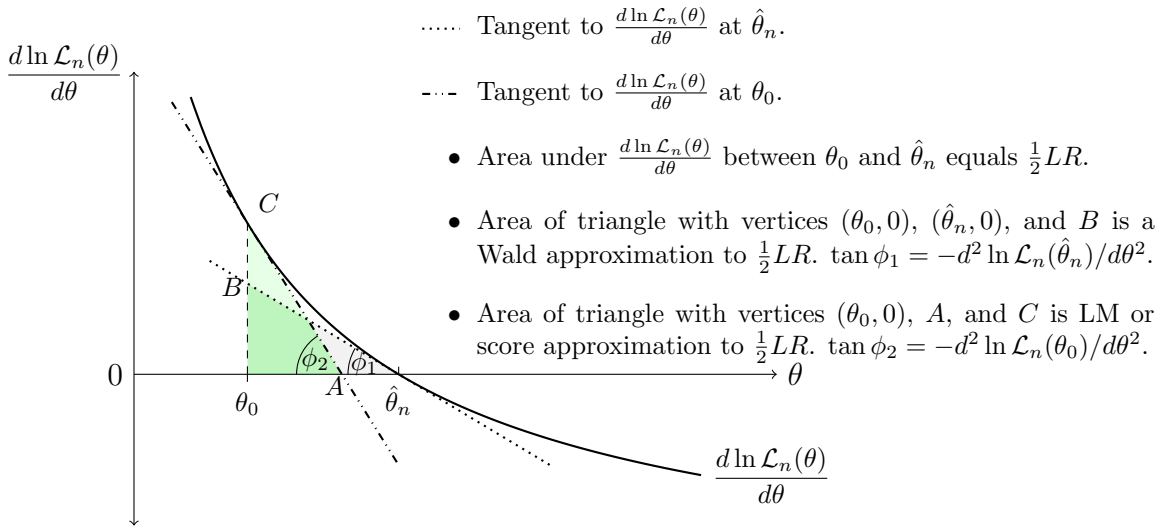

Figure 7: Pagan's Graphical Comparison of the Wald, LM, and LR Tests

1.2.6 An Alternative Graphical Exposition

Although Figure 6 is the most well-known graphical representation of the three classical tests, it is not the only one. Another, originally due to Pagan (1982) is presented in Figure 7. This diagram plots the the score, or derivative of the log-likelihood, against the parameter θ. This is represented in the figure by the thick black curve labelled $d \ln \mathcal{L}_n(\theta)/d\theta$. The relationships between the various tests are then explored by comparing the areas of various regions on the diagram.

To begin, consider the area under the curve, between θ_0 and $\hat{\theta}_n$, which is calculated as

$$\int_{\theta_0}^{\hat{\theta}_n} \frac{d \ln \mathcal{L}_n(\theta)}{d\theta} \, d\theta = \ln \mathcal{L}_n(\hat{\theta}_n) - \ln \mathcal{L}_n(\theta_0) = \frac{1}{2} LR.$$

That is, the nominated area under the curve is proportional to the value of the LR test.

There are two other regions of interest in Figure 7, namely two right-angle triangles which might be used to approximate the area under the curve — such as one might do in welfare analysis — one shaded grey, the other shaded light-green, with their overlap a darker green. Both triangles have the property that their hypotenuse lies on a line that is tangent to the score at some point. Specifically, the hypotenuse of the light green triangle lies on a line that is tangent to $d \ln \mathcal{L}_n(\theta)/d\theta$ at $\theta = \theta_0$ and that of the grey triangle lies on a line that is tangent to $d \ln \mathcal{L}_n(\theta)/d\theta$ at $\theta = \hat{\theta}_n$. An obvious implication of this collinearity is that the hypotenuse and the corresponding tangent have the same slope, which will be exploited below. Furthermore, the tangent to $d \ln \mathcal{L}_n(\theta)/d\theta$ at any point, θ^* say, has the property that

$$\left[\frac{d}{d\theta} \left(\frac{d \ln \mathcal{L}_n(\theta)}{d\theta} \right) \right]_{\theta=\theta^*} = \frac{d^2 \ln \mathcal{L}_n(\theta^*)}{d\theta^2} = \mathcal{H}_n(\theta^*).$$

Looking first at the grey triangle, with vertices $(\theta_0, 0)$, $(\hat{\theta}_n, 0)$, and $B = (\theta_0, b)$ (say), we see that it has area $A_W = b(\hat{\theta}_n - \theta_0)/2$. The slope of the hypotenuse of this triangle (and hence the slope of the tangent) is given by

$$\mathcal{H}_n(\hat{\theta}_n) = \frac{0 - b}{\hat{\theta}_n - \theta_0} = -\frac{b}{\hat{\theta}_n - \theta_0} \Rightarrow b = \left[-\mathcal{H}_n(\hat{\theta}_n) \right] (\hat{\theta}_n - \theta_0).$$

That is,[33]

$$A_W = \frac{1}{2}\left[-\mathcal{H}_n(\hat{\theta}_n)\right](\hat{\theta}_n - \theta_0)^2. \tag{23}$$

If we use $-\mathcal{H}_n(\hat{\theta}_n)$ as an estimator for $\mathcal{I}_n(\theta)$, rather than $\mathcal{I}_n(\hat{\theta}_n)$, then we see that A_W is approximately equal to $\frac{1}{2}W$. In particular, we can think of the Wald statistic as providing a triangle approximation to LR.

Consider next the light-green triangle with vertices $A = (a, 0)$ (say), $C = (\theta_0, c)$ (say), where $c = d \ln \mathcal{L}_n(\theta_0)/d\theta$, and $(\theta_0, 0)$ with area $A_{LM} = c(a - \theta_0)/2$. The tangent to the score at θ_0 has slope

$$\mathcal{H}_n(\theta_0) = \frac{0 - c}{a - \theta_0} \Rightarrow a - \theta_0 = -\frac{c}{\mathcal{H}_n(\theta_0)}.$$

Making this substitution, the area of the light-green triangle becomes

$$A_{LM} = \frac{c^2}{2[-\mathcal{H}_n(\theta_0)]}.$$

Noting that c is the score evaluated at θ_0, and that $-\operatorname{E}[\mathcal{H}_n(\theta_0)] = \mathcal{I}_n(\theta_0)$, we see that $2A_{LM}$ corresponds to a form of the LM statistic. That is, the LM statistic corresponds to twice that triangle approximation to the LR statistic provided by the light-green triangle.[34]

1.2.7 Summary

The key features to take away from the preceding discussion are:

1. The Likelihood Ratio test requires you to evaluate the model in both its restricted and unrestricted forms. If either one of these models is difficult to estimate then the LR test will be difficult to implement.

2. The Wald test only requires you to find the mles for the unrestricted model. This will be particularly attractive if the restricted model is hard to evaluate. Sometimes it is hard to impose restrictions.

3. The LM test only requires you to estimate the restricted model. If the unrestricted model is difficult to estimate then the LM test will be attractive.

4. All of these tests have the same asymptotic distribution. That does not mean that they will take values close to one another or, more importantly, that they will always lead you to the same decision.

5. Sometimes, you can find exact critical values for these tests, but it may require a bit of work.

[33] As an aside, an alternative derivation of A_W follows on noting that $\tan\phi_1 = b/(\hat{\theta}_n - \theta_0)$, so that $b = (\hat{\theta}_n - \theta_0)\tan\phi_1$. Then $A_W = (\hat{\theta}_n - \theta_0)^2(\tan\phi_1)/2$. That this expression is identical to that given in (23) follows from the fact that $\tan\phi_1 = b/(\hat{\theta}_n - \theta_0) = -\mathcal{H}_n(\hat{\theta}_n)$.

[34] Again, a trigonometric version of the argument follows on noting that $a - \theta_0 = c/\tan\phi_2$, so that $A_{LM} = c^2/(2\tan\phi_2)$, and also noting that

$$\tan\phi_2 = \frac{c}{a - \theta_0} = -\mathcal{H}_n(\theta_0).$$

2 Maximum Likelihood With More Than One Parameter

2.1 Estimation

Almost all of the ideas that we might need to know in respect of maximum likelihood techniques have been introduced in the previous sections. The only purpose of this section is to indicate how things generalize to cases of more than one parameter.

For ease of exposition we shall couch the following discussion in terms of an example. The obvious example for us is that of a Normal distribution where now neither the mean μ nor the variance σ^2 are known. As in our earlier examples we shall base inference on a sample of n independent draws from the population: $y_1, \ldots, y_n$. As before the joint density function for the sample is

$$f(y_1, \ldots, y_n; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

and the mles are the solution to

$$\underset{\mu,\sigma^2}{\operatorname{argmax}}\, \mathcal{L}_n(\mu, \sigma^2),$$

where, as before the likelihood function $\mathcal{L}_n(\mu, \sigma^2)$ is simply the joint density function interpreted as a function of the parameters *given* the data rather than a function of the data given the parameters.

Now, we shall rewrite the problem as follows. Let $y = [y_1, \ldots, y_n]'$, a $n \times 1$ vector. Further, let θ denote the parameter vector $\theta = [\mu, \sigma^2]'$. In our example θ is a 2×1 vector but, in general, it might be a $p \times 1$ vector, where $p \leq n$. We shall, hereafter, use p as the dimension of $\theta = [\theta_1, \ldots, \theta_p]'$ but, in this example, p should be understood to equal 2. We shall write the joint density function as

$$f(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta) \quad \left(= \prod_{i=1}^{n} f_i(y_i; \mu, \sigma^2) \text{ in this case}\right),$$

and we shall write $\mathcal{L}_n(\theta; y) = f(y; \theta)$. Moving forward, we shall typically suppress the data (y) in the notation except for the special case where we are thinking of the functions conditional on a particular observation. For example, we might have reason to note that

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \mathcal{L}_n(\theta; y_i), \qquad \mathcal{L}_n(\theta; y_i) \equiv f_i(y_i; \theta).$$

We shall adopt this convention with all related functions, including the score, the hessian and the information matrix. The estimation problem is then to solve $\operatorname{argmax}_\theta \mathcal{L}_n(\theta; y)$.

We next define the score function as

$$\mathcal{S}_n(\theta) = \frac{\partial \ln \mathcal{L}_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln \mathcal{L}_n(\theta; y_i)}{\partial \theta} = \sum_{i=1}^{n} \mathcal{S}_n(\theta; y_i),$$

where

$$\frac{\partial}{\partial \theta} = \left[\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_p}\right]'$$

denotes a $p \times 1$ vector of partial differential operators. So, in this case,

$$\mathcal{S}_n(\theta) = \begin{bmatrix} \dfrac{\partial \ln \mathcal{L}_n(\theta)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ln \mathcal{L}_n(\theta)}{\partial \theta_p} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \ln \mathcal{L}_n(\theta)}{\partial \mu} \\ \dfrac{\partial \ln \mathcal{L}_n(\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (y_i - \mu) \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \displaystyle\sum_{i=1}^{n} (y_i - \mu)^2 \end{bmatrix}.$$

The first-order conditions are then $\mathcal{S}_n(\theta) = 0$, where here 0 denotes a $p \times 1$ vector with every element equal to zero. The point where $\mathcal{S}_n(\theta) = 0$ defines the mles and this is where we introduce the $\hat{}$'s:

$$0 = \mathcal{S}_n(\hat{\theta}_n) = \begin{bmatrix} \dfrac{1}{\hat{\sigma}_n^2} \displaystyle\sum_{i=1}^{n} (y_i - \hat{\mu}_n) \\ -\dfrac{n}{2\hat{\sigma}_n^2} + \dfrac{1}{2\hat{\sigma}^4} \displaystyle\sum_{i=1}^{n} (y_i - \hat{\mu}_n)^2 \end{bmatrix}.$$

Thus we have two equations in two unknowns which can be solved to yield

$$\hat{\theta}_n = \begin{bmatrix} \hat{\mu}_n \\ \hat{\sigma}_n^2 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (y_i - \hat{\mu}_n)^2 \end{bmatrix} = \begin{bmatrix} \bar{y}_n \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 \end{bmatrix}.$$

Note that the mle for σ^2 differs from the usual unbiased estimator

$$s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2.$$

That is, $\hat{\sigma}_n^2 = (n-1)s_n^2/n$. As s_n^2 is unbiased it follows that $\hat{\sigma}_n^2$ isn't; indeed,

$$\mathrm{E}\left[\hat{\sigma}_n^2\right] = (n-1)\sigma^2/n.$$

Observe, however, that as $n \to \infty$, $(n-1)/n \to 1$, and so the bias in $\hat{\sigma}_n^2$ disappears as the sample size becomes large. This is in line with our earlier observation that mles are consistent.

The second-order conditions require a little more care. We know that they involve derivatives of the score. But here the score is a vector and so we need to be able to differentiate a vector with respect to the vector θ. The convention is that you differentiate column vectors (like the score) with respect to row vectors. Thus,

$$\mathcal{H}_n(\theta) = \frac{\partial \mathcal{S}_n'(\theta)}{\partial \theta} = \frac{\partial^2 \ln \mathcal{L}_n(\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln \mathcal{L}_n(\theta; y_i)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \mathcal{H}_n(\theta; y_i),$$

where $\partial \theta' = (\partial \theta)'$. The important thing to observe about this expression is that $\mathcal{H}_n(\theta)$

is a $p \times p$ matrix whose ijth element is given by $\partial^2 \ln \mathcal{L}_n(\theta)/\partial\theta_i\partial\theta_j$. In our example,

$$\mathcal{H}_n(\theta) = \begin{bmatrix} \dfrac{\partial \mathcal{S}_{n,1}(\theta)}{\partial \theta_1} & \dfrac{\partial \mathcal{S}_{n,2}(\theta)}{\partial \theta_1} \\ \dfrac{\partial \mathcal{S}_{n,1}(\theta)}{\partial \theta_2} & \dfrac{\partial \mathcal{S}_{n,2}(\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \mathcal{S}_{n,1}(\mu,\sigma^2)}{\partial \mu} & \dfrac{\partial \mathcal{S}_{n,2}(\mu,\sigma^2)}{\partial \mu} \\ \dfrac{\partial \mathcal{S}_{n,1}(\mu,\sigma^2)}{\partial \sigma^2} & \dfrac{\partial \mathcal{S}_{n,2}(\mu,\sigma^2)}{\partial \sigma^2} \end{bmatrix}$$
$$= \sum_{i=1}^{n} \begin{bmatrix} \dfrac{\partial \mathcal{S}_{n,1}(\mu,\sigma^2;y_i)}{\partial \mu} & \dfrac{\partial \mathcal{S}_{n,2}(\mu,\sigma^2;y_i)}{\partial \mu} \\ \dfrac{\partial \mathcal{S}_{n,1}(\mu,\sigma^2;y_i)}{\partial \sigma^2} & \dfrac{\partial \mathcal{S}_{n,2}(\mu,\sigma^2;y_i)}{\partial \sigma^2} \end{bmatrix}$$

where $\mathcal{S}_{n,j}(\theta)$ denotes the jth element of $\mathcal{S}_n(\theta)$. Hence

$$\mathcal{H}_n(\theta) = -\begin{bmatrix} \dfrac{n}{\sigma^2} & \dfrac{1}{\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu) \\ \dfrac{1}{\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu) & -\dfrac{n}{2\sigma^4} + \dfrac{1}{\sigma^6}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2 \end{bmatrix}$$

It is still the case that $\mathcal{I}_n(\theta) = -\operatorname{E}[\mathcal{H}_n(\theta)]$. We shall adopt a slightly modified notation for the information matrix. Specifically, while we shall write

$$\mathcal{I}_n(\theta) = \sum_{i=1}^{n} \mathcal{I}_n(\theta, y_i) = \sum_{i=1}^{n} \imath(\theta; y_i).$$

In the event that the data are identically distributed, so that $\imath(\theta; y_1) = \cdots = \imath(\theta; y_n)$, we shall represent the information in a single observation by $\imath(\theta)$, in which case $\mathcal{I}_n(\theta) = n\imath(\theta)$. Returning to our example, on noting that

$$\operatorname{E}\left[\sum_{i=1}^{n}(y_i-\mu)\right] = 0 \quad \text{and} \quad \operatorname{E}\left[\sum_{i=1}^{n}(y_i-\mu)^2\right] = n\sigma^2,$$

we obtain

$$\mathcal{I}_n(\theta) = \begin{bmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & -\dfrac{n}{2\sigma^4} + \dfrac{n}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{bmatrix}$$

Observe that $\mathcal{I}_n(\theta)$ is block diagonal, which implies that $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are independent. Next we see that

$$(\mathcal{I}_n(\theta))^{-1} = \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{bmatrix}$$

Recall that $(\mathcal{I}_n(\theta))^{-1}$ is used as the variance in the asymptotic Normal approximation to the sampling distribution of $\hat{\theta}_n$. If nothing else, a variance of σ^2/n for $\bar{y}_n$ should look familiar. Note that, in this case, $\bar{y}_n \sim N(\mu, \sigma^2/n)$ is the exact distribution, however, this is not the case for $\hat{\sigma}_n^2$. Finally, if you are really keen you might show that the score cannot be factored in a way that allows $\hat{\sigma}_n^2$ to attain the CRLB, although the score for $\bar{y}_n$ can be as we saw in (5).

