Lecture 10

Large Sample Tests.

1 Likelihood Ratio Test

Let $X_1, ..., X_n$ be a random sample from a distribution with pdf $f(x|\theta)$ where θ is some one dimensional (unknown) parameter. Suppose we want to test the null hypothesis, H_0, that $\theta = \theta_0$ against the alternative hypothesis, H_a, that $\theta \neq \theta_0$. Assume the same regularity conditions hold as in the MLE theory. Then likelihood ratio test (LRT) statistic is

$$\lambda(x) = \frac{\mathcal{L}(\theta_0|x)}{\mathcal{L}(\hat{\theta}_{ML}|x)}$$

where $x = (x_1, ..., x_n)$ is a realization of the data set and $\hat{\theta}_{ML}$ is the ML estimator. Then we have

Theorem 1. *Under the same regularity conditions as the MLE theory and if $H_0 : \theta = \theta_0$ holds, we have:*

$$-2 \log \lambda(X) \Rightarrow \chi_1^2.$$

Proof. Denote $\ell(\theta|x) = \log \mathcal{L}(\theta|x)$. By the Taylor theorem, for some $\theta^\star$ between θ_0 and $\hat{\theta}_{ML}$,

$$\begin{aligned}
-2\log\lambda(X) &= -2(\ell(\theta_0|X) - \ell(\hat{\theta}_{ML}|X)) \\
&= -2\left(\frac{\partial \ell(\hat{\theta}_{ML}|X)}{\partial\theta}(\theta_0 - \hat{\theta}_{ML}) + \frac{1}{2}\frac{\partial^2 \ell(\theta^\star|X)}{\partial\theta^2}(\theta_0 - \hat{\theta}_{ML})^2\right) \\
&= -\frac{\partial^2 \ell(\theta^\star|X)}{\partial\theta^2}(\theta_0 - \hat{\theta}_{ML})^2
\end{aligned}$$

since $\partial\ell(\hat{\theta}_{ML}|X)/\partial\theta = 0$ by FOC.

By the MLE theory, $\hat{\theta}_{ML} \to_p \theta_0$. So, $\theta^\star \to_p \theta_0$. As will be shown in 14.385, by the uniform law of large numbers,

$$-\frac{1}{n}\frac{\partial^2 \ell(\theta^\star|X)}{\partial\theta^2} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i|\theta^*)}{\partial\theta^2} \to_p I_1(\theta_0)$$

where $I_1(\theta)$ denotes the information for one observation, i.e. $I_1(\theta) = -E_\theta[\partial^2 \log f(X_i|\theta)/\partial\theta^2]$. By the Slutsky theorem,

$$-\frac{1}{nI_1(\theta_0)}\frac{\partial^2 \ell(\theta^\star|X)}{\partial\theta^2} \to_p 1$$

In addition, from the MLE theory,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I_1^{-1}(\theta_0)).$$

So, by Continuous mapping theorem,

$$I(\theta_0)n(\hat{\theta}_{ML} - \theta_0)^2 \Rightarrow \chi_1^2.$$

By the Slutsky theorem,

$$-2\log\lambda(X) = -\frac{1}{nI(\theta_0)}\frac{\partial^2 \ell(\theta^\star|X)}{\partial\theta^2} I(\theta_0)n(\theta_0 - \hat{\theta}_{ML})^2 \Rightarrow \chi_1^2.$$

□

It follows from this theorem that the large sample LR test of level α rejects the null hypothesis if and only if $-2\log\lambda(x) > \chi_1^2(1-\alpha)$, where $\chi_1^2(1-\alpha)$ denotes $1-\alpha$-quantile of χ_1^2. Note that in finite samples, the size of this test may be greater than α but as the sample size increases, the size will converge to α.

