University of Illinois Department of Economics
Spring 2017 Econ 574 Roger Koenker

# Lecture 11
# "Asymptotic Theory of Testing"
(Introduction to the Holy Trinity)

## The likelihood ratio test

Suppose we begin with a problem of testing a simple null hypothesis

$$H_0 : \theta = \theta_0$$

against

$$H_1 : \theta \neq \theta_0$$

For a specific alternative, $\theta = \theta_1$, we "know" that the best test of $H_0$ has critical regions of the form, (Neyman-Pearson Lemma),

$$\left\{ z \mid \frac{L(\theta_1, z)}{L(\theta_0, z)} > c \right\}$$

When the alternative is not simple, then we can sometimes achieve a uniformly most powerful UMP test by using critical regions of the form

$$\left\{ z \mid \frac{L(\hat{\theta}_n, z)}{L(\theta_0, z)} > c \right\}$$

where $\hat{\theta}_n$ denotes the MLE. Note that we simply replace $L$ evaluated at the specific alternative with $L$ evaluated at the best alternative as selected by the principle of maximimum likelihood. For *finite* $n$, it is typically difficult to find the *exact* distribution of the likelihood ratio, but the situation is nice in the following extended example.

*Remark:* Economist's Heuristics for the Neyman-Pearson Lemma

Cosma Shalizi of Carnegie Mellon has suggested the following nice heuristic argument for the optimality of likelihood ratio tests. We want to choose a rejection region, $R$ such that under the alternative hypothesis, the probability of falling into $R$, which we will denote, $Q(R)$ is as large as possible, given that we satisfy a constraint that the probability falling into $R$ under the null hypothesis, denoted by $P(R)$ doesn't exceed some specified level $\alpha$. Then Lagrange tells us to solve,

$$\max\{Q(R) - \lambda(P(R) - \alpha\}$$

over $R$ and $\lambda$. Note that the first term, usually called the power of the test is a Good Thing, and the second term involving $P(R)$, usually called the type I error is a Bad Thing, hence the minus sign before the Lagrange multiplier to make the terms of commensurate goodness. Now it looks like we have a very high dimensional problem involving choice the the set $R$ and the real value $\lambda$, which would lead us into the thickets of the calculus of variations. But if we think of this problem

as choosing $R$ to maximize benefits minus costs, the inscrutable logic of economics tells us that we need to set marginal benefit equal to marginal cost. And if we interpret this in the present context to mean that small changes in the size of $R$ should equate the change in $Q(R)$ with $\lambda$ times the change in $P(R)$, we see that this is equivalent to the rule that $\Delta Q(R)/\Delta P(R) = \lambda$ on the boundary of $R$. Thus the Lagrange multiplier $\lambda$ is seen to be the critical value of a test, or the shadow price of power in units of Type I error. In differentiable settings the $dQ/dP$ object is just the ratio of the densities giving us the likelihood ratio.

*Example:* Linear Model

Suppose $y = x\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma^2$ *known.* We would like to test $H_0 : \beta = \beta_0$. The MLE is $\hat{\beta} = (X'X)^{-1}X'y$, the likelihood is,

$$L(\beta, y) = (2\pi\sigma^2)^{-n/2} \exp\{-\sum(y_i - x_i\beta)^2/2\sigma^2\}$$

so,

$$\lambda \equiv L(\hat{\beta}, y)/L(\beta_0, y) = \exp\{\frac{1}{2\sigma^2}(\hat{\beta} - \beta_0)'(X'X)(\hat{\beta} - \beta_0)\}$$

and

$$2\log\lambda = (\hat{\beta} - \beta_0)'(X'X)(\hat{\beta} - \beta_0)/\sigma^2$$

But *under* $H_0$, $\hat{\beta} \sim \mathcal{N}_p(\beta_0, \sigma^2(X'X)^{-1})$ so $2\log\lambda \sim \chi^2_p$. Thus, an *exact* level $\alpha$ test may be constructed by rejecting $H_0$ when $2\log\lambda$ exceeds the $\alpha$ critical value of $\chi^2_p$. When $\sigma^2$ is unknown, we do essentially the same thing, but replace $\sigma^2$ with $s^2$ and look up the critical value in $F_{p,n-p}$, rather than $\chi^2_p$.