2.2 Testing

We shall ignore the case of testing a simple null against a simple alternative because it has already served it's purpose of illustrating the ideas but is not otherwise of great practical importance. Instead we will start with the problem of testing a simple null against a composite alternative, namely $H_0 : \mu = \mu_0, \sigma^2 > 0$ against $H_1 : \mu \neq \mu_0, \sigma^2 > 0$.[35]

2.2.1 The Likelihood Ratio Test

Recall that $LR = -2 \ln \hat{\lambda}$, where $\hat{\lambda} = \mathcal{L}_n(\theta_0)/\mathcal{L}_n(\hat{\theta}_n)$. Now,

$$\mathcal{L}_n(\theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

To obtain the maximized value of the likelihood function at θ_0 we set $\hat{\mu}_{n,0} = \mu_0$ and then maximize over the one remaining free parameter, namely σ^2, to obtain $\hat{\sigma}^2_{n,0}$. That is, we restrict our estimator of the mean to take the hypothesized value μ_0 and then estimate our free parameter, σ^2, subject to this restriction on the mean.From the first-order condition $S(\hat{\theta}_{n,0}) = 0$, we obtain

$$\begin{aligned} 0 &= \frac{\partial \ln \mathcal{L}_n(\hat{\theta}_{n,0})}{\partial \sigma^2} \\ &= -\frac{n}{2\hat{\sigma}^2_{n,0}} + \frac{1}{2\hat{\sigma}^4_{n,0}}\sum_{i=1}^{n}(y_i - \hat{\mu}_{n,0})^2 \\ \Rightarrow \hat{\sigma}^2_{n,0} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{n,0})^2 \end{aligned}$$

If we substitute $\hat{\sigma}^2_{n,0}$ into $\mathcal{L}_n(\theta)$ in place of σ^2 we obtain the maximized value of the likelihood function under H_0, $\mathcal{L}_n(\theta_0)$, which is

$$\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{n,0})^2\right)\right)^{-n/2} \exp\left\{-\frac{n\sum_{i=1}^{n}(y_i - \hat{\mu}_{n,0})^2}{2\sum_{i=1}^{n}(y_i - \hat{\mu}_{n,0})^2}\right\}$$
$$= \left(\frac{n}{2\pi\sum_{i=1}^{n}(y_i - \mu_0)^2}\right)^{n/2} e^{-n/2}.$$

The final expression on the right-hand side is the outcome when we replace the restricted estimator for the mean, $\hat{\mu}_{n,0}$, by the hypothesized value μ_0. We have already seen that the unrestricted mles are

$$\hat{\mu}_n = \bar{y}_n \qquad \text{and} \qquad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2.$$

[35] You can restrict as many parameters as you like, and leave as many parameters unrestricted as you wish, and still have a simple hypothesis provided that the restricted parameters can only take a single value. Hence H_0 is a simple hypothesis because the restricted parameter μ can only take a single value μ_0, even though σ^2 is left unrestricted. H_1 is a composite hypothesis because the parameter restriction $\mu \neq \mu_0$ leaves μ free to take an infinite number of possible values.

Making these substitutions into $\mathcal{L}_n(\theta)$ yields the maximized value of the unrestricted likelihood function. We obtain

$$\left(\frac{n}{2\pi\sum_{i=1}^n(y_i-\bar{y}_n)^2}\right)^{n/2}e^{-n/2}.$$

Combining these results we see that

$$\hat{\lambda}=\left(\frac{\sum_{i=1}^n(y_i-\bar{y}_n)^2}{\sum_{i=1}^n(y_i-\mu_0)^2}\right)^{n/2}.$$

Notice that the Likelihood Ratio has reduced to a ratio of variances, or a ratio of sums of squared deviations. This is a common feature of Likelihood Ratio tests. Again using the device that

$$\sum_{i=1}^n(y_i-\mu_0)^2=\sum_{i=1}^n(y_i-\bar{y}_n)^2+n(\bar{y}_n-\mu_0)^2.$$

we obtain

$$\hat{\lambda}=\left(\frac{\sum_{i=1}^n(y_i-\bar{y}_n)^2}{\sum_{i=1}^n(y_i-\bar{y}_n)^2+n(\bar{y}_n-\mu_0)^2}\right)^{n/2} \tag{24}$$

$$=\left(\frac{1}{1+n(\bar{y}_n-\mu_0)^2/\sum_{i=1}^n(y_i-\bar{y}_n)^2}\right)^{n/2}. \tag{25}$$

Clearly, $\hat{\lambda}$ will be small for large values of $n(\bar{y}_n-\mu_0)^2/\sum_{i=1}^n(y_i-\bar{y}_n)^2$, or for large values of $n(n-1)(\bar{y}_n-\mu_0)^2/\sum_{i=1}^n(y_i-\bar{y}_n)^2$. But

$$\frac{n(n-1)(\bar{y}_n-\mu_0)^2}{\sum_{i=1}^n(y_i-\bar{y}_n)^2}=\left(\frac{\bar{y}_n-\mu_0}{\sqrt{\sum_{i=1}^n(y_i-\bar{y}_n)^2/[n(n-1)]}}\right)^2=\left(\frac{\bar{y}_n-\mu_0}{\sqrt{s_n^2/n}}\right)^2=t^2,$$

where $t=(\bar{y}_n-\mu_0)/\sqrt{s_n^2/n}\overset{H_0}{\sim}t_{n-1}$, and so $t^2=F\overset{H_0}{\sim}F_{1,(n-1)}$. Therefore, the Likelihood Ratio test will reject H_0 for large values of F, i.e. $F>F_{1,(n-1),1-\alpha}$, where $F_{1,(n-1),1-\alpha}$ is the critical value that cuts off an upper-tail probability of α in an $F_{1,(n-1)}$ distribution. Note that as $n\to\infty$, $F_{1,(n-1)}\to\chi_1^2$, which is exactly the asymptotic approximation to the distribution of the Likelihood Ratio test.[36]

As a final remark on this problem observe that, even though we have allowed θ to be a $p\times 1$ vector, we have still only tested a single restriction. As a consequence, we have the result $LR\underset{a}{\overset{H_0}{\sim}}\chi_1^2$. Of course, with p parameters we may wish to test more restrictions. Suppose that we wish to test $1\leq j\leq p$ restrictions simultaneously. Then the relevant approximation is $LR\underset{a}{\overset{H_0}{\sim}}\chi_j^2$. This last is a very important result.

[36] If we return to (25) then

$$LR=-2\ln\hat{\lambda}=n\ln\left(1+\frac{n(\bar{y}_n-\mu_0)^2}{\sum_{i=1}^n(y_i-\bar{y}_n)^2}\right)=n\ln\left(1+\frac{t^2}{(n-1)}\right).$$

As before we see that LR is large when t^2 is large and so our decision rule will be to reject for large values of t^2. Finally, recall the result that $\ln(1+x)\approx x$ for x small. Observe that as $n\to\infty$, $t^2/(n-1)$ becomes small. Therefore

$$n\ln\left(1+\frac{t^2}{(n-1)}\right)\approx\frac{nt^2}{(n-1)}\approx t^2.$$

Clearly, the approximation $LR\underset{a}{\overset{H_0}{\sim}}\chi_1^2$ implies that $t^2\underset{a}{\overset{H_0}{\sim}}\chi_1^2$ as we have seen before.

2.2.2 The Wald Test

In the case of single parameter, the relevant statistic was

$$W = \mathcal{I}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \overset{H_0}{\underset{a}{\sim}} \chi^2_1.$$

When there is more than one parameter, the structure of the test is a little more complicated, but not much. I will first give a formal description of W and then go back and look at what it all means.

Suppose that we wish to test a set of restrictions of the form $H_0 : R\theta - r = 0$ against $H_1 : R\theta - r \neq 0$ where θ is a $p \times 1$ vector of parameters, R is a $j \times p$ matrix with full row rank, and r is a $j \times 1$ vector. Neither R nor r depend on θ. The idea of the Wald test is to evaluate $R\theta - r$ at $\hat{\theta}_n$, the unrestricted mle, and test whether or not $R\hat{\theta}_n - r$ is sufficiently close to zero to believe that H_0 is true. Noting that

$$\text{Var}\left[R\hat{\theta}_n - r\right] = \text{Var}\left[R\hat{\theta}_n\right] = R\,\text{Var}\left[\hat{\theta}_n\right]R' = R(\mathcal{I}_n(\hat{\theta}_n))^{-1}R',$$

an obvious statistic is[37]

$$W = (R\hat{\theta}_n - r)'\left(R(\mathcal{I}_n(\hat{\theta}_n))^{-1}R'\right)^{-1}(R\hat{\theta}_n - r) \overset{H_0}{\underset{a}{\sim}} \chi^2_j.$$

In order to understand what is meant by the preceding description of the Wald test it is necessary to understand the way the hypotheses have been written. For example, suppose that $\theta = [\theta_1, \theta_2]'$ and that we wish to test

$$H_0 : \theta_1 + 2\theta_2 = 3, \theta_1 - \theta_2 = 0.$$

against the alternative hypothesis that at least one of these parameter restrictions is satisfied. Then in matrix notation we could write the null hypothesis as[38]

$$\begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The matrix R contains the coefficients of the various (linear) restrictions that we wish to test and $r = [3, 0]'$ is the vector of constants. The assumption that R has full row rank is simply a way of saying that the restrictions are (linearly) distinct, so that we are not trying to impose the same restriction more that once. For example, $\theta_1 = 1$ and $2\theta_1 = 2$ are two versions of a single restriction rather than two distinct restrictions.

[37] Note that the generic structure here is $Z'(\text{Var}[Z])^{-1}Z$, where Z is (at least approximately) distributed as a Normal random variable with mean μ (say) and covariance matrix $\text{Var}[Z]$. The statistic will behave like $\mu'(\text{Var}[Z])^{-1}\mu \geq 0$. Under H_0, $\mu = 0$ and so we would expect the test statistic to take small values when H_0 is true. Equally, if H_0 is not true, so that $\mu \neq 0$ then we would expect the test statistic to take values further away from zero. Consequently, the decision rule is to reject H_0 for big values of the statistic, which is approximately chi-squared under the null. This is a very commonly occurring structure in econometrics.

[38] As an aside, we might solve these equations directly as

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}^{-1}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \frac{1}{1 \times (-1) - 2 \times 1}\begin{bmatrix} -1 & -2 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

So an alternative statement of the null hypothesis is $H_0 : \theta_1 = 1, \theta_2 = 1$. Of course, it won't necessarily be so easy to simplify the restrictions if there are fewer restrictions than parameters.

If we think about our problem of testing $H_0 : \mu - \mu_0 = 0$ while leaving σ^2 unrestricted then $\theta = [\mu, \sigma^2]$, $R = [1, 0]$ and $r = [\mu_0]$. In this example $j = 1$ and $p = 2$. Observe that, because

$$\mathcal{I}_n(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \Rightarrow \mathcal{I}_n(\hat{\theta}_n) = \begin{bmatrix} \frac{n}{\hat{\sigma}_n^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{bmatrix},$$

$R\mathcal{I}_n(\hat{\theta}_n)^{-1}R' = \hat{\sigma}_n^2/n$, which is that block of $\mathcal{I}_n(\hat{\theta}_n)^{-1}$ corresponding to the variance of $\hat{\mu}_n$. Hence,

$$W = (\hat{\mu}_n - \mu_0)'(\hat{\sigma}_n^2/n)^{-1}(\hat{\mu}_n - \mu_0) = \frac{n(\bar{y}_n - \mu_0)^2}{\hat{\sigma}_n^2} \underset{a}{\overset{H_0}{\sim}} \chi_1^2.$$

Note that, from our discussion of LR in this case, $(n-1)W/n \sim F_{1,(n-1)}$, and so, if we wanted to use an exact result rather than the asymptotic approximation then the appropriate size-α decision rule is to reject H_0 if

$$W > \frac{n}{n-1} F_{1,(n-1),(1-\alpha)}.$$

Observe that, as $n \to \infty$, $n/(n-1) \to 1$ and $F_{1,(n-1)} \to \chi_1^2$, which is the asymptotic approximation. Note that $F_{1,(n-1),(1-\alpha)}$ is that critical value which cuts off a $100(1-\alpha)\%$ lower tail probability in an $F_{1,(n-1)}$ distribution or a $100\alpha\%$ upper tail probability in the same distribution.