1.1 Formulation for multi-dimensional case

In general, let θ be a multidimensional parameter (say dimensionality is k). Suppose that the null hypothesis Θ_0 can be written in the form $\{\theta \in \Theta : g_1(\theta) = 0, ..., g_p(\theta) = 0\}$ where $g_1, ..., g_p$ denote some nonlinear functions of θ. Equations $g_1(\theta) = 0$, ..., $g_p(\theta) = 0$ are called restrictions of the model (and $k \geq p$, if $k > p$ then it is a composite hypothesis, if $k = p$ then simple). Assume that restrictions are jointly linear independent in the sense that we cannot drop any subset of restrictions without changing set Θ_0. Then, under some regularity conditions (mainly smoothness of $g_1, ..., g_p$),

$$-2\log\lambda(X) = 2\left(\max_{\theta\in\Theta} \ell(\theta|X) - \max_{\theta\in\Theta_0} \ell(\theta|X)\right) \Rightarrow \chi_p^2$$

under the assumption that the null hypothesis hold. So, large sample LR test of level α rejects the null hypothesis if and only if $-2\log\lambda(X) > \chi_p^2(1-\alpha)$. Often, we denote $LR = -2\log\lambda(X)$. LR is called the likelihood ratio statistic. Let us denote $\hat{\theta}_0 = \arg\max_{\theta\in\Theta_0} \ell(\theta|X)$ to be the restricted estimate (estimates assuming the null is true), then

$$LR = 2(\ell(\hat{\theta}_{ML}|X) - \ell(\hat{\theta}_0|X))$$

Example Let $X_1, ..., X_n$ be a random sample from a Poisson(λ) distribution. Recall that the pmf of the Poisson(λ) distribution is $f(x|\lambda) = \lambda^x e^{-\lambda}/x!$ for $x = 0, 1, 2, ...$ Suppose we want to test the null hypothesis, H_0, that $\lambda = \lambda_0 = 6$ against the alternative hypothesis, H_a, that $\lambda \neq \lambda_0$. Suppose we observe $\overline{X}_n = 5$ while our sample size $n = 100$. Let us derive the result of the large sample LR test. Likelihood function is

$$\mathcal{L}(\lambda|X) = \frac{\lambda^{\sum_{i=1}^n X_i} e^{-n\lambda}}{\prod_{i=1}^n X_i!}$$

where $X = (X_1, ..., X_n)$. The log-likelihood is

$$\ell(\lambda|X) = \sum_{i=1}^{n} X_i \log \lambda - n\lambda - \log \prod_{i=1}^{n} X_i!$$

So, the ML estimator $\hat{\lambda}_{ML}$ solves

$$\sum_{i=1}^{n} X_i/\hat{\lambda}_{ML} - n = 0$$

or, equivalently,

$$\hat{\lambda}_{ML} = \overline{X}_n$$

So, LRT statistic is

$$\lambda(x) = (\lambda_0/\hat{\lambda}_{ML})^{\sum_{i=1}^{n} X_I} e^{-n(\lambda_0 - \hat{\lambda}_{ML})}.$$

Then

$$\begin{aligned}
LR &= -2 \log \lambda(x) \\
&= -2 \left(\sum_{i=1}^{n} X_i \log(\lambda_0/\hat{\lambda}_{ML}) - n(\lambda_0 - \hat{\lambda}_{ML}) \right) \\
&= -2n(\overline{X}_n \log(\lambda_0/\overline{X}_n) - \lambda_0 + \overline{X}_n) \\
&= -200(5 \log(6/5) - 6 + 5) \\
&\approx 17.6,
\end{aligned}$$

while $\chi_1^2(0.95) = 3.98$. So large sample LR test rejects the null hypothesis.

