B9311-020 Introduction to Econometrics
Week 3 Lecture Notes
Estimators and Hypothesis Testing

Let Y denote an $n \times 1$ vector of observations with CDF $F(y, \theta)$. Let $\widehat{\theta} = g(Y)$ denote an *estimator* of θ.

- Example: *Method of Moments* Estimators find $\widehat{\theta}$ so that sample moments of Y match the population moments of Y.
  - Let Y_i, $i = 1, ..., n$ be scalar $NIID(\mu, \sigma^2)$ random variables. Then $E(Y) = \mu$ and $E[(Y - \mu)^2] = \sigma^2$. Natural estimators are therefore

    $$\widehat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i \text{ and } \widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{2} (Y_i - \widehat{\mu})^2$$

    which match sample to population moment conditions.:
  - Consider a set of random variables $X_1, X_2, ..., X_n$. Suppose that these variables are serially correlated so that X_i and X_j are correlated for i close the j. Suppose that we want to forecast X_{n+1} using X_n say as ϕX_n, with some constant ϕ. A good choice of ϕ is one that makes X_n uncorrelated with the forecast error $X_{n+1} - \phi X_n$. That is, ϕ satisfies

    $$E[(X_{n+1} - \phi X_n) X_n] = 0$$

    Thus, a method of moments estimators, say $\widehat{\phi}$ can be constructed to solve

    $$(n-1)^{-1} \sum_{i=1}^{n-1} (X_{i+1} - \widehat{\phi} X_i) X_i = 0$$

    so that

    $$\widehat{\phi} = \frac{\sum_{i=1}^{n-1} X_{i+1} X_i}{\sum_{i=1}^{n-1} X_i^2}$$

1 Properties of Estimators

- A natural question is what constitutes a "good" estimator. One way to answer this question is to define a Loss Function, say $L(\widehat{\theta}, \theta)$ which shows the loss that occurs when $\widehat{\theta}$ is used, when the true value of the parameter is θ. For any $\widehat{\theta} = g(Y)$, we could then calculate

  $$R(\widehat{\theta}, \theta) = E[L(\widehat{\theta}, \theta)] = E[L(g(Y), \theta)]$$

  the expected value of the loss. $R(\widehat{\theta}, \theta)$ is called the "Risk" Function
- A good estimator is an estimator that has small risk. The best estimator has the smallest risk.

- Often the risk of an estimator will depend on the value of θ (hence the notation $R(\widehat{\theta}, \theta)$) and thus the "best" estimator will depend the value of θ. Since θ is unknown we must find an estimator that works well for a range of values of θ. Examples
  - If we know that $\theta \in \Theta$, then we might try to find an estimator that solves
$$\min_{\widehat{\theta}} \max_{\theta \in \Theta} R(\widehat{\theta}, \theta)$$
    This produces a mini-max estimator.
  - We might want to find an estimator that minimizes the weighted average risk using a weight function $w(\theta)$. Thus we could consider
$$r(\widehat{\theta}) = \int R(\widehat{\theta}, \theta) w(\theta) d\theta$$
    which is called the average risk of $\widehat{\theta}$ (or the "Bayes-Risk"). The best estimator is the function $\widehat{\theta}$ that minimizes $r(\widehat{\theta})$.
- A useful loss function is
$$L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$$
  which is quadratic loss. The associated risk is called "mean squared error". Since
$$\widehat{\theta} - \theta = [\widehat{\theta} - E(\widehat{\theta})] + [E(\widehat{\theta}) - \theta)]$$
$$E[(\widehat{\theta} - \theta)^2] = E\{[\widehat{\theta} - E(\widehat{\theta})]^2\} + E\{[E(\widehat{\theta}) - \theta)]^2\} + 2E\{[\widehat{\theta} - E(\widehat{\theta})][E(\widehat{\theta}) - \theta)]\}$$
  so that
$$mse = Var(\widehat{\theta}) + [Bias(\widehat{\theta})]^2$$
  where the bias is defined by
$$Bias(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$
  - An estimator is *unbiased* if $Bias(\widehat{\theta}) = 0$, so that $E(\widehat{\theta}) = \theta$.

Often it is difficult to deduce the exact distribution of an estimator, and so various approximations based on large-sample theory are used. The relevant jargon is

- $\widehat{\theta}$ is consistent if $\widehat{\theta} \xrightarrow{p} \theta$.
- $\widehat{\theta}$ is strongly consistent if $\widehat{\theta} \xrightarrow{as} \theta$.
- Suppose some scaled and centered version of an estimator satisfies a CLT, *i.e.*,
$$a_n(\widehat{\theta} - \gamma) \xrightarrow{d} N(0, 1).$$
  where a_n is sequence of real numbers (like $a_n = \sqrt{n}$) and γ is a constant. We then say that $\widehat{\theta}$ is asymptotically normal.

– If

$$a_n(\widehat{\theta} - \gamma) \xrightarrow{d} N(0,1)$$

then (at least for n large)

$$a_n(\widehat{\theta} - \gamma) \overset{a}{\sim} N(0,1)$$

where I use the symbol $\overset{a}{\sim}$ to denote "approximately distributed as." Thus,

$$\widehat{\theta} \overset{a}{\sim} N(\gamma, \frac{1}{a_n^2})$$

For example, if $Y_i \sim iid(\mu, \sigma^2)$ then

$$\frac{\sqrt{n}}{\sigma}(\widehat{\mu} - \mu) \xrightarrow{d} N(0,1)$$

where $\widehat{\mu} = n^{-1} \sum Y_i$ suggesting

$$\widehat{\mu} \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$$

2 Cramer-Rao Inequality

The key question is how to construct good estimators. One very useful result in this regard is the Cramer-Rao inequality, which gives a lower bound on the variance of any unbiased estimator. First, some preliminaries:

- Suppose $Y \sim F(y, \theta)$ with density $f(y, \theta)$. Then

$$1 = \int f(y, \theta) dy$$

so that, differentiating both sides, and assuming the support of Y does not depend on θ

$$0 = \int \frac{\partial f(y, \theta)}{\partial \theta} dy$$

- Let

$$S(\theta, y) = \frac{\partial \ln f(y, \theta)}{\partial \theta}$$

which is called a *Score function.* (When I want to emphasize dependence of this function on θ I will write this function as $S(\theta)$.)