2.2.3 The Lagrange Multiplier Test

When there are p parameters the LM test takes the form

$$LM = \mathcal{S}_n(\theta_0)'(\mathcal{I}_n(\theta_0))^{-1}\mathcal{S}_n(\theta_0) \underset{a}{\overset{H_0}{\sim}} \chi_j^2,$$

where $1 \leq j \leq p$ is the number of restrictions being tested and θ_0 is the mle under H_0. We will see that there are certain features of the statistic that arise in regression problems but for now we will simply complete our example. For the final time we wish to test $H_0 : \mu = \mu_0$ against a two sided alternative while leaving σ^2 unrestricted. We have seen that the score is

$$\mathcal{S}_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \mu)^2 \end{bmatrix}$$

and the information matrix is

$$\mathcal{I}_n(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Furthermore, we have also seen that the restricted mles are $\hat{\mu}_{n,0} = \mu_0$ and

$$\hat{\sigma}_{n,0}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \mu_0)^2.$$

If we now evaluate the score and information matrices at the restricted mles we get

$$\mathcal{S}_n(\hat{\theta}_{n,0}) = \begin{bmatrix} \frac{1}{\hat{\sigma}^2_{n,0}} \sum_{i=1}^{n} (y_i - \mu_0) \\ -\frac{n}{2\hat{\sigma}^2_{n,0}} + \frac{1}{2\hat{\sigma}^4_{n,0}} \sum_{i=1}^{n} (y_i - \mu_0)^2 \end{bmatrix} = \begin{bmatrix} \frac{n(\bar{y}_n - \mu_0)}{\hat{\sigma}^2_{n,0}} \\ -\frac{n}{2\hat{\sigma}^2_{n,0}} + \frac{n\hat{\sigma}^2_{n,0}}{2\hat{\sigma}^4_{n,0}} \end{bmatrix} = \begin{bmatrix} \frac{n(\bar{y}_n - \mu_0)}{\hat{\sigma}^2_{n,0}} \\ 0 \end{bmatrix}$$

and

$$\mathcal{I}_n(\hat{\theta}_{n,0}) = \begin{bmatrix} \frac{n}{\hat{\sigma}^2_{n,0}} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4_{n,0}} \end{bmatrix}.$$

Combining these results the LM test becomes

$$LM = \begin{bmatrix} \frac{n(\bar{y}_n - \mu_0)}{\hat{\sigma}^2_{n,0}}, & 0 \end{bmatrix} \begin{bmatrix} \frac{n}{\hat{\sigma}^2_{n,0}} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4_{n,0}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{n(\bar{y}_n - \mu_0)}{\hat{\sigma}^2_{n,0}} \\ 0 \end{bmatrix} = \frac{n(\bar{y}_n - \mu_0)^2}{\hat{\sigma}^2_{n,0}} \underset{a}{\overset{H_0}{\sim}} \chi^2_1.$$

A number of comments are in order. First, the LM is very similar to the LR and Wald tests, however, the denominator contains $\hat{\sigma}^2_{n,0}$ rather than $\hat{\sigma}^2_n$, and so the statistics are different. Of course, if H_0 is true then the difference between these two variance estimators is asymptotically irrelevant and so all three statistics have the same asymptotic distribution under the null. Second, the fact that we have a zero in the restricted value of the score is not surprising. If you evaluate any score equation at mles obtained from it then the equation will yield a value of zero. In our example μ is not estimated, its value is dictated by H_0. However, σ^2 is estimated, conditional on the value μ_0. Consequently, if you evaluate the score equation corresponding to σ^2 at the values $\hat{\theta}_{n,0} = [\mu_0, \hat{\sigma}^2_0]'$ it must yield zero by construction. Finally, the fact that, in this case $\mathcal{I}_n(\theta)$ is block diagonal has given LM a fairly simple form. This is not always the case, although in regression models, the regression coefficient estimates tend to be (asymptotically) independent of variance estimates.

3 Maximum Likelihood As GMM

We won't visit re-visit all of maximum likelihood here but, one way of specifying the maximum likelihood estimators is to note that, in a correctly specified model, the score has zero expectation. That is, $\mathrm{E}[\mathcal{S}(\theta)] = 0$. Immediately we can estimate θ by matching the relevant sample moments to the set of population moments implicit in this moment condition. We note that we will have an exactly identified estimator, so no excess moment conditions but, all of a sudden, a wonderful vista has opened up. From the optimality properties of the mles we know know about optimal GMM estimation. We have a theory of hypothesis testing that can be ported directly from the maximum likelihood environment into that of GMM. In models where we have more moment conditions than parameters to estimate, so that the model is over-identified, some modification of these testing procedures may be required but the foundations are all there.

Bibliography

Aitchison, J. and S. D. Silvey (1958). Maximum-likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics* 29, 813–828. 42, 51

Andrews, D. W. K. and J. H. Stock (2007). Inference with weak instruments. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume III*, R. Blundell, W. K. Newey, and T. Persson, editors, 122–173, Cambridge University Press, Cambridge. 21

Box, J. F. (1978). *R.A. Fisher: The Life of a Scientist*. Wiley, New York. 40

Breusch, T. S. and A. R. Pagan (1979). The Lagrange multiplier test and its application to model specification in econometrics. *Review of Economic Studies* 47, 239–253. 17

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician* 36(3, Part 1), 153–157. 19

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London. 13, 53

Cramer, J. S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge University Press, Cambridge. 5, 42

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics* 36(4), 767–808. 21

Gouriéroux, C. and A. Monfort (1995). *Statistics and Econometric Models*, volume 1. Cambridge University Press. 42

Hahn, J. and J. Hausman (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *The American Economic Review: Papers and Proceedings* 93(2), 118–125. 21

Hirschberg, J. G. and J. N. Lye (2005). Inferences for the extremum of quadratic regression models. Working Paper Number 906, Department of Economics, The University of Melbourne. 21

Hurzurbazar, V. S. (1948). The likelihood equation, consistency, and the maxima of the likelihood function. *Annals of Eugenics* 14, 185–200. 38

King, M. L. (1987). Towards a theory of point optimal testing. *Econometric Reviews* 6, 169–218. 13

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Springer-Verlag, New York, second edition. 53

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London. 43

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., New York, third edition. 53

Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A* 231, 289–337. Reprinted in Neyman and Pearson (1967). 13, 53

Neyman, J. and E. S. Pearson (1967). *Joint Statistical Papers*. Cambridge University Press. 33

Pagan, A. R. (1982). Estimation and control of linear econometric models. *IHS-Journal: Zeitschrift des Instituts für Höhere Studien — Wien (Institute for Advanced Studies, Vienna)* 6(4), 247–268. 22

Poskitt, D. S. and C. L. Skeels (2008). Conceptual frameworks and experimental design in simultaneous equations. *Economics Letters* 100, 138–142. 9

Poskitt, D. S. and C. L. Skeels (2013). Inference in the presence of weak instruments: A selected survey. *Foundations and Trends® in Econometrics* 6(1), 1–99, ISSN 1551-3076, doi:10.1561/0800000017. 21

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* 44, 50–57. 17, 51

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, Inc., New York, second edition. 41, 43

Rao, C. R. and S. K. Mitra (1971). *Generalized Inverse of Matrices and Its Applications*. John Wiley & Sons, Inc., New York. 43

Schmidt, P. (1976). *Econometrics*. Marcel Dekker, New York. 21

Searle, S. R. (1982). *Matrix Algebra Useful For Statistics*. John Wiley & Sons, New York. 43

Silvey, S. D. (1959). The Lagrangian multiplier test. *Annals of Mathematical Statistics* 30(2), 389–407. 17, 51

Silvey, S. D. (1970). *Statistical Inference*. Chapman and Hall, London. 5, 42

Stock, J. H., J. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20(4), 518–529. 21

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54(3), 426–482. 48

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20(4), 595–601. 37

Winkelmann, R. and S. Boes (2006). *Analysis of Microdata*. Springer-Verlag, Berlin. 42, 50

A Expectation and Variance of the Mean of a Simple Random Sample

For a simple random sample $y_1, \ldots, y_n$ from a population with mean μ and variance σ^2 we see that

$$\mathrm{E}\left[\bar{y}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

hence $\bar{y}_n$ is unbiased for μ. Similarly, the variance of $\bar{y}_n$ is

$$\mathrm{Var}\left[\bar{y}_n\right] = \mathrm{E}\left[\bar{y}_n^2\right] - \mathrm{E}\left[\bar{y}_n\right]^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[y_i y_j\right] - \mu^2.$$

Now, because we have a simple random sample, the covariance between any two sample draws y_i and y_j is zero, which we write $\mathrm{Cov}\left[y_i, y_j\right] = 0$, and so

$$\mathrm{E}\left[y_i y_j\right] = \begin{cases} \mathrm{Var}\left[y_i\right] + \mathrm{E}\left[y_i\right]^2 = \sigma^2 + \mu^2, & \text{if } i = j, \\ \mathrm{Cov}\left[y_i, y_j\right] + \mathrm{E}\left[y_i\right]\mathrm{E}\left[y_j\right] = \mu^2, & \text{otherwise.} \end{cases}$$

Making these substitutions we see that

$$\mathrm{Var}\left[\bar{y}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu^2 - \mu^2 = \frac{n\sigma^2}{n^2} + \frac{n^2\mu^2}{n^2} - \mu^2 = \frac{\sigma^2}{n}.$$

B The Asymptotics of Maximum Likelihood

We have already stated the key asymptotic results for the MLE and the various likelihood-based tests. In this section we provide some insight into just how these results are obtained. Because likelihood techniques are non-linear, in general, things are more complicated than seen in the context of the linear regression model, although many of the same ideas remain in play. Key to the results is establishing the consistency of the maximum likelihood estimator, which is by far the most difficult task facing us. However, once consistency is established, the limiting distribution of the estimator is relatively straight-forward to deal with. Before exploring the main properties let us start with a clear statement of the underlying assumptions and some of their implications.

B.1 Assumptions and Some Implications

We shall restrict our attention to so-called regular problems, which will be defined by the following regularity conditions.

Assumption 1 (Regularity Conditions)**.** The probability density function $\mathcal{L}(\theta; y)$ is said to be regular if

(a) The set $\mathcal{Y}$ of all values of y for which $\mathcal{L}(\theta; y)$ is strictly positive does not depend on θ.

(b) The density is a smooth function of θ such that, for all $\theta \in \Theta$ and $y \in \mathcal{Y}$, $\mathcal{L}(\theta; y)$ and $\ln \mathcal{L}(\theta; y)$ have finite-valued partial derivatives up to the third order.

(c) The variance of $\partial \ln \mathcal{L}(\theta; y)/\partial\theta$ is positive definite for all $\theta \in \Theta$.

(d) The expression $\int_{\mathcal{Y}} \mathcal{L}(\theta; y)\, \mathrm{d}y$ is twice differentiable under the integral.

Assumption 1(a) precludes the support of the density function depending on a parameter. For example, if a random variable is uniformly distributed on an interval $[0, \theta]$ for some unknown parameter θ then the problem of estimating θ would be an irregular problem.[39]

Assumption 1(b) allows us to differentiate the density function sufficiently often to take a second-order Taylor approximation, which we will need in order to establish the asymptotic normality of the mle.

Assumption 1(c) is in essence an identification assumption, allowing us to obtain a unique solution to our problem. We will make a more explicit assumption about identification below.

Finally, Assumption 1(d) allows us to derive expectations of the various quantities (score and hessians) that we will be working with.

Assumption 2 (Identification Assumption)**.** A parameter θ is said to be identifiable (estimable) if and only if $\theta_1 \neq \theta_2$ implies, for some value of y, that $\mathcal{L}(\theta_1; y) \neq \mathcal{L}(\theta_2; y)$ for almost all $y \in \mathcal{Y}$.

For a parameter to be estimable or identified it must be the case that distinct parameter values lead to distinct density functions. If this is not the case you will never be able to determine which parameter value led to a given density function. The 'for almost all' bit is technical jargon saying that any y not satisfying this condition is a zero probability event, meaning that, for all intents and purposes, it cannot happen.

Finally , although it is not integral to the theory of maximum likelihood, our treatment will assume that our data arise from independent sampling so that

$$\mathcal{L}(\theta; y) = \prod_{j=1}^{n} \mathcal{L}(\theta; y_j).$$

Given these assumptions we can then define the maximum as follows.

Definition 1. The Maximum Likelihood Estimator.
A maximum likelihood estimate $\hat{\theta} = \hat{\theta}(y)$ is an element of the parameter space Θ such that $\mathcal{L}(\hat{\theta}; y) \geq \mathcal{L}(\theta; y)$ for all $\theta \in \Theta$.

Given our regularity conditions, and provided that $\mathcal{L}(\theta; y)$ is not maximized on a boundary value of Θ, $\hat{\theta}$ will be a solution to $\mathcal{S}(\theta; y) = 0$.[40]

Before looking at the asymptotic properties of mles, there are a couple of implications that follow immediately from our assumptions. First, when thought of as a function of

[39] In practice you would probably estimate θ using an order statistic, specifically, your best guess is probably the largest observation in your sample.

[40] Observe that there may be multiple solutions to this equation and that some of these may be correspond to local maxima rather than a global maximum. (Strictly we need to check second-order conditions to ensure that our solutions correspond to maxima rather than other sorts of stationary points.) In summary, all of the problems associated with the use of calculus to find a global maximum apply here too. We shall abstract away from such concerns here.

the data, $\mathcal{L}(\theta; y)$ is a density function. Hence, it follows that[41]

$$\int_{\mathcal{Y}} \mathcal{L}(\theta; y)\, \mathrm{d}y = 1. \tag{26}$$

If we differentiate both sides of this equation with respect to θ we obtain

$$\frac{\partial}{\partial \theta} \int_{\mathcal{Y}} \mathcal{L}(\theta; y)\, \mathrm{d}y = \frac{\partial 1}{\partial \theta} \implies \int_{\mathcal{Y}} \frac{\partial \mathcal{L}(\theta; y)}{\partial \theta}\, \mathrm{d}y = 0.$$

by Assumption 1(d). Noting that

$$\frac{\partial \ln \mathcal{L}(\theta; y)}{\partial \theta} = \frac{1}{\mathcal{L}(\theta; y)} \frac{\partial \mathcal{L}(\theta; y)}{\partial \theta},$$

we see that

$$\frac{\partial \mathcal{L}(\theta; y)}{\partial \theta} = \frac{\partial \ln \mathcal{L}(\theta; y)}{\partial \theta} \mathcal{L}(\theta; y), \tag{27}$$

and hence that[42]

$$\int_{\mathcal{Y}} \frac{\partial \mathcal{L}(\theta; y)}{\partial \theta}\, \mathrm{d}y = \int_{\mathcal{Y}} \frac{\partial \ln \mathcal{L}(\theta; y)}{\partial \theta} \mathcal{L}(\theta; y)\, \mathrm{d}y = \int_{\mathcal{Y}} \mathcal{S}(\theta; y) \mathcal{L}(\theta; y)\, \mathrm{d}y = \mathrm{E}\left[\mathcal{S}(\theta; y)\right] = 0. \tag{28}$$

That is, in regular problems, the expected value of the score is zero. Moreover, on taking the transpose of (28) and letting $\mathbf{0}$ denote a $p \times p$ matrix of zeroes, we obtain

$$\frac{\partial}{\partial \theta} \int_{\mathcal{Y}} \mathcal{S}(\theta; y)' \mathcal{L}(\theta; y)\, \mathrm{d}y = \frac{\partial}{\partial \theta} 0' = \mathbf{0}.$$

Note that there is no transpose attached to $\mathcal{L}(\theta; y)$ because it is a probability density function and hence scalar. Looking more closely at the left-hand side of this expression

$$\begin{aligned}
\frac{\partial}{\partial \theta} \int_{\mathcal{Y}} \mathcal{S}(\theta; y)' \mathcal{L}(\theta; y)\, \mathrm{d}y &= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \mathcal{S}(\theta; y)' \mathcal{L}(\theta; y)\, \mathrm{d}y \\
&= \int_{\mathcal{Y}} \left\{ \left[\frac{\partial}{\partial \theta} \mathcal{S}(\theta; y)' \right] \mathcal{L}(\theta; y) + \left[\frac{\partial}{\partial \theta} \mathcal{L}(\theta; y) \right] \mathcal{S}(\theta; y)' \right\} \mathrm{d}y \\
&= \int_{\mathcal{Y}} \mathcal{H}(\theta; y) \mathcal{L}(\theta; y)\, \mathrm{d}y + \int_{\mathcal{Y}} \mathcal{S}(\theta; y) \mathcal{S}(\theta; y)' \mathcal{L}(\theta; y)\, \mathrm{d}y \\
&= \mathrm{E}\left[\mathcal{H}(\theta; y)\right] + \mathrm{E}\left[\mathcal{S}(\theta; y) \mathcal{S}(\theta; y)'\right].
\end{aligned}$$

That is,

$$\begin{aligned}
\mathrm{E}\left[\mathcal{H}(\theta; y)\right] + \mathrm{E}\left[\mathcal{S}(\theta; y) \mathcal{S}(\theta; y)'\right] &= \mathbf{0} \\
\mathrm{E}\left[\mathcal{S}(\theta; y) \mathcal{S}(\theta; y)'\right] &= -\mathrm{E}\left[\mathcal{H}(\theta; y)\right] = \mathcal{I}(\theta)
\end{aligned} \tag{29}$$

Equation (29) is the so-called *information matrix equality*. It tells us that the information matrix may be written as either minus the expected hessian or the expectation of the

[41] In this sections I am not going to swap between the notations of $f(y; \theta)$ denoting a density function and $\mathcal{L}(\theta; y)$ denoting the likelihood function. The functions are identical, differing only in interpretation, and it will just be easier to stick with the latter notation for now.