> **Digression on non-central** $\chi^2$. If $Z_1, \ldots, Z_p$ are independently distributed as $\mathcal{N}(\xi_i, \sigma^2)$, then $\sum_{i=1}^p Z_i^2/\sigma^2 \sim \chi^2_p(\delta)$ non-central $\chi^2$ with $p$ degrees of freedom and noncentrality parameter $\delta$ where $\delta = \sum \xi_i^2/\sigma^2$. *Caveat:* Sometimes $\sqrt{\delta}$ is used, or
>
> $\delta^2/2$, depending upon what the author had for breakfast. More generally, if the $p$-vector $Z$ is $\mathcal{N}(\xi, \Omega)$, then $Z'\Omega^{-1}Z \sim \chi^2_p(\delta)$, where $\delta = \xi'\Omega^{-1}\xi$.

**Q.** What happens in this case as $n \to \infty$? **A.** When $H_0$ is true, then nothing much happens, i.e., $2\log\lambda$ is $\chi^2_p(0)$, i.e. central chi-square for all $n$. But, if $H_0$ is false, so $\beta = \beta_1 \neq \beta_0$, then, $\hat{\beta} \sim \mathcal{N}(\beta_1, \sigma^2(X'X)^{-1})$ and therefore, $\hat{\beta} - \beta_0 \sim \mathcal{N}(\beta_1 - \beta_0, \sigma^2(X'X)^{-1})$ we have $2\log\lambda \sim \chi^2_p(\delta_n)$ where $\delta_n = (\beta_1 - \beta_0)'(X'X)(\beta_1 - \beta_0)/\sigma^2$. What happens to $\delta_n$ as $n \to \infty$. Of course, this depends upon what happens to $(X'X)$. Our standard assumption on $X$ is

$$\lim_{n\to\infty} n^{-1}X'X = D >> 0.$$

so for large $n$ we have

$$\delta_n \approx n(\beta_1 - \beta_0)'D(\beta_1 - \beta_0)/\sigma^2$$

Thus, for larger and larger $n$, $E\chi^2_p(\delta) = p + \delta$ and $V\chi^2_p(\delta) = 2p + 4\delta$ so both mean and variance are blowing up. What does this imply about Probability of Type II error, i.e., probability of saying $H_0$ is true when it isn't? It goes to zero, and we say the test is *consistent.*

Suggested Exercise: Use Chebyshev and the moments above to prove this.

This isn't a very satisfactory state of affairs, since it does not really provide sufficient detail about what the probability of a type II error is. To compare tests we'd like to investigate power under a more challenging state of affairs. For fixed alternatives and fixed significance levels, life is too easy, and all we can say is that eventually we will reject $H_0$ if it is false.

There are three approaches to studying the asymptotic power of tests:

| | | |
|---|---|---|
| (i) | let $\alpha \to 0$ | (Bayes) |
| (ii) | look at rates at which power $\to 1$ | (Bahadur) |
| (iii) | let the alternative shrink toward $\theta_0$ | (Pitman) |

We will focus on the third. Suppose instead of fixing the alternative $\beta = \beta_1$ we choose instead a "sequence of local alternatives," so $\beta_1 = \beta_0 + \xi/a_n$ where $\{a_n\}$ is a sequence $\to \infty$, chosen to stabilize the power of the test. In the present example the obvious choice is $a_n = \sqrt{n}$, i.e., $\beta_0 + \xi/\sqrt{n}$. Since, then

$$2\log\lambda = (\hat{\beta} - \beta_0)(X'X)(\hat{\beta} - \beta_0)/\sigma^2 \to \chi^2_p(\delta_0)$$

where $\delta_n = n^{-1}\xi'(X'X)\xi \to \xi' D \xi \equiv \delta_0$. This concludes the normal linear model example.

Wilks Test **General theory of Likelihood Ratios $2\log\lambda$.**

We will begin by considering the scalar case, and then generalize. Expanding around $\hat{\theta}$, under $H_0$,

$$l(\theta_0) = l(\hat{\theta}) + (\theta_0 - \hat{\theta})l(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2 l''(\theta^*)$$

where $|\theta^* - \theta_0| \le |\hat{\theta} - \theta_0|$ so

$$\begin{aligned} 2\log\lambda &= 2(l(\hat{\theta}) - l(\theta_0)) \\ &= 2\{l(\hat{\theta}) - l(\hat{\theta}) - (\theta_0 - \hat{\theta})l'(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^2 l''(\theta^*)\} \\ &= -(\hat{\theta} - \theta_0)^2 l''(\theta^*). \end{aligned}$$

But, since $\hat{\theta} \xrightarrow{p} \theta_0, n^{-1}l''(\theta^*) \xrightarrow{p} -I(\theta_0)$ and $\sqrt{n}(\hat{\theta} - \theta_0) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, I(\theta_0)^{-1})$ we have

$$2\log\lambda = -n(\hat{\theta} - \theta_0)^2(\frac{1}{n}l''(\theta^*)) \rightsquigarrow \chi^2_1(0)$$

**The Wald Test**

The Wald test based on the discrepancy $(\hat{\theta} - \theta_0)$ obviously has the same asymptotic behavior under $H_0$ and $H_1$ as $2\log\lambda$.