- Let

$$\mathcal{I}(\theta) = -E(\frac{\partial S(\theta, Y)}{\partial \theta}) = -E(\frac{\partial^2 \ln[f(Y, \theta)]}{\partial \theta^2})$$

denote the *Information*

- Note

$$\frac{\partial f(y,\theta)}{\partial \theta} = S(\theta,y) \times f(y,\theta)$$

thus

$$0 = \int \frac{\partial f(y,\theta)}{\partial \theta} dy = \int S(\theta,y) f(y,\theta) dy = E[S(\theta,Y)]$$

and thus the Score function has an expected value of 0. (Note the randomness in the score function comes from evaluating the function at the random value Y.)

- Differentiating again, yields:

$$0 = \int \frac{\partial S(\theta,y)}{\partial \theta} f(y,\theta) dy + \int S(\theta,y)^2 f(y,\theta) dy$$

so that

$$\mathcal{I}(\theta) = -E(\frac{\partial S(\theta,Y)}{\partial \theta}) = E(S(\theta,Y)^2) = var(S(\theta,Y))$$

- Now, let $\widehat{\theta} = g(Y)$ denote an unbiased estimator of θ. Then

$$\theta = \int g(y) f(y,\theta) dy$$

so that (differentiating both sides with respect to θ)

$$1 = \int g(y) S(\theta,y) f(y,\theta) dy$$

with $\widehat{\theta} = g(Y)$, this implies

$$E(\widehat{\theta}, S(\theta,Y)) = Cov(\widehat{\theta}, S(\theta,Y)) = 1$$

so that

$$Var(\widehat{\theta}) Var(S(\theta,Y)) \geq 1$$

and thus

$$Var(\widehat{\theta}) \geq \frac{1}{var(S(\theta,Y))} = \mathcal{I}(\theta)^{-1}$$

which is the Cramer-Rao inequality

- The same set of results obtain when θ is a $k \times 1$ vector
  - $S(\theta,Y)$ is a $k \times 1$ Score vector with $E(S(\theta,Y)) = 0$
  - $Var(S(\theta,Y)) = E(S(\theta,Y)S(\theta,Y)') = -E(\frac{\partial S(\theta,Y)}{\partial \theta'}) = \mathcal{I}(\theta)$ a $k \times k$ Information matrix
  - If $\widehat{\theta}$ is an unbiased estimator, the $E[(\theta - \widehat{\theta})(\theta - \widehat{\theta})'] \geq \mathcal{I}(\theta)^{-1}$

3 Properties of Maximum Likelihood Estimators

- Let Y denote a random vector with density $f(y, \theta)$. Then

$$\mathcal{L}(\theta) = f(Y, \theta),$$

the density of Y evaluated at $y = Y$ and viewed as function of θ is referred to as the *Likelihood Function.*

  - Let $Y_1, Y_2, ..., Y_n$ be *iid*, each with density $f(y, \theta)$. Then

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(Y_i, \theta)$$

  is the likelihood function.

- Let

$$L_n(\theta) = \ln(\mathcal{L}_n(\theta))$$

denote the log-likelihood function.

- Suppose that θ is a $k \times 1$ vector and let

$$s_i(\theta) = \frac{\partial \ln f(Y_i, \theta)}{\partial \theta}$$

and

$$S_n(\theta) = \sum_{i=1}^{n} s_i(\theta)$$

denote the Score. (Note that these functions are evaluated at the random value Y. For notational simplicity I write $s_i(\theta)$ instead of $s_i(\theta, Y_i)$, etc.)

- Let

$$\mathcal{I}_i(\theta) = -E(\frac{\partial s_i(\theta)}{\partial \theta'}) = E(s_i(\theta)s_i(\theta)'),$$

denote the information in the $i'th$ observation,

$$\mathcal{I}(\theta) = E(S_n(\theta)S_n(\theta)')$$

denote the information in the sample, and

$$\overline{\mathcal{I}}_n(\theta) = n^{-1} \sum \mathcal{I}_i(\theta) = n^{-1}\mathcal{I}(\theta)$$

denote the average information.

  - With independent sampling, the $s_i's$ are independent and so $\mathcal{I}(\theta) = \sum \mathcal{I}_i(\theta)$. With *iid* samling $\mathcal{I}_i(\theta) = \mathcal{I}_j(\theta) = \overline{\mathcal{I}}_n(\theta) = \overline{\mathcal{I}}(\theta)$, say

- Let $\widehat{\theta}_{mle}$ solve

$$\max_{\theta} L_n(\theta)$$

- Some asymptotic properties of MLEs
Given a set of "regularity" conditions:

  –

$$\widehat{\theta}_{mle} \xrightarrow{p} \theta_o$$

  and

$$\overline{\mathcal{I}}_n(\theta_o)^{1/2}\sqrt{n}(\widehat{\theta}_{mle} - \theta_o) \xrightarrow{d} N(0, I)$$

  so that

$$\widehat{\theta}_{mle} \overset{a}{\sim} N(\theta_o, \mathcal{I}(\theta_o)^{-1})$$

  where θ_o is the true value of θ.