[42] Note that the zero at the right-hand side of this expression is a $p \times 1$ vector of zeroes, where p is the dimension of θ.

outer product of the score vector. Because the score is a derivative it is also called a gradient. This leads to the expected outer product of the scores also being called the outer product of the gradient (OPG) form of the information matrix. The significance of the result is that we need the scores in order to calculate the mle and so we can also obtain the information matrix without further differentiation. Computationally this is often very convenient and is particularly important in the calculation of LM test statistics, which is explored further in the OPG handout. That said, estimates of the information matrix based on the OPG form tend to not be as good as those based on the expected value of the hessian.

B.2 Consistency of the MLE

Given the assumptions of the preceding section, we wish to prove the following result.

Theorem 1 (Consistency of the MLE)**.** *If $\hat{\theta}_n$ is the maximum likelihood estimator satisfying $\ln \mathcal{L}(\hat{\theta}_n; y) \geq \ln \mathcal{L}(\theta; y)$ for all n and for every $\theta \in \Theta$, then $\hat{\theta}_n \underset{p}{\to} \theta_0$ where θ_0 is the true parameter value. That is, the maximum-likelihood estimator is consistent.*

A complete proof of this result is difficult and so we shall only sketch the result originally due to Wald (1949). The key to the proof is to establish that the expectation of $\ln \mathcal{L}(\theta; y)$ is maximized by the true parameter value θ_0. Once we have this then consistency follows directly. The proof itself is a two step process. The first step is to establish that

$$\mathrm{E}_0 \left[\ln \mathcal{L}(\theta_0; y) \right] \geq \mathrm{E}_0 \left[\ln \mathcal{L}(\theta; y) \right],$$

where the notation $\mathrm{E}_0 [\cdot]$ indicates that the expectation is taken with respect to the true density, i.e. the density evaluated at θ_0, which is $\mathcal{L}(\theta_0; y)$. The second step is then to establish that

$$\operatorname*{plim}_{n \to \infty} n^{-1} \ln \mathcal{L}(\theta_0; y) \geq \operatorname*{plim}_{n \to \infty} n^{-1} \ln \mathcal{L}(\theta; y).$$

for all $\theta \in \Theta$ and, in particular for $\theta = \hat{\theta}$, which must also be an element of Θ. The conclusion is reached on recognizing that these two inequalities can only be reconciled if $\hat{\theta}$ is consistent for θ. Having laid out or objective, let us establish the result.

To begin, consider the expectation

$$\mathrm{E}_0 \left[- \ln \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\theta_0; y)} \right].$$

Because the natural logarithm is a strictly concave function, this is the expectation of a strictly convex function (as a consequence of the minus sign) and so, by Jensen's inequality,

$$\mathrm{E}_0 \left[- \ln \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\theta_0; y)} \right] \geq - \ln \mathrm{E}_0 \left[\frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\theta_0; y)} \right]$$

where equality arises if and only if $\theta = \theta_0$. But notice that

$$\mathrm{E}_0 \left[\frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\theta_0; y)} \right] = \int \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\theta_0; y)} \mathcal{L}(\theta_0; y) \, \mathrm{d}y = 1,$$

as a consequence of (26). Hence

$$\mathrm{E}_0\left[-\ln\frac{\mathcal{L}(\theta;y)}{\mathcal{L}(\theta_0;y)}\right] \geq -\ln 1 = 0.$$

We can rearrange this expression to obtain

$$\mathrm{E}_0\left[-\ln\frac{\mathcal{L}(\theta;y)}{\mathcal{L}(\theta_0;y)}\right] = \mathrm{E}_0\left[\ln\mathcal{L}(\theta_0;y) - \ln\mathcal{L}(\theta;y)\right] = \mathrm{E}_0\left[\ln\mathcal{L}(\theta_0;y)\right] - \mathrm{E}_0\left[\ln\mathcal{L}(\theta;y)\right] \geq 0.$$

which, in turn, implies that

$$\mathrm{E}_0\left[\ln\mathcal{L}(\theta_0;y)\right] \geq \mathrm{E}_0\left[\ln\mathcal{L}(\theta;y)\right], \tag{30}$$

which completes the first step of the proof.

The second step of the proof proceeds as follows. Given our assumption of independent sampling, for any value of θ we see that

$$n^{-1}\ln\mathcal{L}(\theta;y) = n^{-1}\sum_{j=i}^{n}\ln\mathcal{L}(\theta;y_j).$$

is the sample mean of independent random variables with expectation $n^{-1}\mathrm{E}_0\left[\ln\mathcal{L}(\theta;y)\right]$ for all n. Hence, by the Law of Large Numbers

$$n^{-1}\ln\mathcal{L}(\theta;y) \xrightarrow{p} \lim_{n\to\infty} n^{-1}\mathrm{E}_0\left[\ln\mathcal{L}(\theta;y)\right]. \tag{31}$$

In particular, on setting $\theta = \theta_0$, we see that

$$n^{-1}\ln\mathcal{L}(\theta_0;y) \xrightarrow{p} \lim_{n\to\infty} n^{-1}\mathrm{E}_0\left[\ln\mathcal{L}(\theta_0;y)\right] \geq \lim_{n\to\infty} n^{-1}\mathrm{E}_0\left[\ln\mathcal{L}(\theta;y)\right] = \operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\theta;y),$$

where the inequality follows from (30). That is,

$$\operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\theta_0;y) \geq \operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\theta;y), \tag{32}$$

with equality if and only if $\theta = \theta_0$.

However, by definition, for any n, the maximum likelihood estimator is that value $\hat{\theta}$ such that $n^{-1}\ln\mathcal{L}(\hat{\theta};y) \geq n^{-1}\ln\mathcal{L}(\theta;y)$ and, in particular,

$$\operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\hat{\theta};y) \geq \operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\theta_0;y). \tag{33}$$

The inequalities (32) and (33) are in conflict unless

$$\operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\theta_0;y) = \operatorname*{plim}_{n\to\infty} n^{-1}\ln\mathcal{L}(\hat{\theta};y),$$

which implies, by what we called the Continuous Mapping Theorem in the Asymptotics handout, $\hat{\theta} \xrightarrow{p} \theta_0$.

Two final comments. First, given that we have ended up working with probability limits, we have implicitly used a Weak Law of Large Numbers. We might equally have applied a Strong Law of Large Numbers to obtain almost sure convergence, which is a stronger result; that is $\hat{\theta} \xrightarrow{a.s.} \theta_0$. Second, we haven't addressed the question of what to do if there is more than one solution to the score equation ($\mathcal{L}(\theta;y) = 0$). Hurzurbazar (1948) has shown that, under certain regularity conditions a unique consistent estimator emerges in the sense that if there are, say, two contenders $\hat{\theta}_1$ and $\hat{\theta}_2$ then $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2) \xrightarrow{a.s.} 0$, making the problem of multiple solutions a small-sample problem.

B.3 Asymptotic Normality

Having established consistency, establishing a Normal asymptotic distribution in regular problems is relatively straight-forward. Central to the development is the multivariate Taylor theorem presented in Appendix A.2.2 of the Asymptotics handout. Consequently, in addition to the assumptions underlying our proof of consistency, we will make such assumptions as required for a Taylor expansion to be valid. Specifically, we will require that over a domain $\mathcal{D}$ the log-likelihood function is twice continuously differentiable,[43] where $\mathcal{D}$ is an interval entirely containing the line segment joining θ_0 and $\hat{\theta}_n$, for all values of n. This implies (i) continuity in θ of the log-likelihood function and its first two derivatives, and (ii) that θ_0 is not a boundary point of Θ. We shall also assume that both the Hessian and its expectation with respect to $f(y;\theta_0)$ are non-singular. Given that we have now assumed differentiability of the log-likelihood function, the solution to the problem

$$\underset{\theta\in\Theta}{\operatorname{argmax}}\; n^{-1}\ln\mathcal{L}_n(\theta)$$

is obtained by solving the first order conditions[44]

$$n^{-1}\mathcal{S}_n(\hat{\theta}_n;y) = \left.\frac{\partial n^{-1}\ln\mathcal{L}_n(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_n} = 0. \tag{34}$$

If we expand $n^{-1}\mathcal{S}_n(\hat{\theta}_n;y)$ about $n^{-1}\mathcal{S}_n(\theta_0;y)$ then we obtain[45]

$$n^{-1}\mathcal{S}_n(\hat{\theta}_n;y) = n^{-1}\mathcal{S}_n(\theta_0;y) + \left[n^{-1}\frac{\partial\mathcal{S}_n(\theta;y)}{\partial\theta'}\right]_{\theta=\theta_n^*}(\hat{\theta}_n-\theta_0), \tag{35}$$

where $\theta_n^* = \lambda_n\hat{\theta}_n + (1-\lambda_n)\theta_0$, $0<\lambda_n<1$, lies on the line segment joining the points θ_0 and $\hat{\theta}_n$, so that $\theta_n^*\in\mathcal{D}$. By definition $\mathcal{S}_n(\hat{\theta}_n;y)=0$ and so we can rearrange the equation to obtain[46]

$$\hat{\theta}_n-\theta_0 = \left[-n^{-1}\frac{\partial\mathcal{S}_n(\theta;y)}{\partial\theta'}\right]^{-1}_{\theta=\theta_n^*} n^{-1}\mathcal{S}_n(\theta_0;y). \tag{36}$$

[43] A function with k continuous derivatives is called a C^k function. In order to specify a C^k function on a domain X, the notation $C^k(X)$ is used. The most common C^k space is C^0, the space of continuous functions, whereas C^1 is the space of continuously differentiable functions. Note that we did not assume differentiability in our proof of consistency.

[44] The carets, or hats, appear when we set the derivatives equal to zero, making them first-order conditions.

[45] This form of Taylor series expansion coincides with equation (A.3) of the Asymptotics handout.

[46] As an aside we observe that (36) provides a basis for the iterative solution of the potentially non-linear equations in (34). If we replace θ_0 and θ_n^* by θ_i and $\hat{\theta}_n$ by θ_{i+1} then (36) can be written

$$\theta_{i+1} = \theta_i - \left[\frac{\partial\mathcal{S}_n(\theta;y)}{\partial\theta'}\right]^{-1}_{\theta=\theta_i}\mathcal{S}_n(\theta_i;y).$$

It is clear that, given some starting value θ_1, we can generate a sequence of revised estimates θ_2, θ_3, ... until such time as we believe that the sequence has converged to some limit point which is then our maximum likelihood estimate $\hat{\theta}_n$. There are different criteria that we might use to reach such a decision, including (i) $|\theta_{i+1}-\theta_i|<\epsilon_1$ for some arbitrarily small ϵ_1, or perhaps (ii) $\mathcal{S}_n(\theta_i;y)'\mathcal{S}_n(\theta_i;y)<\epsilon_2$ for some arbitrarily small ϵ_2, which implies that the score is close to zero. In practice, one might check both of these criteria and claim convergence if either is satisfied. It would also be sensible to impose a limit on the maximum number of iterations i (often something like $i\le 25$) to ensure that your computer does not sink into an infinite loop (and hence refuse to speak to you again) if it is unable to attain converge. A lack of convergence might occur if $\partial\mathcal{S}_n(\theta_i)/\partial\theta'$ is singular or close to singular which implies that the log-likelihood function is very flat in the neighbourhood of θ_i making it difficult to find a maximum.

The broad outline of what follows is that we will appeal to a law of large numbers to show that

$$\left[-n^{-1}\frac{\partial \mathcal{S}_n(\theta; y)}{\partial \theta'}\right]^{-1}_{\theta=\theta_n^*} = \imath_{\theta_0},$$

where $\imath_{\theta_0} = \mathrm{V}_0\left[\mathcal{S}_n(\theta_0; y_i)\right]$, so that $I(\theta_0) = n\,\imath_{\theta_0}$. We shall then appeal to a central limit theorem to show that

$$n^{-1/2}\mathcal{S}_n(\theta_0; y) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}).$$

Together these results will allow us to conclude that

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}^{-1})$$

and hence that

$$\hat{\theta}_n \underset{a}{\sim} \mathrm{N}(\theta_0, n^{-1}\imath_{\theta_0}^{-1})\,\mathrm{N}(\theta_0, [I(\theta_0)]^{-1}).$$

Let us now make these ideas concrete. First, recall the following results:[47]

1. The score has zero expectation with respect to $f(\theta_0; y)$:

$$\mathrm{E}_0\left[\mathcal{S}_n(\theta_0; y_i)\right] = 0. \tag{37}$$

2. The information matrix equality:

$$I(\theta_0) = \mathrm{V}_0\left[\mathcal{S}_n(\theta_0; y)\right] = \mathrm{E}_0\left[\mathcal{S}_n(\theta_0; y)\mathcal{S}_n(\theta_0; y)'\right] = -\mathrm{E}_0\left[H_n(\theta_0; y)\right].$$

   Note that the the quantity

$$I(\theta_0) = \sum_{i=1}^{n} \mathrm{E}_0\left[\mathcal{S}_n(\theta_0; y_i)\mathcal{S}_n(\theta_0; y_i)'\right] = \sum_{i=1}^{n} I_i(\theta_0)$$

   is Fisher's information for the entire sample,[48] with

$$I_i(\theta_0) = \mathrm{E}_0\left[\mathcal{S}_n(\theta_0; y_i)\mathcal{S}_n(\theta_0; y_i)'\right].$$

   Fisher's information for the ith observation. For identically distributed data, $I_i(\theta_0) = \imath_{\theta_0}$ (say) is constant for all observations and $I(\theta_0) = n\imath_{\theta_0}$. It can become quite confusing as to which measure of information is actually being discussed and so you need to be aware of these distinctions.