**The Rao Test**

This form of the test is also known as the score or Lagrange-multiplier test and is based on the size of the gradient in the direction of the alternative when evaluated at the null. Expand $l'(\theta_0)$ around $\theta = \hat{\theta}$ to obtain

$$l'(\theta_0) = l'(\hat{\theta}) + (\theta_0 - \hat{\theta})l''(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2 l'''(\theta^*)$$

or

$$\begin{aligned}\sqrt{n}\frac{1}{n}l'(\theta_0) &= -\sqrt{n}(\hat{\theta}-\theta_0)(\frac{1}{n}l''(\hat{\theta})) + \frac{1}{2}\sqrt{n}(\theta_0-\hat{\theta})^2\frac{1}{n}l'''(\theta^*) \\ &\xrightarrow{p} \sqrt{n}(\hat{\theta}-\theta_0)I(\theta_0) \\ &\rightsquigarrow \mathcal{N}(I(\theta_0)\xi, I(\theta_0))\end{aligned}$$

hence,

$$\frac{1}{n}(l'(\theta_0))'I(\theta_0)^{-1}l'(\theta_0) \overset{\mathcal{D}}{\rightsquigarrow} \chi^2_p(\delta)$$

or equivalently,

$$-(l'(\theta_0))'(\frac{1}{n}l''(\theta_0))^{-1}l'(\theta_0) \overset{\mathcal{D}}{\rightsquigarrow} \chi^2_1(\delta).$$

The expansion shows that this is asymptotically equivalent to the other two tests under $H_0$ *and* $H_1$.

Why is the LM test called the Lagrange multiplier test? Suppose we consider the simple case

$$\max l(\theta) - \lambda(\theta - \theta_0)$$

we get first order condition $l'(\theta_0) = \lambda$ so the test based on the score function is the same as the test based on the Lagrange multiplier. This generalizes in a nice way, see GM section 17.2.2.

**Multivariate Extensions**

Now suppose that $\Theta \in \mathbb{R}^2$, and $H_0 : \theta = \theta_0$ and $H_n : \theta = \theta_n = \theta_0 + \xi/\sqrt{n}$.

*Theorem:* (Asymptotic Equivalence of the Holy Trinity) Under the Lehmann Conditions, and $H_n$, the test statistics

$$\begin{aligned} W &= n(\hat{\theta}_n - \theta_0)' I(\theta_0)(\hat{\theta}_n - \theta_0) \\ LR &= 2(l(\hat{\theta}_n) - l(\theta_0)) \\ LM &= n^{-1}(\nabla l(\theta_0))' I(\theta_0)^{-1} \nabla l(\theta_0) \end{aligned}$$

all converge in distribution to $\chi^2_p(\delta)$, with $\delta = \xi' I(\theta_0)\xi$

*Proof:* Under $H_n$, $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(\xi, I(\theta_0)^{-1})$ so it follows immediately that $W \rightsquigarrow \chi^2_p(\delta)$. Expanding $l(\cdot)$ around $\hat{\theta}_n$ as in the scalar case yields $LR \rightsquigarrow \chi^2_p(\delta)$. Expanding $\nabla l(\cdot)$ around $\hat{\theta}_n$ as in the scalar case yields

$$n^{-1/2}\nabla l(\theta_0) = n^{-1/2}\nabla l(\hat{\theta}_n) + \sqrt{n}(\theta_0 - \hat{\theta})'(\nabla^2 l(\hat{\theta}_n)/n) + o_p(1) \rightsquigarrow \mathcal{N}(\xi' I(\theta_0), I(\theta_0))$$

so $LM \rightsquigarrow \chi^2_p(\delta)$.

**Composite Hypotheses I**

Often the null hypothesis involves only a subset of the parameters, and the remaining ones must be treated as nuisance parameters. Partition $\theta = (\theta_1, \theta_2)$ and consider $H_0 : \theta_1 = \theta_{10}$ vs. $H_n : \theta_1 = \theta_{10} + \xi/\sqrt{n}$, with no restriction on $\theta_2$.

Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ denote the unrestricted and restricted mle's respectively, so $\hat{\theta}_n = (\theta_{10}, \tilde{\theta}_2)$. Partition $\nabla l = (s_1(\theta), s_2(\theta))'$ and $I(\theta) = (I_{ij}(\theta))$. Let $I^{ij}$ denote the $ij$ block of $I^{-1}$.

*Theorem:* (AE/HT - II ) Under Lehmann's Conditions, and $H_n$,

$$\begin{aligned} W &= n(\hat{\theta}_1 - \theta_{10})' I^{11}(\theta_0)^{-1}(\hat{\theta}_1 - \theta_{10}) \\ LR &= 2(l(\hat{\theta}_n) - l(\tilde{\theta}n)) \\ LM &= n^{-1} s_1(\tilde{\theta}_n)' I^{11}(\tilde{\theta}_n) s_1(\tilde{\theta}_n) \end{aligned}$$

all converge in distribution to $\chi^2_q(\xi' I^{11}(\theta_0)^{-1}\xi)$ where $q = \dim \theta_1 \leq p$.

*Proof:* Recall that

$$I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{12})^{-1}$$

which equals $I_{11}$ if $I_{12} = 0$, otherwise it is strictly *larger*, i.e., the unknown nuisance parameters *increase* the variability of the mle of the full model relative to what it would be if the nuisance parameters were known.

The Wald case is easy

$$\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \rightsquigarrow \mathcal{N}(\xi, I^{11}(\theta_0))$$

so $W \rightsquigarrow \chi^2_1(\delta)$. The LR, LM cases are more tedious and we will defer the proof to the next result which includes this one. Clearly $I^{11}(\theta_0)$ may be replaced by $I^{11}(\hat{\theta}_n)$ or $I^{11}(\tilde{\theta}_n)$ under $H_n$ and can be estimated by either $n^{-1}\nabla l(\nabla l)'$ or $-n^{-1}\nabla^2 l$.

**General Composite Hypothesis II**

Now consider the general nonlinear hypotheses,

$$\begin{array}{ll} H_0: & H(\theta) = 0 \\ H_n: & \theta = \theta_n = \theta_0 + \xi/\sqrt{n} \qquad\qquad \text{where } H(\theta_0) = 0. \end{array}$$

Again, let $\hat{\theta}_n, \tilde{\theta}_n$ denote the unrestricted and restricted mle's. We will denote the Jacobian of the function describing the null hypothesis by $\mathcal{J} = \nabla H(\theta_0)$, and we will assume that $rank(\mathcal{J}) = q \leq p$, finally at the risk of some confusion, we will denote $\mathcal{I} = \mathcal{I}(\theta_0)$.

*Theorem:* (AE/HT - III) Under Lehmann conditions and $H_n$,

$$\begin{aligned} W &= nH(\hat{\theta}_n)'(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}H(\hat{\theta}_n) \\ LR &= 2(l(\hat{\theta}_n) - l(\tilde{\theta}_n)) \\ LM &= n^{-1}\nabla l(\tilde{\theta}_n)'\mathcal{I}^{-1}\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'\mathcal{I}^{-1}\nabla l(\tilde{\theta}_n) \end{aligned}$$

all converge in distribution to $\chi^2_q(\delta)$, where $\delta = \xi\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'\xi$.

*Proof:* Wald, again, is easy. By the $\delta$-method $\sqrt{n}H(\hat{\theta}_n) \rightsquigarrow \mathcal{N}(\mathcal{J}'\xi, \mathcal{J}'\mathcal{I}^{-1}\mathcal{J})$ so we have immediately, $W \rightsquigarrow \chi^2_q(\delta)$. For LR, we may expand, as in the one dimensional case to get

$$LR = n(\hat{\theta}_n - \tilde{\theta}_n)'\mathcal{I}(\hat{\theta}_n - \tilde{\theta}_n) + o_p(1).$$

To connect this expression with $W$ note that $\sqrt{n}H(\hat{\theta}_n) = \sqrt{n}(\hat{\theta} - \tilde{\theta})\mathcal{J} + o_p(1)$ since $\sqrt{n}\hat{H} = \sqrt{n}\tilde{H} + \sqrt{n}(\hat{\theta} - \tilde{\theta})'\mathcal{J} + o_p(1)$, but $\tilde{H} = H(\tilde{\theta}_n) = 0$ by definition, so another way to express $W$ is

$$W^* = n(\hat{\theta} - \tilde{\theta})'\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'(\hat{\theta} - \tilde{\theta})$$