  – Sketch of consistency proof under *iid* sampling:
  Let

$$C(\delta) = E_{\theta_o}[\ln(f(Y, \theta_o + \delta) - \ln(f(Y, \theta_o))]$$

  where θ_o is the true value of θ and E_{θ_o} means taking the expected value using the density $f(y, \theta_o)$. Then $C(\delta) < 0$ for $\delta \neq 0$ (the inequality will be weak only if the distribution of Y is degenerate.) To see this

$$E_{\theta_o} \ln[\frac{f(y, \theta_o + \delta)}{f(y, \theta_o)}] < \ln E_{\theta_o}[\frac{f(y, \theta_o + \delta)}{f(y, \theta_o)}] = \ln(1) = 0$$

  where the first inequality follows from Jensen's inequality (Rao, page 149) since the log function is concave. Clearly then $C(\delta)$ is maximized at $\delta = 0$. Also

$$C_n(\delta) = n^{-1} \sum \{\ln(f(Y_i, \theta_o + \delta) - f(Y_i, \theta_o)\} \xrightarrow{p} C(\delta)$$

  uniformly in δ. (This is *Uniform* LLN result — see Gallant, A. R. (1997), *An Introduction to Econometric Theory*, Princeton University Press., page 135). Thus the minimizer of $C_n(\delta)$ converges to the minimizer of $C(\delta)$, which we just showed was 0. Thus the minimizer of $n^{-1} \sum \ln(f(Y_i, \theta)$ converges to θ_o.

  – Sketch of Asymptotic Normality

    * First

$$\frac{1}{\sqrt{n}} S_n(\theta_o) \xrightarrow{d} N(0, \overline{\mathcal{I}}(\theta_o))$$

    follows immediately from applying the CLT to $\sum s_i(\theta)$.

    * Next

$$S_n(\widehat{\theta}_{mle}) = S_n(\theta_o) + \frac{\partial S_n(\widetilde{\theta})}{\partial \theta}(\widehat{\theta}_{mle} - \theta_o)$$

where $\widetilde{\theta}$ is between θ and $\widehat{\theta}_{mle}$. Since $S_n(\widehat{\theta}_{mle}) = 0$,

$$\sqrt{n}(\widehat{\theta}_{mle} - \theta_o) = [-\frac{1}{n}\frac{\partial S_n(\widetilde{\theta})}{\partial \theta}]^{-1}[\frac{1}{\sqrt{n}}S_n(\theta_o)]$$

and

$$[-\frac{1}{n}\frac{\partial S_n(\widetilde{\theta})}{\partial \theta}] \xrightarrow{p} \overline{\mathcal{I}}(\theta_o)$$

(LLN, CMT, Consistency of $\widehat{\theta}_{mle}$). Thus

$$\sqrt{n}(\widehat{\theta}_{mle} - \theta_o) \xrightarrow{d} N(0, \overline{\mathcal{I}}(\theta_o)^{-1})$$

by Slutsky's Theorem.

– These results also hold for vector $\widehat{\theta}_{mle}$ and vector values $S_n(\theta_o)$, etc.

3.0.1 Examples (to be worked out in class)

- $iidN(\mu, \sigma^2)$
- Binomial (n, p)
- Uniform $[0, \theta]$

4 Method of Moment Estimators

Suppose $Y_i, i = 1, ..., n$ is a sequence of $iid(\mu, \Sigma)$ random $l \times 1$ vectors.

- The method of moments estimator of μ is

$$\widehat{\mu}_{mm} = n^{-1}\sum Y_i.$$

  From the LLN and CLT, we have

$$\widehat{\mu}_{mm} \xrightarrow{as} \mu$$

  and

$$\sqrt{n}(\widehat{\mu}_{mm} - \mu) \xrightarrow{d} N(0, \Sigma).$$

  Notice that the estimator can be constructed and these properties obtained without knowing very much about the probability distribution of Y.

- Now suppose that $\mu = h(\theta_o)$ where μ is $l \times 1$, θ_o is $k \times 1$ with $k \leq l$. Our goal is the estimate θ_o. A Method of Moments estimator can be obtained by solving

$$\min_{\theta} J_n(\theta)$$

where

$$\begin{aligned} J_n(\theta) &= [\frac{1}{n}\sum_{i-1}^{n}(Y_i - h(\theta))]'[\frac{1}{n}\sum_{i-1}^{n}(Y_i - h(\theta))] \\ &= (\overline{Y} - h(\theta))'(\overline{Y} - h(\theta)) \end{aligned}$$

Let $\widehat{\theta}_{mm}$ denote the method of moments estimator. The properties of $\widehat{\theta}_{mm}$ can be derived in a way that parallels the discussion of the maximum likelihood estimator.

- Consistency follows by arguing that $J_n(\theta) \to J(\theta)$ and that $J(\theta)$ is minimized at $\theta = \theta_o$.
- Asymptotic normality is proved using the following steps
  * 1. Show the gradient evaluated at θ_o satisfies a CLT.
    The gradient is

$$g_n(\theta) = \frac{\partial J_n(\theta)}{\partial \theta} = -2[\frac{\partial h(\theta)}{\partial \theta'}]'(\overline{Y} - h(\theta))$$

so that

$$\sqrt{n} g_n(\theta_o) = -2[\frac{\partial h(\theta_o)}{\partial \theta'}]'[\sqrt{n}(\overline{Y} - h(\theta_o))] \xrightarrow{d} N(0, 4[\frac{\partial h(\theta_o)}{\partial \theta'}]'\Sigma[\frac{\partial h(\theta)}{\partial \theta'}]$$

2. Linearize $g_n(\widehat{\theta}_{mm})$ around $g_n(\theta_o)$ and solve for $\widehat{\theta}_{mm}$.

$$g_n(\widehat{\theta}_{mm}) = g_n(\theta_o) + \frac{\partial g_n(\widetilde{\theta})}{\partial \theta'}(\widehat{\theta}_{mm} - \theta_o)$$

where $\widetilde{\theta}$ is between θ_o and $\widetilde{\theta}$.