[47] These results are proved in Appendix B.1.

[48] Sir Ronald Aylmer Fisher (1890–1962) was a British statistician and geneticist. He is viewed as one of the founding fathers of both fields. In the area of statistics he made important contributions to distribution theory, maximum likelihood theory, hypothesis testing, analysis of variance (ANOVA), multivariate analysis, and experimental design. He had strong views on genetics which, it is fair to say, would not be deemed politically correct in the current day. In 1957 he moved to the University of Adelaide, where he became a Senior Research Fellow with the CSIRO, a position that he retained until his death. A biography of Fisher was penned by one of his daughters (Box, 1978), who was for a time married to the famous time series analyst and Bayesian statistician George Edward Pelham Box (1919–2013).

With these two results in hand our task is nearly done. If we consider first the term

$$\mathcal{S}_n(\theta_0; y) = \sum_{i=1}^{n} \mathcal{S}_n(\theta_0; y_i.$$

we know that each of the elements $\mathcal{S}_n(\theta_0; y_i)$ are independent and identically distributed, because the y_i are independent and identically distributed, with mean zero and covariance matrix $\imath_{\theta_0}$. These results allow us to appeal to the Lindeberg-Levy central limit theorem to establish that[49]

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \imath_{\theta_0}^{-1/2} [\mathcal{S}_n(\theta_0; y_i) \xrightarrow{d} \mathrm{N}(0, I_k)$$

or that

$$n^{-1/2} \mathcal{S}_n(\theta_0; y) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}). \tag{38}$$

Moreover, because $\hat{\theta}_n \xrightarrow{p} \theta_0$ it follows that, as $n \to \infty$, $\theta_n^* = \lambda_n \hat{\theta}_n + (1 - \lambda_n)\theta_0 \xrightarrow{p} \theta_0$, so that

$$\operatorname*{plim}_{n\to\infty} \left\{ \left[-n^{-1} \frac{\partial \mathcal{S}_n(\theta; y)}{\partial \theta'} \right]_{\theta=\theta_n^*} \right\} = \operatorname*{plim}_{n\to\infty} \left\{ \left[-n^{-1} \sum_{i=1}^{n} H_n(\theta; y_i) \right]_{\theta=\theta_n^*} \right\} = \imath_{\theta_0},$$

where the final equality follows from Khintchine's Theorem (or Kolmogorov's Strong Law of Large Numbers 1 (Theorem 7 of the Asymptotics handout). Combining these results we see that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \left[-n^{-1} \frac{\partial \mathcal{S}_n(\theta; y)}{\partial \theta'} \right]_{\theta=\theta_n^*}^{-1} n^{-1/2} \mathcal{S}_n(\theta_0; y) \\
&= \left\{ \imath_{\theta_0}^{-1} + \underbrace{\left(\left[-n^{-1} \frac{\partial \mathcal{S}_n(\theta; y)}{\partial \theta'} \right]_{\theta=\theta_n^*}^{-1} - \imath_{\theta_0}^{-1} \right)}_{\xrightarrow{p} \mathbf{0}} \right\} n^{-1/2} \mathcal{S}_n(\theta_0; y) \\
&= \imath_{\theta_0}^{-1} \underbrace{n^{-1/2} \mathcal{S}_n(\theta_0; y)}_{\xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0})} + o_p(1) \qquad (39) \\
&\xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}^{-1}). \qquad (40)
\end{aligned}$$

There is a number of important remarks that can be made about this result.

1. We see that in the limit the maximum likelihood estimator attains the Cramér-Rao Lower Bound;[50] see the example in Section B.4.1.

[49] Strictly we would need to appeal to the Multivariate CLT (Theorem 14 in the Asymptotics handout) in conjunction with Lindeberg-Levy CLT but we have demonstrated this line of argument in Section 4.1.3 of the Asymptotics handout and won't repeat it here. In the result which follows please be aware of the distinction between an identity matrix of order k, denoted I_k, and the information matrix for this problem, denoted $I(\theta_0)$.

[50] We met Cramér in Footnote 17 of the Asymptotics handout. C.R. Rao (1920 – 2023), is a statistician of Indian origin who made enormous contributions to the study of statistics. He obtained his PhD from Cambridge under the supervision of R. A. Fisher (see Footnote 48). He is known, among other things for the efficient score test (also known as the LM test), the CRLB which he developed independently of Cramér (and others), the Rao-Blackwell Theorem (which we shall encounter in the theory of estimation, an area where he has made numerous contributions. He is also an important figure in the area of multivariate analysis. One of his best known contributions is his classic text Rao (1973).

2. There is one implicit assumption that we have made in the derivation of this result, namely that $\imath_{\theta_0}$ and $H_n(\theta_0; y)$ are non-singular (invertible). If this is not the case then the multivariate Taylor expansion upon which our derivation is based will be invalid. Singularity of the information matrix corresponds to a lack of identification; Silvey (1970, Section 4.7.4) provides a good heuristic discussion of why this is so. Consequently, we might expect a lack of identification, or weakness of identification, which implies either singularity or near singularity of the information matrix, to lead to the Normal distribution providing a poor approximation to the true sampling distribution of the maximum likelihood estimator. This is, in fact, the case and is currently a topic of much interest in the literature.

Typically our ultimate purpose in exploring limiting distributional results is to obtain useful approximations that can be used in practice. It remains therefore to observe that if

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}^{-1})$$

then

$$\hat{\theta}_n \underset{a}{\sim} \mathrm{N}(\theta_0, n^{-1}\imath_{\theta_0}^{-1}) = \mathrm{N}(\theta_0, [I(\theta_0]^{-1}).$$

In practice $I(\theta_0)$ will need to be estimated and a variety of options are available; see Winkelmann and Boes (2006, Section 3.3.5) for a good discussion of this topic.

B.4 Constrained Maximum Likelihood Estimators

When looking at the LR and LM tests, we saw that restricted mles were required in order to construct the test statistics. If the restrictions implied by the null hypothesis are linear then restricted, or constrained, mles can typically be obtained by substitution on writing the restricted parameters in terms of the unrestricted parameters. Such a strategy is unlikely to be successful if the restrictions are non-linear. In this section we consider the problem of imposing arbitrary restrictions, or constraints, during maximum likelihood estimation.[51] Under our assumptions the standard approach is to form the Lagrangian

$$\ln \mathcal{L}_n(\theta) - \lambda_n' g(\theta),$$

where λ_n is a j-vector of Lagrange multipliers. The restricted maximum likelihood estimator is then the solution to the set of first-order conditions

$$\left.\frac{\partial \ln \mathcal{L}_n(\theta) - \lambda_n' g(\theta)}{\partial \theta}\right|_{\theta=\tilde{\theta}_n, \lambda_n=\tilde{\lambda}_n} = \mathcal{S}_n(\tilde{\theta}_n; y) - G(\tilde{\theta}_n)'\tilde{\lambda}_n = 0 \tag{41}$$

$$\left.\frac{\partial \ln \mathcal{L}_n(\theta) - \lambda_n' g(\theta)}{\partial \lambda}\right|_{\theta=\tilde{\theta}_n, \lambda_n=\tilde{\lambda}_n} = g(\tilde{\theta}_n) = 0 \tag{42}$$

where the $j \times p$ matrix

$$G(\tilde{\theta}_n) = \left[\left.\frac{\partial g(\theta)'}{\partial \theta}\right|_{\theta=\tilde{\theta}_n}\right].$$

[51] The definitive treatment is that of Aitchison and Silvey (1958). Good textbook treatments can be found in, inter alia, Silvey (1970), Cramer (1986) and Gouriéroux and Monfort (1995).

is assumed to have full row rank.[52] Regardless of the veracity of the statement $g(\theta_0) = 0$, (42) ensures that the restrictions are satisfied by $\tilde{\theta}_n$.

In most circumstances one might expect to find that if the restrictions that are imposed are valid then $\tilde{\theta}_n \xrightarrow{p} \theta_0$ and this can indeed be shown. Consider the steps followed in the proof of consistency of the maximum likelihood estimator in the unrestricted model. First, we established the information equality, i.e. for any sample size n the expected log-likelihood function is maximized at the true parameter value θ_0. There is nothing in our proof of that result that is in anyway affected by the veracity of the statement $g(\theta_0) = 0$ and so it is still true whether or not such restrictions hold. Second, we showed that in the limit, and as a consequence of the information equality, the average log-likelihood was maximized at θ_0 with probability one. Again, this result was established without precluding the possibility of relationships existing amongst the elements of θ_0. The final step was then to show that the unrestricted maximum likelihood estimator converged with probability one to that value of θ that maximized the expected log-likelihood, namely θ_0. It is this last step that we need to think about because in the constrained estimation problem we are working with the Lagrangian rather than the log-likelihood function. The role of any constraints is to restrict the set of possible θ from which the maximum likelihood estimate can be chosen as a solution to the optimization problem. If the true value θ_0 is not part of that set then the estimator must necessarily be inconsistent. However, if θ_0 is part of the feasible set then, given that this is the value at which the expected log-likelihood function is maximized, it follows that the maximum likelihood estimator will converge to θ_0 almost surely. That is, if $g(\theta_0) = 0$ is true in the population then $\tilde{\theta}_n \xrightarrow{a.s.} \theta_0$, so that $\tilde{\theta}_n$ is consistent for θ_0.

The asymptotic distribution for $\tilde{\theta}_n$ and $\tilde{\lambda}_n$ is a little more work because there is a redundancy in the model specification. To illustrate what is meant by this consider a situation where a model has two parameters θ_1 and θ_2, say, and that we wish to impose the restriction $\theta_1 + \theta_2 = 0$. Then the restricted estimator will be $\tilde{\theta}_n' = [\tilde{\theta}_{n,1}, \tilde{\theta}_{n,2}]$ $(= [\tilde{\theta}_{n,1}, -\tilde{\theta}_{n,1}]$ say). We see that there is only one parameter to be freely estimated, even though we notionally have two parameters that we are estimating. The covariance matrix of $\tilde{\theta}_n$ will then have the form

$$\operatorname{Var}\left[\tilde{\theta}_n\right] = \begin{bmatrix} \operatorname{Var}\left[\tilde{\theta}_{n,1}\right] & \operatorname{Cov}\left[\tilde{\theta}_{n,1}, \tilde{\theta}_{n,2}\right] \\ \operatorname{Cov}\left[\tilde{\theta}_{n,2}, \tilde{\theta}_{n,1}\right] & \operatorname{Var}\left[\tilde{\theta}_{n,2}\right] \end{bmatrix} = \begin{bmatrix} \operatorname{Var}\left[\tilde{\theta}_{n,1}\right] & -\operatorname{Var}\left[\tilde{\theta}_{n,1}\right] \\ -\operatorname{Var}\left[\tilde{\theta}_{n,1}\right] & \operatorname{Var}\left[\tilde{\theta}_{n,1}\right] \end{bmatrix}.$$

This matrix is positive semi-definite and hence singular. You can see that it is rank deficient, and therefore singular, because one row (or column) can be written as multiple of the other, where the scale factor is -1. Alternatively it is easy to see that its determinant is equal to zero, so that its inverse is undefined. A singular covariance matrix is different from what we are used to, although it is not something that we need to think about too much.[53]

[52] The rank assumption on $G(\theta_0)$ amounts to the assumption that there are no redundant assumptions. Another way of saying this is that each of the assumptions are assumed to be distinct, so that you are not testing the same assumption more than once. For example, $\theta_1 = 6$ and $\theta_2 = 4$ are two distinct restrictions whereas $\theta_1 = 6$ and $3\theta_1 = 18$ are not.

[53] A good treatment of the multivariate Normal distribution with a singular covariance matrix, the context of interest to us, can be found in Mardia, Kent, and Bibby (1979). The treatment by Searle (1982) is also very accessible. Rao and Mitra (1971) and Rao (1973) also provide an extensive treatment of many useful distributional results for singular distributions, but their treatments are quite technical.

Given that $\tilde{\theta}_n$ is consistent for θ_0 when H_0 is true, application of Taylor's Theorem yields the following three results. First,

$$\begin{aligned} n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y) &= n^{-1/2}\mathcal{S}_n(\theta_0; y) + n^{-1}H_n(\theta^*_{1,n}; y)n^{1/2}(\tilde{\theta}_n - \theta_0) \\ &= n^{-1/2}\mathcal{S}_n(\theta_0; y) - \imath_{\theta_0} n^{1/2}(\tilde{\theta}_n - \theta_0) \\ &\qquad + \underbrace{n^{-1}\left[H_n(\theta^*_{1,n}; y) + \imath_{\theta_0}\right]}_{=o_p(1)} \underbrace{n^{1/2}(\tilde{\theta}_n - \theta_0)}_{=O_p(1)} \\ &= n^{-1/2}\mathcal{S}_n(\theta_0; y) - \imath_{\theta_0} n^{1/2}(\tilde{\theta}_n - \theta_0) + o_p(1), \end{aligned} \tag{43}$$

Second,

$$\begin{aligned} n^{-1/2}G(\tilde{\theta}_n)'\tilde{\lambda}_n &= G(\theta_0)' n^{-1/2}\tilde{\lambda}_n + \underbrace{\left[G(\theta^*_{2,n}) - G(\theta_0)\right]'}_{=o_p(1)} n^{-1/2}\tilde{\lambda}_n \\ &= G(\theta_0)' n^{-1/2}\tilde{\lambda}_n + o_p(1), \end{aligned} \tag{44}$$

where the second equality follows because (i) $G(\theta_0) = O(1)$ and so the term

$$\left[G(\theta^*_{2,n}) - G(\theta_0)\right]' n^{-1/2}\tilde{\lambda}_n.$$

is of smaller order than is $G(\theta_0)' n^{-1/2}\tilde{\lambda}_n$ and (ii) from (41) we see that

$$n^{-1/2}G(\tilde{\theta}_n)'\tilde{\lambda}_n = n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y).$$

As the right-hand side of this expression is of stochastic order $O_p(1)$ it follows that $n^{-1/2}\tilde{\lambda}_n = O_p(1)$ because $G(\theta_0) = O(1)$. Our third expansion is

$$\begin{aligned} n^{1/2}g(\tilde{\theta}_n) &= n^{1/2}\underbrace{g(\theta_0)}_{=0} + G(\theta_0)n^{1/2}(\tilde{\theta}_n - \theta_0) + \underbrace{\left[G(\theta^*_{3,n}) - G(\theta_0)\right]}_{=o_p(1)} \underbrace{n^{1/2}(\tilde{\theta}_n - \theta_0)}_{=O_p(1)} \\ &= G(\theta_0)n^{1/2}(\tilde{\theta}_n - \theta_0) + o_p(1), \end{aligned} \tag{45}$$

where, in the derivations of equations (43)–(45), we have used the notation

$$\theta^*_{k,n} = \alpha_{k,n}\tilde{\theta}_n + (1 - \alpha_{k,n})\theta_0, \quad 0 < \alpha_{k,n} < 1, \quad k = 1, 2, 3.$$

Note that the matrices G, which are not functions of n, do not require scaling whereas the hessian matrices H_n do.