with $W = W^* + o_p(1)$. This formulation is related to the Hausman-Wu testing strategy. To complete the connection between LR and W it is convenient to have a representation of the restricted estimator $\tilde{\theta}_n$ in terms of the score evaluated at the true parameter. To this end, consider the problem,

$$\max_{(\theta,\lambda)} l(\theta) - \lambda' H(\theta),$$

expand the the first order conditions,

$$\nabla \ell(\tilde{\theta}_n) = \tilde{\lambda}_n J$$

$$H(\tilde{\theta}_n) = 0$$

around $\theta_0$ to obtain,

$$\begin{aligned} 0 &= \nabla l(\theta_0) + \nabla^2 l(\theta_0)(\tilde{\theta} - \theta_0) - \mathcal{J}\tilde{\lambda}_n + o_p(1), \\ 0 &= \mathcal{J}(\tilde{\theta}_n - \theta_0) + H(\theta_0) + o_p(1). \end{aligned}$$

and since, under the null $H(\theta_0) = 0$ we have

$$\begin{pmatrix} \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ \frac{1}{\sqrt{n}}\lambda \end{pmatrix} = \begin{pmatrix} \mathcal{I} & \mathcal{J} \\ \mathcal{J} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}}\nabla l(\theta_0) \\ 0 \end{pmatrix} + o_p(1).$$

Using standard formulae for partitioned inverses we have, under $H_0$,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = [\mathcal{I}^{-1} - \mathcal{I}^{-1}\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'\mathcal{I}^{-1}]\frac{1}{\sqrt{n}}\nabla l(\theta_0) + o_p(1)$$

so

$$\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) = [\mathcal{I}^{-1}\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'\mathcal{I}]\frac{1}{\sqrt{n}}\nabla l(\theta_0) + o_p(1)$$

Now,

$$\frac{1}{\sqrt{n}}\nabla l(\theta_0) \rightsquigarrow \mathcal{N}(0, \mathcal{I})$$

so

$$\frac{1}{\sqrt{n}}\mathcal{I}^{-1/2}\nabla l(\theta_0) \rightsquigarrow \mathcal{N}(0, I_p).$$

Set $K \equiv \mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'$ and let $G$ satisfy $GG' = \mathcal{I}^{-1}$ then $G'KG$ is idempotent with rank $q$, by direct computation and,

$$\text{rank } (G'KG) = \text{ Tr } (G'KG) = \text{ Tr } (KGG') = \text{ Tr } [(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}'\mathcal{I}^{-1}\mathcal{J}] = q.$$

Thus,

$$G'KG\frac{1}{\sqrt{n}}\mathcal{I}^{-1/2}\nabla l(\theta_0) \rightsquigarrow \mathcal{N}(0, G'KG)$$

and since

$$\sqrt{n}(\hat{\theta} - \tilde{\theta})\mathcal{I}^{1/2} = G'KG\frac{1}{\sqrt{n}}\mathcal{I}^{-1/2}\nabla l(\theta_0)$$

it follows that LR is asymptotically equivalent to

$$n(\hat{\theta} - \tilde{\theta})'\mathcal{I}(\hat{\theta} - \tilde{\theta}) \rightsquigarrow \chi^2_q.$$

under $H_0$. The details under $H_n$ are left as an exercise.

For the LM case, write

$$n^{-1/2}\nabla l(\tilde{\theta}_n) = n^{-1/2}\nabla l(\hat{\theta}_n) + \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)\nabla^2 l(\tilde{\theta}_n)/n + o_p(1)$$

so

$$\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) = -\mathcal{I}(\theta_0)^{-1}n^{-1/2}\nabla l(\tilde{\theta}_n).$$

so $W^*$ is equivalent to

$$LM^* = n^{-1}\nabla l(\tilde{\theta}_n)'\mathcal{I}^{-1}\mathcal{J}(\mathcal{J}'\mathcal{I}^{-1}\mathcal{J})^{-1}\mathcal{J}\mathcal{I}^{-1}\nabla l(\tilde{\theta}_n)$$

which completes the proof.

*Remark*: Another useful way to connect LM more directly to Lagrange multipliers is to observe that since

$$\tilde{\theta}_n = \arg\min_{\theta}\{l(\theta - \lambda' H(\theta)\}$$

we have at the restricted estimated, $\tilde{\theta}_n$,

$$\nabla l(\tilde{\theta}_n) = \mathcal{J}(\tilde{\theta}_n)\lambda,$$

so we can write,

$$LM^{**} = n^{-1}\lambda'\mathcal{J}'\mathcal{I}^{-1}\mathcal{J}\lambda.$$

where all the matrices are evaluated at $\tilde{\theta}_n$.