3. Show

$$\frac{\partial g_n(\widetilde{\theta})}{\partial \theta'} \xrightarrow{p} 2H$$

where

$$H = [\frac{\partial h(\theta_o)}{\partial \theta'}]'[\frac{\partial h(\theta_o)}{\partial \theta'}]$$

a constant, non-singular matrix.

$$\frac{\partial g_n(\theta)}{\partial \theta'} = 2[\frac{\partial h(\theta)}{\partial \theta'}]'[\frac{\partial h(\theta)}{\partial \theta'}] + m_n(\theta)(\overline{Y} - h(\theta))$$

where $m_n(\theta)$ denotes the derivatives of $\partial h(\theta)/\partial \theta'$ with respect to θ. Evaluating this expression at $\theta = \theta_o$, the second term vanishes in probability and the first term is $2H$.

4. Write

$$\sqrt{n}(\widehat{\theta}_{mm} - \theta_o) = [\frac{\partial g_n(\widetilde{\theta})}{\partial \theta'}]^{-1}[\sqrt{n} g_n(\theta_o)] \xrightarrow{d} N(0, H^{-1}[\frac{\partial h(\theta_o)}{\partial \theta'}]'\Sigma[\frac{\partial h(\theta)}{\partial \theta'}]H^{-1})$$

so that

$$\widehat{\theta}_{mm} \overset{a}{\sim} N(\theta_o, V_n)$$

where

$$V_n = (1/n)H^{-1}[\frac{\partial h(\theta_o)}{\partial \theta'}]'\Sigma[\frac{\partial h(\theta)}{\partial \theta'}]H^{-1}$$

1 General Framework

Suppose that we have two competing hypotheses about the distribution of a random variable Y :

- Hypothesis 1 will be called the *Null* and is written as

$$H_o : Y \sim F_o(Y)$$

- Hypothesis 2 will be called the *Alternative* and is written as

$$H_a : Y \sim F_a(Y)$$

- It is useful to categorize the errors in inference that we can make
  - We can say that H_a is true when H_o is true. This is called *Type 1 Error*
  - We can say that H_o is true when H_a is true. This is called *Type 2 Error*
- We will consider tests based on realizations of the random variable Y.
  - Specifically, we will define a region of the sample space, say W, and
    * Reject H_o (Accept H_a) if $Y \in W$
    * Otherwise Reject H_a (Accept H_o)
  - W is called a *Critical Region*
- Our goal is to find procedures for choosing W to minimize the probability of making errors. However, we can also always make the probability of type 1 error smaller by making W smaller, and make the probability of type 2 error smaller by making W larger.
  - A standard procedure in test design (procedures for choosing W) is therefore to fix the probability of type 1 error at some pre-specified value, and choose the critical region to minimize the probability of type 2 error.
  - The pre-chosen probability of type 1 error is called the *size* of the test
  - The probability of accepting H_a when H_a is true is called the *power* of the test.
    * $Power = 1 - P(type\ 2\ error)$
  - The hypothesis testing design problem is: Choose a test to maximize power subject to a pre-specified size.

2 Likelihood Ratio Tests and the Neyman-Pearson Lemma

The Neyman-Pearson Lemma says that power is maximized, subject to a size constraint, by choosing the critical region based on the likelihood ratio

$$LR(Y) = \frac{\mathcal{L}_a(Y)}{\mathcal{L}_o(Y)}$$

where $\mathcal{L}_a(Y)$ and $\mathcal{L}_o(Y)$ are the likelihoods under the alternative and null, respectively. The critical region for a test with size a is

$$W_\alpha = \{Y | LR(Y) > c_\alpha\}$$

where c_a is chosen so that

$$P\{LR(Y) > c_\alpha | Y \sim F_o\} = \alpha$$

The proof of this result is easy:

Suppose the random variables have a continuous distribution with density f_a and f_o under the alternative and null. Then $\mathcal{L}_o(Y) = f_o(Y)$ and $\mathcal{L}_a(Y) = f_a(Y)$. Let W_α denote the NP critical region. Let X_α denote any other critical region with size a. Note

$$W_\alpha = (W_\alpha \cap X_\alpha) \cup (W_\alpha \cap \widetilde{X}_\alpha)$$

and

$$X_\alpha = (X_\alpha \cap W_\alpha) \cup (X_\alpha \cap \widetilde{W}_\alpha)$$

Now:

$$\alpha = \int_{W_\alpha} f_o(y)dy = \int_{X_\alpha} f_o(y)dy$$

which implies

$$\int_{W_\alpha \cap \widetilde{X}_\alpha} f_o(y)dy = \int_{X_\alpha \cap \widetilde{W}_\alpha} f_o(y)dy$$

But, for any $Y \in W_\alpha$ (and hence in $Y \in (W_\alpha \cap \widetilde{X}_\alpha)$), $f_a(Y) > c_a f_o(Y)$, and for any $Y \in \widetilde{W}_\alpha$ (and hence in $Y \in (X_\alpha \cap \widetilde{W}_\alpha)$), $f_a(Y) < c_a f_o(Y)$. Thus

$$\int_{W_\alpha \cap \widetilde{X}_\alpha} f_a(y)dy > \int_{X_\alpha \cap \widetilde{W}_\alpha} f_a(y)dy$$

adding back in $\int_{W_\alpha \cap X_\alpha} f_a(y)dy$ yields

$$P(Y \in W_\alpha | Y \sim F_a) = \int_{W_\alpha} f_a(y)dy > \int_{X_\alpha} f_a(y)dy = P(Y \in X_a | Y \sim F_a)$$

3 Parametric Restrictions

Write the density of Y as $f(y, \theta)$, where θ is a $k \times 1$ vector of parameters. Suppose $\theta \in \Theta$, where

$$H_o : \theta \in \Theta_o$$

$$H_a : \theta \in \Theta_a$$

where $\Theta = \Theta_o \cup \Theta_a$ with $\Theta_o \cap \Theta_a = \emptyset$.