Combining equations (43)–(45) with suitably scaled versions of equations (41) and (42) yields

$$\begin{gathered} n^{-1/2}\mathcal{S}_n(\theta_0; y) - \imath_{\theta_0} n^{1/2}(\tilde{\theta}_n - \theta_0) - G(\theta_0)' n^{-1/2}\tilde{\lambda}_n + o_p(1) = 0 \\ G(\theta_0)n^{1/2}(\tilde{\theta}_n - \theta_0) + o_p(1) = 0 \end{gathered}$$

or, in matrix notation,

$$\begin{bmatrix} \imath_{\theta_0} & G(\tilde{\theta}_n)' \\ G(\theta_0) & \mathbf{0} \end{bmatrix} \begin{bmatrix} n^{1/2}(\tilde{\theta}_n - \theta_0) \\ n^{-1/2}\tilde{\lambda}_n \end{bmatrix} + o_p(1) = \begin{bmatrix} n^{-1/2}\mathcal{S}_n(\theta_0; y) \\ 0 \end{bmatrix}.$$

Using equation (4.3.2) from the Matrices handout, we can solve this equation to obtain

$$\begin{aligned}\begin{bmatrix} n^{1/2}(\tilde{\theta}_n - \theta_0) \\ n^{-1/2}\tilde{\lambda}_n \end{bmatrix} &= \begin{bmatrix} \imath_{\theta_0} & G(\tilde{\theta}_n)' \\ G(\theta_0) & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} n^{-1/2}\mathcal{S}_n(\theta_0; y) \\ 0 \end{bmatrix} + o_p(1) \\ &= \begin{bmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{bmatrix} \begin{bmatrix} n^{-1/2}\mathcal{S}_n(\theta_0; y) \\ 0 \end{bmatrix} + o_p(1) \qquad (46) \\ &= \begin{bmatrix} A_{11} \\ A_{12}' \end{bmatrix} n^{-1/2}\mathcal{S}_n(\theta_0; y) + o_p(1) \qquad (47)\end{aligned}$$

where

$$\begin{aligned} A_{11} &= \imath_{\theta_0}^{-1} - \imath_{\theta_0}^{-1}G(\tilde{\theta}_n)'(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1}G(\tilde{\theta}_n)\imath_{\theta_0}^{-1} \\ &= \imath_{\theta_0}^{-1} - \imath_{\theta_0}^{-1}G(\theta_0)'(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}G(\theta_0)\imath_{\theta_0}^{-1} + o_p(1) \\ A_{12} &= \imath_{\theta_0}^{-1}G(\tilde{\theta}_n)'(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1} = \imath_{\theta_0}^{-1}G(\theta_0)'(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1} + o_p(1) \end{aligned}$$

and

$$A_{22} = -(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1} = -(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1} + o_p(1).$$

Given that

$$n^{1/2}\mathcal{S}_n(\theta_0; y) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0})$$

it follows that, when H_0 is true,

$$\begin{bmatrix} n^{1/2}(\tilde{\theta}_n - \theta_0) \\ n^{-1/2}\tilde{\lambda}_n \end{bmatrix} \xrightarrow{d} \mathrm{N}(0, B_0), \qquad (48)$$

where

$$\begin{aligned} B_0 &= \operatorname*{plim}_{n\to\infty} \begin{bmatrix} A_{11}\imath_{\theta_0}A_{11} & A_{11}\imath_{\theta_0}A_{12} \\ A_{12}'\imath_{\theta_0}A_{11} & A_{12}'\imath_{\theta_0}A_{12} \end{bmatrix} \\ &= \begin{bmatrix} \imath_{\theta_0}^{-1} - \imath_{\theta_0}^{-1}G(\theta_0)'(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}G(\theta_0)\imath_{\theta_0}^{-1} & \mathbf{0} \\ \mathbf{0}' & (G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1} \end{bmatrix}. \end{aligned}$$

Our first observation is that their joint asymptotic distribution is Normal which, when coupled with a zero covariance, implies that the terms $n^{1/2}(\tilde{\theta}_n - \theta_0)$ and $n^{-1/2}\tilde{\lambda}_n$ are asymptotically independent. Second, on defining the $j \times p$ matrix $T = \imath_{\theta_0}^{-1/2}G(\theta_0)'$, which is of rank j, we can write

$$\imath_{\theta_0}^{-1} - \imath_{\theta_0}^{-1}G(\theta_0)'(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}G(\theta_0)\imath_{\theta_0}^{-1} = \imath_{\theta_0}^{-1/2}[I_p - T(T'T)^{-1}T']\imath_{\theta_0}^{-1/2}.$$

Now the matrix $J_T = I_p - T(T'T)^{-1}T'$ is idempotent and has rank $p - j$, as must $\imath_{\theta_0}^{-1/2}J_T\imath_{\theta_0}^{-1/2}$ (because $\imath_{\theta_0}^{-1/2}$ has full rank). This is as it should be because, there are j restrictions amongst the p elements of $\tilde{\theta}_n$ and so there are only $p-j$ free elements. Noting that $(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}$ has rank equal to j, it follows that, although B_0 is of dimension $(p+j) \times (p+j)$ its rank is p, the rank of $\imath_{\theta_0}$, and so is singular. As described earlier, this means that the joint density function of $n^{1/2}(\tilde{\theta}_n - \theta_0)$ and $n^{-1/2}\tilde{\lambda}_n$ is not a proper density function but, as it is a useful representation, we shall not pursue this point further.

B.4.1 The Cramér-Rao Lower Bound

The Exact Result Let $\tilde{\theta}_n$ denote any estimator for θ_0. This estimator may or may not be biased. If $\tilde{\theta}_n$ is biased for θ_0 then we can write

$$\mathrm{E}_0\left[\tilde{\theta}_n\right] = \theta_0 + b_n(\theta_0), \tag{49}$$

where $b_n(\theta_0)$ represents the bias in $\tilde{\theta}_n$ and the notation is chosen to indicate that the bias may vary with both the sample size and the value of θ_0. It will be convenient in what follows to introduce the notation $B_n(\theta_0) = [\partial b_n(\theta_0)'/\partial\theta_0]$. Note that neither $b_n(\theta_0)$, which is the expected value of $\tilde{\theta}_n - \theta_0$, nor, consequently, $B_n(\theta_0)$ depend upon y.

Our first task is to find the covariance matrix for $\tilde{\theta}_n$ and $\mathcal{S}_n(\theta_0; y)$. If we differentiate both sides of (49) with respect to θ_0, so that we are exploring how the expected value of $\tilde{\theta}_n$ changes as the value of θ_0 changes, we obtain

$$\frac{\partial \mathrm{E}_0\left[\tilde{\theta}_n'\right]}{\partial\theta_0} = \frac{\partial\theta_0'}{\partial\theta_0} + \frac{\partial b_n(\theta_0)'}{\partial\theta_0} = I_p + B_n(\theta_0).$$

Writing the expectation explicitly as an integral yields

$$\frac{\partial}{\partial\theta_0}\int_{\mathcal{Y}} \tilde{\theta}_n' f(y;\theta_0)\, dy = I_p + B_n(\theta_0).$$

If we now assume sufficient regularity to interchange the orders of differentiation and integration we have

$$\int_{\mathcal{Y}} \frac{\partial f(y;\theta_0)}{\partial\theta_0}\, \tilde{\theta}_n'\, dy = I_p + B_n(\theta_0).$$

Note that $\tilde{\theta}_n$ is a constant with respect to θ_0 because, being an estimator, it cannot be a function of an unknown parameter. Therefore, only the density $f(y;\theta_0)$ depends upon θ_0, which is why it is the only one of the two terms differentiated with respect to θ_0. Applying (27) yields

$$\int_{\mathcal{Y}} \frac{\partial \ln \mathcal{L}_n(y;\theta_0)}{\partial\theta_0}\, \tilde{\theta}_n'\, f(y;\theta_0)\, dy = I_p + B_n(\theta_0),$$

or, recognizing that the integral is simply the covariance matrix between $\tilde{\theta}_n$ and $\mathcal{S}_n(\theta_0; y)$,

$$\mathrm{E}_0\left[\mathcal{S}_n(\theta_0;y)\tilde{\theta}_n'\right] = \mathrm{Cov}_0\left[\mathcal{S}_n(\theta_0;y), \tilde{\theta}_n\right] = I_p + B_n(\theta_0), \tag{50}$$

where the first equality follows because $\mathrm{E}_0\left[\mathcal{S}_n(\theta_0;y)\right] = 0$.

Next define the mean squared error of an estimator $\tilde{\theta}_n$ about θ_0 to be

$$\mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] = \mathrm{E}_0\left[(\tilde{\theta}_n - \theta_0)(\tilde{\theta}_n - \theta_0)'\right] \geq \mathbf{0}.$$

We see that the mean squared error is a simple generalization of variance allowing for the possibility that the point θ_0 about which squared deviations are taken may not be the expected value of $\tilde{\theta}_n$. Clearly if $\tilde{\theta}_n$ is unbiased for θ_0 then $\mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] = \mathrm{V}_0\left[\tilde{\theta}_n\right]$.

Finally, because $(\tilde{\theta}_n - \theta_0)(\tilde{\theta}_n - \theta_0)'$ is an outer product, and hence positive semi-definite, it follows that $\mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] \geq \mathbf{0}$.

Now let $\tau = [\tilde{\theta}_n', \mathcal{S}_n(\theta_0; y)']'$ and $\mu = [\theta_0', 0']'$. It follows that

$$\begin{aligned}\mathrm{MSE}_0\left[\tau, \mu\right] &= \begin{bmatrix} \mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] & \mathrm{Cov}_0\left[\mathcal{S}_n(\theta_0; y), \tilde{\theta}_n\right] \\ \mathrm{Cov}_0\left[\tilde{\theta}_n, \mathcal{S}_n(\theta_0; y)\right] & \mathrm{V}_0\left[\mathcal{S}_n(\theta_0; y)\right] \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] & I_p + B_n(\theta_0) \\ I_p + B_n(\theta_0)' & I(\theta_0) \end{bmatrix} \geq \mathbf{0}.\end{aligned}$$

By definition, for any positive semi-definite matrix A, it follows that $X'AX$ is also positive semi-definite. If we set $A = \mathrm{MSE}_0\left[\tau, \mu\right]$ and choose

$$X = \left[I_p, -\left(I_p + B_n(\theta_0)\right)\left[I(\theta_0)\right]^{-1}\right]',$$

it follows that

$$\mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] \geq \left(I_p + B_n(\theta_0)\right)\left[I(\theta_0)\right]^{-1}\left(I_p + B_n(\theta_0)\right)'. \tag{51}$$

The Cramér-Rao Lower Bound (CRLB) states that if $\tilde{\theta}_n$ is unbiased for θ_0 then $\mathrm{V}_0\left[\tilde{\theta}_n\right] \geq \left[I(\theta_0)\right]^{-1}$. This result is a trivial application of (51). If $\tilde{\theta}_n$ is unbiased for θ_0 then $b_n(\theta_0) = 0$ and hence $B_n(\theta_0) = \mathbf{0}$. Furthermore, $\mathrm{MSE}_0\left[\tilde{\theta}_n, \theta_0\right] = \mathrm{V}_0\left[\tilde{\theta}_n\right]$. If we make these two substitutions into (51) we obtain the desired result immediately. We say that an estimator is efficient if its variance attains the CRLB, so that $\mathrm{V}_0\left[\tilde{\theta}_n\right] - \left[I(\theta_0)\right]^{-1}$.

The CRLB is a lower bound only. It doesn't provide any information as to whether the lower bound is attainable nor does it provide any indication as to just how close one might get if the lower bound is unattainable. It is possible to say something about this former question. From (50) we see that the covariance between an unbiased estimator and the score is I_p. This implies that we can write one as a linear function of the other, for example, $\mathcal{S}_n(\theta_0; y) = \alpha + \Gamma\tilde{\theta}_n$, where neither the p-vector α nor the $p \times p$ matrix Γ depend upon y. Now we know that $\mathrm{E}_0\left[\mathcal{S}_n(\theta_0; y)\right] = 0$ and so this implies that

$$0 = \mathrm{E}_0\left[\alpha + \Gamma\tilde{\theta}_n\right] = \alpha + \Gamma\mathrm{E}_0\left[\tilde{\theta}_n\right] = \alpha + \Gamma\theta_0.$$

Solving for α we see that $\mathcal{S}_n(\theta_0; y) = \Gamma(\tilde{\theta}_n - \theta_0)$. If we make this substitution in (50) we obtain

$$\mathrm{E}_0\left[\Gamma(\tilde{\theta}_n - \theta_0)\tilde{\theta}_n'\right] = \Gamma\mathrm{E}_0\left[(\tilde{\theta}_n - \theta_0)\tilde{\theta}_n'\right] = \Gamma\mathrm{V}_0\left[\tilde{\theta}_n\right] = I_p,$$

which implies that $\Gamma = \left[\mathrm{V}_0\left[\tilde{\theta}_n\right]\right]^{-1}$.[54] If, and only if, the CRLB is attainable then $\mathrm{V}_0\left[\tilde{\theta}_n\right] = \left[I(\theta_0)\right]^{-1}$. We conclude that a necessary and sufficient condition for the CRLB to be attainable is that the score factorizes according to $\mathcal{S}_n(\theta_0; y) = I(\theta_0)(\tilde{\theta}_n - \theta_0)$. If you repeat the treatment of the linear regression model in a likelihood context then you will see that this factorization applies for the regression coefficients, β, but not for the variance, σ^2 and so the CRLB is unattainable for unbiased estimators of this parameter. Consequently OLS, which is the maximum likelihood estimator in this context, attains the CRLB but the unbiased variance estimator s_n^2 does not.

[54] The second equality exploits the standard result that if $\mathrm{E}[X] = \mu_x$ and $\mathrm{E}[Y] = \mu_y$ then $\mathrm{E}[(X - \mu_x)(Y - \mu_y)'] = \mathrm{E}[(X - \mu_x)Y']$.

The Limiting Case In many practical problems unbiased estimators are the exception rather than the rule. In such cases we typically focus on the use of consistent estimators for which we can construct asymptotic distributions that provide good approximations to the true sampling distributions. Up to, and including, equation (51), we explicitly made allowance for the possibility that the estimator was biased and so there is nothing in the analysis that requires significant change. We note that of $\tilde{\theta}_n - \theta_0 \xrightarrow{p} 0$ it follows that $b_n(\theta_0) \xrightarrow{p} 0$ and, by implication, $B_n(\theta_0) \xrightarrow{p} \mathbf{0}$. This might suggest that (51) becomes something like

$$\lim_{n\to\infty} \mathrm{V}_0\left[\tilde{\theta}_n\right] \geq \lim_{n\to\infty} \left[I(\theta_0)\right]^{-1}.$$

This is, in fact, an empty statement because we know that, on the left-hand-side, the variance of a consistent estimator collapses to zero and, on the right-hand-side, $I(\theta_0)$ is a sum of n terms each equal to $\imath_{\theta_0}$ and so it diverges, which is the same as its inverse converging to zero. Consequently we need to apply the scaling necessary to obtain a valid limiting distribution for some transformed function of $\tilde{\theta}_n$. From our earlier developments, we see that when working with *iid* data, it is $\sqrt{n}\,(\tilde{\theta}_n - \theta_0)$ which has the valid limiting distribution. Therefore, the result that we need is

$$\lim_{n\to\infty} n\mathrm{V}_0\left[\tilde{\theta}_n\right] \geq \lim_{n\to\infty} n\left[I(\theta_0)\right]^{-1} = \lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}\imath_{\theta_0}\right]^{-1} = \imath_{\theta_0}^{-1}.$$

This is the limiting form of the Cramér-Rao inequality and we see from the previous section that, under our assumption, the maximum likelihood estimator is *asymptotically efficient* by which we mean that the variance of its limiting distribution is equal to the limit of the average information in the sample.