- Example: $Y_i \sim iidN(\mu, 1), i = 1, ..., n$

$$H_o : \mu = \mu_o$$

$$H_a : \mu = \mu_a$$

with $\mu_o \neq \mu_a$. Note

$$f(y, \mu) = (2\pi)^{-\frac{n}{2}} \exp[-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2]$$

and thus

$$\begin{aligned} lr(Y) &= \ln(LR(Y)) = \frac{1}{2}[\sum (Y_i - \mu_o)^2 - \sum (Y_i - \mu_a)^2] \\ &= a(\mu_o, \mu_a) + \sum Y_i(\mu_a - \mu_o) \end{aligned}$$

and thus when $\mu_a > \mu_o$, the LR test rejects for large values of $\sum Y_i$, or equivalently large values of $\overline{Y} = n^{-1} \sum Y_i$. Thus we can write the LR testing procedure as

  - Reject H_o when $\overline{Y} > c_a$ where c_a is chosen so that

$$P(\overline{Y} > c_a | \overline{Y} \sim N(\mu_o, \frac{1}{n}) = \alpha$$

    That is, the probability is calculated under the assumption that the sample was drawn from the null distribution.

  - Notice that the critical region is the same for any H_a with $\mu_a > \mu_o$. That is, we use the same critical region for

$$H_o : \mu = \mu_o$$

$$H_a : \mu > \mu_o$$

    Since the LR critical regions are the same for all of the simple hypotheses making up H_a and each is most powerful, then the LR procedure is said to be *Uniformly Most Powerful* for H_o vs. H_a in this instance. This is a general property of LR tests for simple null hypotheses versus "one-sided" alternatives.

- A useful summary of testing procedure for the case

$$H_o : \theta = \theta_o$$

$$H_a : \theta \neq \theta_a$$

is a "Power Function" which shows how the power of the test changes as a function of θ.

- Some Jargon:
  - When Θ_o contains a single point, then the null hypothesis is said to be *simple.* When Θ_o contains more than one point, then the null is said to be a *composite.* Similarly for the alternative. ($H_a : \mu > \mu_o$ is a composite alternative.)
  - The general form of the likelihood ratio used for testing is

$$LR = \frac{\max_{\theta \in \Theta_a} \mathcal{L}(\theta)}{\max_{\theta \in \Theta_o} \mathcal{L}(\theta)}$$

  - A test is *consistent* for $H_o : \theta = \theta_o$ vs. $H_a : \theta = \theta_a$ if $Power \to 1$ as $n \to \infty$.
    * Exercise: show the LR test for the normal mean is consistent
  - A test is *biased* if $power < size$ for some $\theta \in \Theta_a$
  - Suppose θ, a vector, is partitioned as $\theta = (\theta_1, \theta_2)$, where $\theta_1 \in \Theta_{1,o}$ under H_o, but θ_2 is unrestricted. A critical region (or test) is *Similar* if

$$P(Y \in W_a | \theta_{1,o}, \theta_2)$$

    does not depend on θ_2.
    * In the normal mean example, suppose that σ^2 is unknown. The LR test is not similar, since the distribution of $\overline{Y}$ depends on σ^2
    * A $t - test$ for a normal mean is similar. (The distribution of the test statistic does not depend on σ^2).
    * A test statistic with a distribution that does not depend on *nuisance* parameters is said to be *pivotal*
  - A test is *Invariant* if the results are invariant pre-specified transformations of the data.
    * The t-test for the normal mean problem

$$H_o : \mu = 0 \text{ versus } H_o : \mu > 0$$

      is invariant to transformations of the form $X = aY$ where $a > 0$.

4 Likelihood Ratio Test Statistics

We are interested in testing

$$H_o : \theta = \theta_o \,\text{versus}\; H_a : \theta \neq \theta_o$$

where θ is a $k \times 1$ vector using a likelihood ratio test. To carry the test we have to choose the critical region W or equivalently, the critical value c_α. Recall c_a is determined by the requirement that $P(LR > c_a | H_o) = \alpha$, and thus to determine the critical value we need to know the probability distribution of LR when the null hypothesis is true. We now develop a large-sample approximation to solve this problem.

Let $\widehat{\theta} = \arg\max_\theta \mathcal{L}(\theta)$ denote the MLE of θ and write the maximized likelihood ratio as

$$LR = \frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\theta_o)}$$

Define the Likelihood Ratio Statistic as

$$\xi_{LR} = 2(\ln(LR)) = 2[L_n(\widehat{\theta}) - L_n(\theta_o)]$$

where $L_n(\theta) = \ln(\mathcal{L}(\theta))$. Since ξ_{LR} is a monotonic transformation of the likelihood ratio, the LR test can be implemented by rejecting the null when ξ_{LR} exceeds a pre-specified critical value.

To derive the approximate distribution of ξ_{LR} under the null hypothesis, write

$$L_n(\theta_o) = L_n(\widehat{\theta}) + (\theta_o - \widehat{\theta})' \frac{\partial L_n(\widehat{\theta})}{\partial \theta} + \frac{1}{2}(\theta_o - \widehat{\theta})' \frac{\partial^2 L_n(\widetilde{\theta})}{\partial \theta \partial \theta'}(\theta_o - \widehat{\theta})$$

where $\widetilde{\theta}$ is between θ_o and $\widehat{\theta}$. Since

$$\frac{\partial L_n(\widehat{\theta})}{\partial \theta} = 0$$

$$\begin{aligned}
\xi_{LR} &= -(\widehat{\theta} - \theta_o)' \frac{\partial^2 L_n(\widetilde{\theta})}{\partial \theta \partial \theta'}(\widehat{\theta} - \theta_o) \\
&= [\sqrt{n}(\widehat{\theta} - \theta_o)]'[-\frac{1}{n}\frac{\partial^2 L_n(\widetilde{\theta})}{\partial \theta \partial \theta'}][\sqrt{n}(\widehat{\theta} - \theta_o)]
\end{aligned}$$

From our earlier results

$$\sqrt{n}(\widehat{\theta} - \theta_o)] \overset{d;H_o}{\rightarrow} N(0, \overline{\mathcal{I}}(\theta_o)^{-1})$$

and

$$[-\frac{1}{n}\frac{\partial^2 L_n(\widetilde{\theta})}{\partial \theta \partial \theta'}] = -\frac{1}{n}\sum \frac{\partial^2 \ln f(Y_i, \widetilde{\theta})}{\partial \theta \partial \theta'} \overset{p;H_o}{\rightarrow} \overline{\mathcal{I}}(\theta_o)$$

so that

$$\xi_{LR} \overset{d;H_o}{\rightarrow} \xi \sim \chi^2_k.$$

(This final result follows from noting that ξ_{LR} is asymptotically a quadratic form of a $N(0, \overline{\mathcal{I}})$ variable around the inverse of it's covariance matrix.)

5 Wald Test Statistics

A close cousin of the LR statistic is the Wald statistic

$$\xi_W = (\widehat{\theta} - \theta_o)'[-\frac{\partial^2 L_n(\widehat{\theta})}{\partial\theta\partial\theta'}](\widehat{\theta} - \theta_o)$$

which differs from ξ_{LR} only because the estimated information matrix is evaluated at $\widehat{\theta}$ rather than $\widetilde{\theta}$. Since both $\widehat{\theta}$ and $\widetilde{\theta}$ converge in probability to θ_o under the null,

$$\xi_W \overset{p;H_o}{\rightarrow} \xi_{LR}$$

The motivation behind the Wald test is straightforward

$$\sqrt{n}(\widehat{\theta} - \theta_o) \overset{p;H_o}{\rightarrow} N(0, \overline{\mathcal{I}}^{-1})$$

so that

$$\widehat{\theta} \overset{a;H_o}{\sim} N(\theta_o, \frac{1}{n}\overline{\mathcal{I}}^{-1})$$

Recall

$$\overline{\mathcal{I}} = -E[\frac{1}{n}\frac{\partial^2 L_n(\theta_o)}{\partial\theta\partial\theta'}],$$

thus

$$n\overline{\mathcal{I}} = -E[\frac{\partial^2 L_n(\theta_o)}{\partial\theta\partial\theta'}]$$

Dropping the expectation operator E and evaluating the second derivative matrix at $\widehat{\theta}$ yields the approximation

$$\widehat{\theta} \overset{a;H_o}{\sim} N(\theta_o, [-\frac{\partial^2 L_n(\widehat{\theta})}{\partial\theta\partial\theta'}]^{-1})$$

so that

$$\xi_W = (\widehat{\theta} - \theta_o)'[-\frac{\partial^2 L_n(\widehat{\theta})}{\partial\theta\partial\theta'}](\widehat{\theta} - \theta_o) \overset{a;H_o}{\sim} \chi^2_k$$

When the alternative is true, then $\widehat{\theta} \approx \theta_a \neq \theta_o$ and so we expect large positive values of ξ_W and ξ_{LR} and hence the null is rejected in favor of the alternative for large values of the test statistics.

6 Score/Lagrange Multiplier Test Statistic

Another approximation to ξ_{LR} is give by the Score or Lagrange Multiplier test statistic:

$$\begin{aligned}
\xi_{LM} &= [S_n(\theta_o)]'[-\frac{\partial^2 L_n(\theta_o)}{\partial\theta\partial\theta'}]^{-1}[S_n(\theta_o)] \\
&= [\frac{1}{\sqrt{n}}S_n(\theta_o)]'[-\frac{1}{n}\frac{\partial^2 L_n(\theta_o)}{\partial\theta\partial\theta'}]^{-1}[\frac{1}{\sqrt{n}}S_n(\theta_o)]
\end{aligned}$$

Since

$$\frac{1}{\sqrt{n}}S_n(\theta_o) \overset{d,H_o}{\longrightarrow} N(0,\overline{\mathcal{I}})$$

and

$$[-\frac{1}{n}\frac{\partial^2 L_n(\theta_o)}{\partial\theta\partial\theta'}] \overset{p,H_o}{\longrightarrow} \overline{\mathcal{I}}$$

then

$$\xi_{LM} \overset{a,H_o}{\sim} \chi^2_k$$

follows directly. It is also straightforward to show (you should) that

$$\xi_{LM} \overset{p,H_o}{\longrightarrow} \xi_{LR} \overset{p,H_o}{\longrightarrow} \xi_W$$

An alternative form of the LM statistic uses another approximation for $\overline{\mathcal{I}}$

$$\overline{\mathcal{I}} \approx \frac{1}{n}\sum s_i(\theta_o)s_i(\theta_o)'$$

Since

$$S_n(\theta_o) = \sum s_i(\theta_o)$$

this version of LM test statistic can then be written as

$$\xi_{LM} = [\sum s_i(\theta_o)]'[\sum s_i(\theta_o)s_i(\theta_o)']^{-1}[\sum s_i(\theta_o)]$$

In the second part of this course, you will recognize this as the fitted sum of squares from the regression of a vector of $1's$ onto $s_i(\theta_o)$.