As a final comment, some writers are so wedded to the notion of consistent estimation that when they write of the CRLB they will, in fact, be talking about this limiting case. It is important that you know exactly what is meant, especially if you find yourself working with non-*iid* data where the limit of the average information will not be the same as the information in each observation.

B.5 The Classical Tests

In what follows we shall concentrate on the problem of testing the simple null hypothesis $H_0 : g(\theta_0) = 0$ against the two-sided alternative $H_1 : g(\theta_0) \neq 0$. The theme that runs through all of the derivations in this section is to write the statistic of interest as a quadratic form, something like $x'Ax$, where x is a random variable having a Normal distribution and A is some (symmetric) matrix that is not a function of x.[55] We shall see that these quadratic forms have chi-squared distributions. Sometimes this will be easy to see, when $A = \mathrm{Var}\,[x]^{-1}$, other times A will be singular which will make things a little more complicated. Nevertheless the underlying approach will remain the same in each case and this is the framework on which all the detail is hung.

B.5.1 The Wald Test

The idea of the Wald test (Wald, 1943) is to obtain maximum likelihood estimates from an unrestricted model and to see whether they satisfy the restrictions being tested. If

[55] If A is not symmetric then we note that we can always write $x'Ax = x'A^*x$, where $A^* = \frac{1}{2}(A + A')$, which is symmetric by construction.

they do then this implies that the restrictions hold in the sample data, from which we infer that they hold in the population. The approach is the following.

1. We know that, in regular problems, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}^{-1})$.

2. Applying the Delta method we find that, in general, if

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}^{-1}) \tag{52}$$

then

$$n^{1/2}[g(\hat{\theta}_n) - g(\theta_0)] \xrightarrow{d} \mathrm{N}\left(0, G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right), \tag{53}$$

where, as usual, the $j \times p$ matrix

$$G(\theta_0) = \left[\left.\frac{\partial g(\theta)'}{\partial \theta}\right|_{\theta=\theta_0}\right].$$

is assumed to have full row rank.[56]

3. If $H_0 : g(\theta_0) = 0$ is true then (53) reduces to

$$n^{1/2}g(\hat{\theta}_n) \xrightarrow{d} \mathrm{N}\left(0, G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right),$$

4. At this stage we have two equivalent paths to our final result.

   (a) The first is to exploit the result:

   If the n-vector $y \sim \mathrm{N}(\mu, \Sigma)$ then $\Sigma^{-1/2}(y - \mu) \sim \mathrm{N}(0, I_n)$.

   (b) The other exploits the following result:

   If the n-vector $y \sim N(\mu, \Sigma)$ then $(y - \mu)'\Sigma^{-1}(y - \mu) \sim \chi^2_n$.

5. Using the preceding results we see that if $H_0 : g(\theta_0) = 0$ is true then

   (a) $Z_0 = n^{1/2}\left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1/2} g(\hat{\theta}_n) \xrightarrow{d} \mathrm{N}(0, I_j)$.

   (b) $W_0 = ng(\hat{\theta}_n)'\left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1} g(\hat{\theta}_n) \xrightarrow{d} \chi^2_j$.

6. Unfortunately, neither Z_0 nor W_0 are operational because both require knowledge of θ_0 in order to evaluate $G(\theta_0)I(\theta_0)^{-1}G(\theta_0)'$. Recognizing that $\hat{\theta}_n$ is always consistent for θ_0, under both H_0 and H_1, we see that

$$\operatorname*{plim}_{n\to\infty} n\left[G(\hat{\theta}_n)\imath_{\hat{\theta}_n}^{-1}G(\hat{\theta}_n)' - G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right] = \mathbf{0}.$$

   Given $G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)' > \mathbf{0}$,[57] the functions

$$\left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1/2} \quad \text{and} \quad \left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1}$$

[56] See discussion of Footnote 52 on page 43.

[57] By assumption, $G(\theta_0)$ has full column rank, which implies that $G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)' > \mathbf{0}$ if $\imath_{\theta_0}^{-1} > \mathbf{0}$. But, $\imath_{\theta_0}^{-1} > \mathbf{0}$ if $\imath_{\theta_0} > \mathbf{0}$. Now $\imath_{\theta_0} > \mathbf{0}$ is a condition for identification, see the discussion on page 42, and is something that we are assuming to be true.

are everywhere continuous, which allows us to write the following:

$$\begin{aligned} Z_0 &= Z_n + n^{1/2}\left\{\left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1/2} - \left[G(\hat{\theta}_n)\imath_{\hat{\theta}_n}^{-1}G(\hat{\theta}_n)'\right]^{-1/2}\right\} g(\hat{\theta}_n) \\ &= Z_n + o_p(1), \end{aligned}$$

where

$$Z_n = n^{1/2}\left[G(\hat{\theta}_n)\imath_{\hat{\theta}_n}^{-1}G(\hat{\theta}_n)'\right]^{-1/2} g(\hat{\theta}_n),$$

and

$$\begin{aligned} W_0 &= W_n + ng(\hat{\theta}_n)'\left\{\left[G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)'\right]^{-1} - \left[G(\hat{\theta}_n)\imath_{\hat{\theta}_n}^{-1}G(\hat{\theta}_n)'\right]^{-1}\right\} g(\hat{\theta}_n) \\ &= W_n + o_p(1), \end{aligned}$$

where

$$W_n = ng(\hat{\theta}_n)'\left[G(\hat{\theta}_n)\imath_{\hat{\theta}_n}^{-1}G(\hat{\theta}_n)'\right]^{-1} g(\hat{\theta}_n)$$

and, in an obvious notation, $\imath_{\hat{\theta}_n}$ denotes $\imath_\theta$ evaluated at $\hat{\theta}_n$. We see that Z_0 and Z_n have the same limiting distribution, namely $\mathrm{N}(0, I_j)$, as do W_0 and W_n, which both have χ^2_j limiting distributions when H_0 is true. Both Z_n and W_n are (operational) statistics and both depend upon the unrestricted estimator $\hat{\theta}_n$.

7. The α-level Wald test is then defined by its critical region, which is the set of values for the statistic that are deemed inconsistent with the null hypothesis being true. The size, or level, of the test is then the probability of observing a value for the statistic which falls in the critical region when the null hypothesis is true. Using the set notation $\{x : \text{condition}\}$, which should be read as 'the set of x such that the condition is satisfied', we define the Wald test as follows.

   (a) If using Z_n then the α-level Wald test of H_0 against H_1 is:

   $$\left\{Z_n : |Z_n| \geq (z_{\alpha/2})\ell\right\}$$

   where, the j-vector $\ell = [1, \ldots, 1]'$ and, for $z \sim \mathrm{N}(0,1)$, $z_{1-\alpha/2}$ is defined by the equation

   $$P\left[z \leq z_{1-\alpha/2}\right] = 1 - \alpha/2.$$

   That is, $z_{1-\alpha/2}$ is the critical value which cuts off a lower tail probability of $1-\alpha/2$ in a standard Normal distribution or, equivalently, a size $\alpha/2$ upper tail. We reject H_0 against H_1 is every element of Z_n lies in the two-tailed critical region. For example, if $Z_n = [Z_{n,1}, Z_{n,2}]'$ then our rejection region is the set of possible Z_n such that $|Z_{n,1}| \geq z_{1-\alpha/2}$ and $|Z_{n,2}| \geq z_{1-\alpha/2}$. This is not a commonly used form of the test but we see that if $j = 1$ Z_n is just the usual t statistic which is being compared to critical values from a standard Normal distribution, which is the usual asymptotic approximation.[58]

   (b) When $j > 1$ it is more common to see W_n employed as the Wald test statistic. If using W_n then the α-level Wald test of H_0 against H_1 is:

   $$\left\{W : W \geq \chi^2_{j,(1-\alpha)}\right\}.$$

[58] The special case of $j = 1$ is the example of a Wald test used by Winkelmann and Boes (2006, Section 3.6).

where, for $C \sim \chi^2_j$, $\chi^2_{j,(1-\alpha)}$ is defined by the equation

$$P\left[C \leq \chi^2_{j,(1-\alpha)}\right] = 1 - \alpha.$$

That is, $\chi^2_{j,(1-\alpha)}$ is the critical value which cuts off a lower tail probability of $1-\alpha$ in a χ^2_j distribution or, equivalently, a size α upper tail.

More broadly, one can imagine constructing Wald-type tests on the basis of any consistent estimator for which we can find a limiting distribution, such as generalized methods of moments estimators. The arguments to be followed are exactly those outlined above and will not be pursued further.

B.5.2 The (Efficient) Score and Lagrange Multiplier Tests

The score test (or, more completely, the efficient score test) was originally proposed by Rao (1948). Subsequently, Aitchison and Silvey (1958) introduced the Lagrange multiplier test, whose properties were explored in Silvey (1959). As we shall see these tests are asymptotically equivalent.

The motivations for the two test procedures are as follows:

Score test We know that the score has zero expectation when the expectation is evaluated at the true parameter θ_0.[59] If, in the population, $g(\theta_0) = 0$ then the constrained maximum likelihood estimator $\tilde{\theta}_n$ is consistent for θ_0. Consequently, for sufficiently large samples, we would expect the score to be close to zero when evaluated at $\tilde{\theta}_n$ if H_0 is true. The score test is a test of this restriction.

LM test As we show below, the Lagrange multiplier is equal to zero when the restrictions are satisfied. The idea of the LM test is then to test this restriction. That this ends up being equivalent to the score test should not be surprising in light of (47) where we see $\tilde{\lambda}_n$ expressed as a function of the score.

Let us now address the details of each variant of the test.

The Score Test Combining (43) and (47) we can write

$$\begin{aligned} n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y) &= n^{-1/2}\mathcal{S}_n(\theta_0; y) - \imath_{\theta_0} n^{1/2}(\tilde{\theta}_n - \theta_0) + o_p(1), \\ &= n^{-1/2}\mathcal{S}_n(\theta_0; y) - \imath_{\theta_0} A_{11} n^{-1/2}\mathcal{S}_n(\theta_0; y) + o_p(1) \\ &= [I_p - \imath_{\theta_0} A_{11}] n^{-1/2}\mathcal{S}_n(\theta_0; y) + o_p(1), \end{aligned}$$

where we recall that

$$A_{11} = \imath^{-1}_{\theta_0} - \imath^{-1}_{\theta_0} G(\theta_0)'(G(\theta_0)\imath^{-1}_{\theta_0} G(\theta_0)')^{-1} G(\theta_0)\imath^{-1}_{\theta_0} + o_p(1).$$

From (38) we have

$$n^{-1/2}\mathcal{S}_n(\theta_0; y) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0}),$$

and so it follows that

$$n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y) \xrightarrow{d} \mathrm{N}(0, V_s(\theta_0)).$$

[59] The score evaluated at the true parameter is known as the efficient score.

where, after a little tedious algebra, we find that

$$\begin{aligned} V_s(\theta_0) &= \text{plim}[I_p - \imath_{\theta_0} A_{11}]\imath_{\theta_0}[I_p - \imath_{\theta_0} A_{11}]' \\ &= G(\theta_0)'(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}G(\theta_0). \end{aligned}$$

Our problem is that $V_s(\theta_0)$ has less than full rank and is, therefore, singular. Even though $V_s(\theta_0)$ is of dimension $p \times p$, the matrices $(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}$ and $G(\tilde{\theta}_n)$ are both of rank j and hence so is $V_s(\theta_0)$. Consequently we are unable to appeal to the result 4(b) (on page 49) that we used with the Wald statistic. However, Corollary 2 of Appendix C.2 of the Normality handout provides the necessary extension of that result that we need. Specifically we see that

$$n^{-1}\mathcal{S}_n(\tilde{\theta}_n; y)' V_s(\theta_0)^- \mathcal{S}_n(\tilde{\theta}_n; y) \xrightarrow{d} \chi^2_j.$$

Our outstanding task is to find some matrix $V_s(\theta_0)^-$. From the discussion of generalized inverses in the Matrices handout we see that there is no unique choice but, for the purposes of Corollary 2, any choice will do. It is readily apparent that one choice for $V_s(\theta_0)^-$ is $\imath_{\theta_0}^{-1}$. This yields

$$S_0 = n^{-1}\mathcal{S}_n(\tilde{\theta}_n; y)'\imath_{\theta_0}^{-1}\mathcal{S}_n(\tilde{\theta}_n; y) \xrightarrow{d} \chi^2_j.$$

Of course, S_0 is not a statistic because $\imath_{\theta_0}^{-1}$ depends upon the unknown parameter θ_0. However, as $\tilde{\theta}_n$ is consistent for θ_0 when H_0 is true, an operational statistic is

$$S_n = n^{-1}\mathcal{S}_n(\tilde{\theta}_n; y)'\imath_{\tilde{\theta}_n}^{-1}\mathcal{S}_n(\tilde{\theta}_n; y).$$

Because

$$\begin{aligned} S_n &= S_0 + (S_n - S_0) \\ &= S_0 + \underbrace{n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y)}_{=O_p(1)} \underbrace{\left[\imath_{\theta_0}^{-1} - \imath_{\tilde{\theta}_n}^{-1}\right]}_{=o_p(1)} \underbrace{n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y)}_{=O_p(1)} \\ &= S_0 + o_p(1), \end{aligned}$$

we see that S_n and S_0 have the same limiting distribution. That is,

$$S_n = \mathcal{S}_n(\tilde{\theta}_n; y)\left[n\imath_{\tilde{\theta}_n}\right]^{-1}\mathcal{S}_n(\tilde{\theta}_n; y) = \mathcal{S}_n(\tilde{\theta}_n; y)I_n(\tilde{\theta}_n)^{-1}\mathcal{S}_n(\tilde{\theta}_n; y) \xrightarrow{d} \chi^2_j. \tag{54}$$

We note in passing that S_n has the same limiting distribution as does the Wald test W_n. Even though these limiting distributions are the same it does not follow that the two tests will necessarily yield the same conclusions in finite samples.

The LM Test Our first task is to establish that the Lagrange multipliers have zero mean when the restrictions hold in the population. This follows immediately from (47) and (48) where we see that if H_0 is true then[60]

$$n^{-1/2}\tilde{\lambda}_n = (G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}G(\theta_0)\imath_{\theta_0}^{-1}n^{-1/2}\mathcal{S}_n(\tilde{\theta}_n; y) + o_p(1) \tag{55a}$$

$$\xrightarrow{d} \text{N}\left(0, (G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}\right). \tag{55b}$$

[60] An alternative path to this point is to re-arrange (41) to obtain

$$G(\tilde{\theta}_n)'\tilde{\lambda}_n = \mathcal{S}_n(\tilde{\theta}_n; y). \tag{*}$$

Now, $G(\tilde{\theta}_n)'$ is rectangular rather than square and so we cannot 'solve' the equation for $\tilde{\lambda}_n$ by pre-multiplying both sides of the equation by $[G(\tilde{\theta}_n)']^{-1}$. However, if we pre-multiply both sides of (*) by

Now, we should also establish that $n^{-1/2}\tilde{\lambda}_n$ has a non-zero mean if H_0 is false but, at this stage, I will ask you to take my word for it.