7 Confidence Intervals

A $(1-\alpha)\times 100\%$ confidence interval for θ is the set of values of θ that cannot be rejected, when taken as the null values for a test with size α. These are easily calculated from the Wald Statistic. Let

$$\widehat{V} = [-\frac{\partial^2 L_n(\widehat{\theta})}{\partial\theta\partial\theta'}]^{-1}$$

denote the estimated covariance matrix from $\widehat{\theta}$. Then the Wald statistic is

$$\xi_W = (\widehat{\theta}-\theta_o)'\widehat{V}^{-1}(\widehat{\theta}-\theta_o)$$

and $H_o : \theta = \theta_o$ is not rejected using a test of size α if

$$\xi_W \leq \chi^2_{\kappa,1-\alpha}$$

where $\chi^2_{k,1-\alpha}$ denotes the $1-\alpha$ quantile of the χ^2_k distribution. The confidence interval is therefore

$$\{\theta|(\widehat{\theta}-\theta)'\widehat{V}^{-1}(\widehat{\theta}-\theta) \leq \chi^2_{\kappa,1-\alpha}\}$$

which is recognized as the interior of an ellipse centered at $\theta = \widehat{\theta}$.

In the one dimensional case ($k=1$), the normal distribution can be used in the place of the χ^2 yielding

$$\{\theta|\widehat{\theta} - Z_{1-\frac{\alpha}{2}}\widehat{V}^{-\frac{1}{2}} \leq \theta \leq \widehat{\theta} + Z_{1-\frac{\alpha}{2}}\widehat{V}^{-\frac{1}{2}}\}$$

where $Z_{1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$ ordinate of the $N(0,1)$ distribution.

8 Nuisance Parameters in Testing

In many application a null hypothesis specifies values for some the parameters but leaves the other unknown parameters unrestricted. How does this affect the testing procedures discussed above?

Let θ denote a $p \times 1$ vector of unknown parameters partitioned as

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

where θ_1 is $k \times 1$ and θ_2 is $(p-k) \times 1$. Suppose that the hypotheses of interest are

$$H_o : \theta_1 = \theta_{1,o} \text{ versus } H_a : \theta_1 \neq \theta_{1,o}$$

with θ_2 unspecified under the null and alternative.

Let

$$\widehat{\theta} = \arg\max_{\theta} L(\theta)$$

and

$$\widetilde{\theta} = \arg\max_{\theta} L(\theta) \text{ subject to } \theta_1 = \theta_{1,o}$$

and partition these as

$$\widehat{\theta} = \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{bmatrix} \text{ and } \widetilde{\theta} = \begin{bmatrix} \theta_{1,o} \\ \widetilde{\theta}_2 \end{bmatrix}$$

The LR statistic is

$$\xi_{LR} = 2[L(\widehat{\theta}) - L(\widetilde{\theta})]$$

and the Wald statistic is

$$\xi_W = [\sqrt{n}(\widehat{\theta}_1 - \theta_{1,o})]'[\overline{\mathcal{I}}^{11}]^{-1}[\sqrt{n}(\widehat{\theta}_1 - \theta_{1,o})]$$

where

$$\overline{\mathcal{I}}^{-1} = \begin{bmatrix} \overline{\mathcal{I}}^{11} & \overline{\mathcal{I}}^{12} \\ \overline{\mathcal{I}}^{21} & \overline{\mathcal{I}}^{22} \end{bmatrix}$$

Since

$$\sqrt{n}(\widehat{\theta} - \theta_o) \xrightarrow{d,H_o} N(0, \overline{\mathcal{I}}^{-1})$$

$$\xi_W \xrightarrow{d,H_o} \chi^2_k$$

follows immediately.

Before working out the distribution of the LR statistics, first some preliminary calculations. Let

$$S_n(\theta_o) = \begin{bmatrix} S_{n1}(\theta_o) \\ S_{n2}(\theta_o) \end{bmatrix}$$

where

$$S_{n1}(\theta_o) = \frac{\partial L_n(\theta_o)}{\partial \theta_1} \text{ and } S_{n2}(\theta_o) = \frac{\partial L_n(\theta_o)}{\partial \theta_2}$$

and recall

$$\begin{bmatrix} \frac{1}{\sqrt{n}} S_{n1}(\theta_o) \\ \frac{1}{\sqrt{n}} S_{n2}(\theta_o) \end{bmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \overline{\mathcal{I}}_{11} & \overline{\mathcal{I}}_{12} \\ \overline{\mathcal{I}}_{21} & \overline{\mathcal{I}}_{22} \end{bmatrix} \right.$$

From earlier we have

$$\sqrt{n}(\widehat{\theta} - \theta_o) = \overline{\mathcal{I}}^{-1}[\frac{1}{\sqrt{n}} S_n(\theta_o)] + o_p(1)$$

Also

$$S_{n2}(\theta_{1,o}, \widetilde{\theta}_2) = S_{n2}(\theta_{1,o}, \theta_{2,o}) + \frac{\partial S_{n2}(\theta_{1,o}, \overline{\theta}_2)}{\partial \theta_2}(\widetilde{\theta}_2 - \theta_{2,o})$$

where $\overline{\theta}_2$ is between $\widetilde{\theta}_2$ and $\theta_{2,o}$. Thus,

$$\sqrt{n}(\widetilde{\theta}_2 - \theta_{2,o}) = \overline{\mathcal{I}}_{22}{}^{-1}[\frac{1}{\sqrt{n}} S_{n2}(\theta_{1,o}, \theta_{2,o})] + o_p(1)$$

and

$$S_{n1}(\theta_{1,o}, \widetilde{\theta}_2) = S_{n1}(\theta_{1,o}, \theta_{2,o}) + \frac{\partial S_{n1}(\theta_{1,o}, \overline{\theta}_2)}{\partial \theta_2}(\widetilde{\theta}_2 - \theta_{2,o})$$

so that

$$\frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \widetilde{\theta}_2) = \frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \theta_{2,o}) + [\frac{1}{n}\frac{\partial S_{n1}(\theta_{1,o}, \overline{\theta}_2)}{\partial \theta_2}]\sqrt{n}(\widetilde{\theta}_2 - \theta_{2,o})$$

$$\frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \widetilde{\theta}_2) = [I \quad -\overline{\mathcal{I}}_{12}\overline{\mathcal{I}}_{22}{}^{-1}] \begin{bmatrix} \frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \theta_{2,o}) \\ \frac{1}{\sqrt{n}} S_{n2}(\theta_{1,o}, \theta_{2,o}) \end{bmatrix}$$

So that

$$\frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \widetilde{\theta}_2) \xrightarrow{d, H_o} N(0, ABA')$$

where

$$A = [I \quad \overline{\mathcal{I}}_{12}\overline{\mathcal{I}}^{22}] \text{ and } B = \overline{\mathcal{I}}$$

A straightforward calculation shows

$$ABA' = [\overline{\mathcal{I}}_{11} - \overline{\mathcal{I}}_{12}\overline{\mathcal{I}}_{22}{}^{-1}\overline{\mathcal{I}}_{21}]$$

The Score Test can then be formed as

$$\xi_{LM} = [\frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \widetilde{\theta}_2)]'[\overline{\mathcal{I}}_{11} - \overline{\mathcal{I}}_{12}\overline{\mathcal{I}}_{22}{}^{-1}\overline{\mathcal{I}}_{21}]^{-1}[\frac{1}{\sqrt{n}} S_{n1}(\theta_{1,o}, \widetilde{\theta}_2)]$$

Using any of the estimators of $\overline{\mathcal{I}}$ that we discussed last time.

Finally, from last time

$$\begin{aligned} 2[L_n(\widehat{\theta}) - L_n(\theta_o)] &= [\sqrt{n}(\widehat{\theta} - \theta_o)]'\overline{\mathcal{I}}[\sqrt{n}(\widehat{\theta} - \theta_o)] + o_p(1) \\ &= [\frac{1}{\sqrt{n}} S_n(\theta_o)]'\overline{\mathcal{I}}^{-1}[\frac{1}{\sqrt{n}} S_n(\theta_o)] + o_p(1) \end{aligned}$$

and a similar calculation shows

$$\begin{aligned} 2[L_n(\tilde{\theta}) - L_n(\theta_o)] &= [\sqrt{n}(\tilde{\theta}_2 - \theta_{2,o})]'\mathcal{I}_{22}[\sqrt{n}(\tilde{\theta}_2 - \theta_{2,o})] + o_p(1) \\ &= [\frac{1}{\sqrt{n}}S_{n2}(\theta_{1,o}, \theta_{2,o})]'[\mathcal{I}_{22}]^{-1}[\frac{1}{\sqrt{n}}S_{n2}(\theta_{1,o}, \theta_{2,o})] + o_p(1) \end{aligned}$$

so that

$$\begin{aligned} \xi_{LR} &= 2[L_n(\hat{\theta}) - L_n(\tilde{\theta})] = \\ & [\frac{1}{\sqrt{n}}S_n(\theta_o)]'\mathcal{I}^{-1}[\frac{1}{\sqrt{n}}S_n(\theta_o)] - [\frac{1}{\sqrt{n}}S_{n2}(\theta_{1,o}, \theta_{2,o})]'[\mathcal{I}_{22}]^{-1}[\frac{1}{\sqrt{n}}S_{n2}(\theta_{1,o}, \theta_{2,o})] + o_p(1) \end{aligned}$$

and a straightforward calculation using the partitioned inverse formula, shows that this is the same as the LM statistic (and Wald statistic) up to a term that is $o_p(1)$.

9 Testing Restrictions on Parameters

Thus far we have considered testing restrictions on θ that take the form

$$\theta = \theta_o$$

which restricts all of the elements or θ, or

$$[I_k \quad 0_{k\times p}]\theta = \theta_{1,o}$$

which restricts the first k elements. Suppose that instead we are interesting in the restriction

$$R\theta = r$$

where R is a $k \times p$ matrix with full row rank and $k \leq p$. If $k = p$, then since $R\theta = r$ implies that $\theta = R^{-1}r$, we are just in the first situation with $\theta_o = R^{-1}r$.

When $k < p$ then we are in the second situation. To see this, consider multiplying θ by a full rank matrix G with

$$G = \begin{bmatrix} R \\ R^{\perp} \end{bmatrix}$$

where $R^{\perp}$ is a $(p-k) \times p$ matrix with rows with full row rank and with rows orthogonal to the rows of R. (There are a variety or ways to compute the rows of $R^{\perp}$.) Then we can reparametrize the likelihood using

$$\delta = G\theta$$

instead of θ. Partitioning δ as

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$$

where $\delta_1 = R\theta$ and $\delta_2 = R^{\perp}\theta$. The hypothesis that $R\theta = r$ corresponds to $\delta_1 = r$ with δ_2 unrestricted.

There is not need to explicitly re-parameterize the model to carry out the test, once you note that the reparamerization will lead to

$$\widehat{\delta} = G\widehat{\theta}$$

a straightforward calculation (that you should verify) shows that the Wald statistic is given by:

$$\xi_W = (R\widehat{\theta} - r)'[R\{-\frac{\partial^2 L_n(\widehat{\theta})}{\partial\theta\partial\theta'}\}^{-1}R']^{-1}(R\widehat{\theta} - r)$$

and the formula for ξ_{LR} is unchanged except for the fact that $\widetilde{\theta}$ is now the MLE subject to the constraint that $R\theta = r$. Finally

$$\xi_{LM} = [\sum s_i(\widetilde{\theta})]'[\sum s_i(\widetilde{\theta})s_i(\widetilde{\theta})']^{-1}[\sum s_i(\widetilde{\theta})]$$

can be used to construct the LM statistic. (Exercise: show this.)