Appealing again to the result of 4(b) (on page 49), and using the distributional result (55b), we see that a natural test statistic is based on

$$LM_0 = n^{-1/2}\tilde{\lambda}_n' \left[(G(\theta_0)\imath_{\theta_0}^{-1}G(\theta_0)')^{-1}\right]^{-1} n^{-1/2}\tilde{\lambda}_n \xrightarrow{d} \chi_j^2.$$

This statistic is not operational but writing

$$\begin{aligned} LM_n &= n^{-1}\tilde{\lambda}_n' G(\tilde{\theta}_n)\imath_{\tilde{\theta}_n}^{-1}G(\tilde{\theta}_n)'\tilde{\lambda}_n \\ &= LM_0 + \underbrace{(LM_n - LM_0)}_{=o_{p(1)}} \end{aligned} \tag{56}$$

we see that LM_n and LM_0 have the same limiting distribution so that, just like W_n and S_n, $LM_n \xrightarrow{d} \chi_j^2$ when H_0 is true.

If we substitute for $n^{-1/2}G(\tilde{\theta}_n)'\tilde{\lambda}_n$ from (41) then LM_n becomes

$$LM_n = n^{-1}\mathcal{S}_n(\tilde{\theta}_n; y)'\imath_{\tilde{\theta}_n}^{-1}\mathcal{S}_n(\tilde{\theta}_n; y) = \mathcal{S}_n(\tilde{\theta}_n; y)' I_n(\tilde{\theta}_n)^{-1}\mathcal{S}_n(\tilde{\theta}_n; y) = S_n$$

and we see that LM_n and S_n are numerically identical.

Of course, we tend not to have infinitely large samples and so, for any finite sample size, an asymptotic approximation is provided by basing a test on critical values from a χ_j^2 distribution as follows. If using LM_n then the α-level LM test of H_0 against H_1 is:

$$\left\{LM : LM \geq \chi_{j,(1-\alpha)}^2\right\}.$$

where, for $C \sim \chi_j^2$, $\chi_{j,(1-\alpha)}^2$ is defined by the equation

$$P\left[C \leq \chi_{j,(1-\alpha)}^2\right] = 1 - \alpha.$$

That is, $\chi_{j,(1-\alpha)}^2$ is the critical value which cuts off a lower tail probability of $1-\alpha$ in a χ_j^2 distribution or, equivalently, a size α upper tail.

B.5.3 The Likelihood Ratio Test

Neyman and Pearson (1933) showed that, when testing a simple null hypothesis against a simple alternative,[61] the likelihood ratio test has at least as much power as any other test. This so-called Neyman-Pearson lemma may not seem a very strong statement but it is possibly the strongest statement of its type in all of the theory of hypothesis testing.[62]

$(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1}G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}$ then we obtain

$$\tilde{\lambda}_n = (G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1}G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}\mathcal{S}_n(\tilde{\theta}_n; y)$$

and, on scaling by $n^{-1/2}$, reach the same end-point. We note in passing that $(G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}G(\tilde{\theta}_n)')^{-1}G(\tilde{\theta}_n)\imath_{\theta_0}^{-1}$ is a reflexive generalized inverse of $G(\tilde{\theta}_n)'$.

[61] Recall that a simple hypothesis is one where the the parameter can only take a single value, e.g. $\mathcal{H}_* : \theta = \theta_*$.

[62] Mood, Graybill, and Boes (1974) provide a good introductory treatment of this material. A step up is provided by Cox and Hinkley (1974), with the definitive treatment being that of Lehmann (1986).

Moreover, although most testing problems do not fall within this class, the Neyman-Pearson lemma is nevertheless used to motivate the use of the likelihood ratio test in a wide variety of testing problems.

The likelihood ratio statistic is defined to be twice the difference between the unrestricted and restricted log-likelihoods. Like W_n, S_n and S_n we shall see that $LR_n \xrightarrow{d} \chi^2_j$. That is,

$$LR_n = 2[\ln \mathcal{L}_n(\hat{\theta}_n; y) - \ln \mathcal{L}_n(\tilde{\theta}_n; y)] \xrightarrow{d} \chi^2_j. \tag{57}$$

To establish this result we first take Taylor series expansions of both the restricted and unrestricted log-likelihoods about θ_0. Thus,

$$\ln \mathcal{L}_n(\hat{\theta}_n; y) = \mathcal{L}_n(\theta_0; y) + \mathcal{S}_n(\theta_0; y)'(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)' H_n(\theta^*_{1,n}; y)(\hat{\theta}_n - \theta_0),$$

where $\theta^*_{1,n} = \alpha_{1,n}\hat{\theta}_n + (1 - \alpha_{1,n})\theta_0$ with $0 < \alpha_{1,n} < 1$, and, for $\theta^*_{2,n} = \alpha_{2,n}\hat{\theta}_n + (1 - \alpha_{2,n})\theta_0$ with $0 < \alpha_{2,n} < 1$,

$$\ln \mathcal{L}_n(\tilde{\theta}_n; y) = \mathcal{L}_n(\theta_0; y) + \mathcal{S}_n(\theta_0; y)'(\tilde{\theta}_n - \theta_0) + \frac{1}{2}(\tilde{\theta}_n - \theta_0)' H_n(\theta^*_{2,n}; y)(\tilde{\theta}_n - \theta_0).$$

Taking twice the difference between these two results yields

$$\begin{aligned} LR_n &= 2[\ln \mathcal{L}_n(\hat{\theta}_n; y) - \ln \mathcal{L}_n(\tilde{\theta}_n; y)] \\ &= 2\mathcal{S}_n(\theta_0; y)'(\hat{\theta}_n - \tilde{\theta}_n) + (\hat{\theta}_n - \theta_0)' H_n(\theta^*_{1,n}; y)(\hat{\theta}_n - \theta_0) \\ &\qquad - (\tilde{\theta}_n - \theta_0)' H_n(\theta^*_{2,n}; y)(\tilde{\theta}_n - \theta_0). \end{aligned} \tag{58}$$

Next we observe that

$$n^{-1} H_n(\theta^*_{1,n}; y) = -\imath_{\theta_0} + \left[n^{-1} H_n(\theta^*_{1,n}; y) + \imath_{\theta_0}\right] = -\imath_{\theta_0} + o_p(1)$$

and

$$n^{-1} H_n(\theta^*_{2,n}; y) = -\imath_{\theta_0} + \left[n^{-1} H_n(\theta^*_{2,n}; y) + \imath_{\theta_0}\right] = -\imath_{\theta_0} + o_p(1),$$

so that

$$\begin{aligned} (\hat{\theta}_n - \theta_0)' H_n(\theta^*_{1,n}; y)(\hat{\theta}_n - \theta_0) &= \underbrace{n^{1/2}(\hat{\theta}_n - \theta_0)'}_{=O_p(1)} n^{-1} H_n(\theta^*_{1,n}; y) \underbrace{n^{1/2}(\hat{\theta}_n - \theta_0)}_{=O_p(1)} \\ &= -n(\hat{\theta}_n - \theta_0)' \imath_{\theta_0} (\hat{\theta}_n - \theta_0) + o_p(1) \end{aligned}$$

and

$$\begin{aligned} (\tilde{\theta}_n - \theta_0)' H_n(\theta^*_{1,n}; y)(\tilde{\theta}_n - \theta_0) &= \underbrace{n^{1/2}(\tilde{\theta}_n - \theta_0)'}_{=O_p(1)} n^{-1} H_n(\theta^*_{1,n}; y) \underbrace{n^{1/2}(\tilde{\theta}_n - \theta_0)}_{=O_p(1)} \\ &= -n(\tilde{\theta}_n - \theta_0)' \imath_{\theta_0} (\tilde{\theta}_n - \theta_0) + o_p(1), \end{aligned}$$

respectively. Substituting these results into (58) yields

$$LR_n = 2n^{-1/2}\mathcal{S}_n(\theta_0; y)' n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n) + n(\tilde{\theta}_n - \theta_0)' \imath_{\theta_0}(\tilde{\theta}_n - \theta_0) - n(\hat{\theta}_n - \theta_0)' \imath_{\theta_0}(\hat{\theta}_n - \theta_0) + o_p(1). \tag{59}$$

Recall from (39) that

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \imath^{-1}_{\theta_0} n^{-1/2} \mathcal{S}_n(\theta_0; y) + o_p(1).$$

Similarly, from (47),

$$n^{1/2}(\tilde{\theta}_n - \theta_0) = B_{11} n^{-1/2} \mathcal{S}_n(\theta_0; y) + o_p(1).$$

where

$$B_{11} = \imath_{\theta_0}^{-1} - \imath_{\theta_0}^{-1} G(\theta_0)' (G(\theta_0) \imath_{\theta_0}^{-1} G(\theta_0)')^{-1} G(\theta_0) \imath_{\theta_0}^{-1}.$$

Substituting these two results into (59), and noting that

$$n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n) = n^{1/2}(\hat{\theta}_n - \theta_0 + \theta_0 - \tilde{\theta}_n) = n^{1/2}(\hat{\theta}_n - \theta_0) - n^{1/2}(\tilde{\theta}_n - \theta_0),$$

yields

$$\begin{aligned} LR_n &= n^{-1} \mathcal{S}_n(\theta_0; y)' [2(\imath_{\theta_0}^{-1} - B_{11}) + B_{11}' \imath_{\theta_0} B_{11} - \imath_{\theta_0}^{-1} \imath_{\theta_0} \imath_{\theta_0}^{-1}] \mathcal{S}_n(\theta_0; y) + o_p(1) \\ &= n^{-1} \mathcal{S}_n(\theta_0; y)' (\imath_{\theta_0}^{-1} - B_{11}) \mathcal{S}_n(\theta_0; y) + o_p(1), \end{aligned} \tag{60}$$

where we have exploited the fact that $B_{11}' \imath_{\theta_0} B_{11} = B_{11}$.

The final step of our derivation is to recall from (38) that

$$n^{-1/2} \mathcal{S}_n(\theta_0; y) \xrightarrow{d} \mathrm{N}(0, \imath_{\theta_0})$$

and then appeal to Corollary 1 of Appendix C.2 of the Normality handout on noting that the matrix

$$\imath_{\theta_0} (\imath_{\theta_0}^{-1} - B_{11}) \imath_{\theta_0} (\imath_{\theta_0}^{-1} - B_{11}) \imath_{\theta_0} = \imath_{\theta_0} (\imath_{\theta_0}^{-1} - B_{11}) \imath_{\theta_0}.$$

and so satisfies condition (i) of the Corollary. Once again, the asymptotic approximation is to use the statistic LR_n, based on finite n, with critical values from a χ_j^2 distribution.

B.5.4 Linear Restrictions

So far we have adopted the implicit representation $g(\theta_0) = 0$ for our hypotheses of interest. The advantage of doing so is that it makes our treatment quite general. The cost, of course, is that it makes the treatment quite general and perhaps more difficult to understand than it might otherwise be. Far and away the most common form of restrictions are linear restrictions, which can be written in the form $R\theta - r = 0$, where R is a $j \times p$ matrix with full row rank and r is a j-vector. Note that neither R nor r can be functions of θ_0. For example, if we wished to test an hypothesis of the form $\theta_1 = 0$ then $R = [1, 0, \ldots, 0]$ and $r = 0$. Equally if we wanted to test $\theta_2 = \theta_3$ then either $R = [0, 1, -1, 0, \ldots, 0]$ or $R = [0, -1, 1, 0, \ldots, 0]$, it doesn't matter which, and $r = 0$. Finally, if we wished to test $\theta_4 = -1$ then $R = [0, 0, 0, 1, 0, \ldots, 0]$ and $r = 1$. If we suppose that $p = 5$ and we wished to test all of the preceding hypotheses simultaneously then we would have

$$g(\theta_0) = R\theta - r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_{0,1} \\ \theta_{0,2} \\ \theta_{0,3} \\ \theta_{0,4} \\ \theta_{0,5} \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

as the hypothesis of interest. We note that $G(\theta_0) = R$ so, for example, an expression like

$$\imath_{\theta_0}^{-1} G(\theta_0)' (G(\theta_0) \imath_{\theta_0}^{-1} G(\theta_0)')^{-1} G(\theta_0) \imath_{\theta_0}^{-1}.$$

reduces to

$$\imath_{\theta_0}^{-1} R'(R\imath_{\theta_0}^{-1}R')^{-1}R\imath_{\theta_0}^{-1}.$$

In the special case where only exclusion restrictions are of interest there is even greater simplification. Suppose that we order the elements of θ_0 so that the parameters potentially set to zero appear first. That is, write $\theta_0 = [\theta_1', \theta_2']'$ with the hypothesis of interest then $\theta_1 = 0$. If θ_1 has $j \leq p$ elements then $R = [I_j, \mathbf{0}]$ and $r = 0$. Continuing our previous example, on writing

$$\imath_{\theta_0}^{-1} = \begin{bmatrix} \imath_{\theta_0}^{11} & \imath_{\theta_0}^{12} \\ \imath_{\theta_0}^{21} & \imath_{\theta_0}^{22} \end{bmatrix}$$

where $\imath_{\theta_0}^{11}$ is of dimension $j \times j$ and $\imath_{\theta_0}^{21} = (\imath_{\theta_0}^{12})'$, we see that

$$\imath_{\theta_0}^{-1} R'(R\imath_{\theta_0}^{-1}R')^{-1}R\imath_{\theta_0}^{-1} = \imath_{\theta_0}^{-1}[I_j, \mathbf{0}]' \left([I_j, \mathbf{0}]\imath_{\theta_0}^{-1}[I_j, \mathbf{0}]'\right)^{-1} [I_j, \mathbf{0}]\imath_{\theta_0}^{-1} = \begin{bmatrix} \imath_{\theta_0}^{11} \\ \imath_{\theta_0}^{21} \end{bmatrix} \left(\imath_{\theta_0}^{11}\right)^{-1} [\imath_{\theta_0}^{11}\imath_{\theta_0}^{12}]$$

$$= \begin{bmatrix} \imath_{\theta_0}^{11} & \imath_{\theta_0}^{12} \\ \imath_{\theta_0}^{21} & \imath_{\theta_0}^{21}(\imath_{\theta_0}^{11})^{-1}\imath_{\theta_0}^{12} \end{bmatrix}.$$

Even greater simplification is available using our results for partitioned matrices (Section 4 of the Matrices handout), although this is left as an exercise for the reader.

B.5.5 A Final Word

The analysis of Section B.5 has proceeded on the basis of the maximum likelihood estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$. However, it should be recognized that these results will still hold if the maximum likelihood estimators is replaced by some other consistent, asymptotically Normal estimators, and scores and likelihoods are replaced by their analogues for these other estimators. Specifically, these general testing principles can be applied almost immediately to GMM estimators, etc, to generate testing procedures as required. We will not pursue this further here but it is something of which you should be aware.