2 Large Sample Tests: Wald

2.1 Simplistic 1-dimensional case

Once we know the asymptotic distribution of some statistic, say, $\delta(X_1, ..., X_n)$, we can construct a large sample test based on this asymptotic distribution. Suppose we can show that

$$\sqrt{n}(\delta(X_1, ..., X_n) - \tau) \Rightarrow N(0, \sigma^2)$$

where τ is some 1-dimensional parameter. Suppose we have a consistent estimator $\hat{\sigma}^2$ of σ^2, i.e. $\hat{\sigma}^2 \rightarrow_p \sigma^2$. By the Slutsky theorem,

$$\sqrt{n}(\delta(X_1, ..., X_n) - \tau)/\hat{\sigma} \Rightarrow N(0, 1)$$

Suppose we want to test the null hypothesis, H_0, that $\tau = \tau_0$ against the two-sided alternative. Under the null hypothesis,

$$\sqrt{n}(\delta(X_1, ..., X_n) - \tau_0)/\hat{\sigma} \Rightarrow N(0, 1).$$

So, one test of level α will be to reject the null hypothesis if $t = \sqrt{n}\frac{\delta(X_1,\dots,X_n)-\tau_0}{\hat{\sigma}}$ is smaller than $z_{\alpha/2}$ or larger than $z_{1-\alpha/2}$. Which is equivalent to calculating statistic

$$W = n\left(\frac{\delta(X)-\tau_0}{\hat{\sigma}}\right)^2$$

and comparing it to $1-\alpha$ quantile of χ_1^2 distribution.

2.2 Multi-dimensional case

Notice that this logic could be easily extended to multi-dimensional parameters. Assume that τ is p-dimensional and

$$\sqrt{n}(\delta(X)-\tau) \Rightarrow N(0,\Sigma),$$

and we can construct a consistent estimate $\hat{\Sigma}$ of the covariance matrix Σ, that is $\hat{\Sigma} \to^p \Sigma$, then

$$W = n(\delta(X)-\tau_0)'\hat{\Sigma}^{-1}(\delta(X)-\tau_0) \Rightarrow \chi_p^2$$

if $H_0 : \tau = \tau_0$.

2.3 Special case: 1-dimensional MLE

We can specialize this to the MLE case, for example, and see how this test compares to the LR introduced before. Let $\hat{\theta}_{ML}$ be the ML estimator of 1-dimensional parameter $\theta \in \mathbb{R}$. We know that, under some regularity conditions,

$$\sqrt{n}(\hat{\theta}_{ML}-\theta) \Rightarrow N(0, I_1^{-1}(\theta))$$

Under some regularity conditions, $I_1^{-1}(\theta)$ may be consistently estimated by $I_1^{-1}(\hat{\theta}_{ML})$. Suppose that our null hypothesis is $H_0 : \theta = \theta_0$. Then, under the null hypothesis,

$$\sqrt{n}I^{1/2}(\hat{\theta}_{ML})(\hat{\theta}_{ML}-\theta_0) \Rightarrow N(0,1)$$

Under the null hypothesis

$$W = nI_1(\hat{\theta}_{ML})(\hat{\theta}_{ML}-\theta_0)^2 \Rightarrow \chi_1^2.$$

Recall that LR-statistic is given by

$$LR = n(\hat{\theta}_{ML}-\theta_0)^2\left(-\frac{1}{n}\frac{\partial^2\ell(\theta^\star|X)}{\partial\theta^2}\right)$$

where $\theta^\star$ is between θ_0 and $\hat{\theta}_{ML}$. As in the case of Wald statistic, under the null hypothesis,

$$LR \Rightarrow \chi_1^2$$

Moreover,

$$W - LR \to_p 0$$

since $I(\hat{\theta}_{ML}) \to_p I(\theta_0)$ and $-(1/n)\partial^2 \ell(\theta^\star|X)/\partial\theta^2 \to_p I(\theta_0)$. Thus, LR and Wald statistics are asymptotically equivalent. They are different in finite samples though. In particular, it is known that $W \geq LR$ in the case of normal likelihood.

An advantage of the Wald statistic in comparison with the LR statistic is that it only includes calculations based on the unrestricted estimator $\hat{\theta}_{ML}$. On the other hand, in order to calculate the Wald statistic, we have to estimate the information matrix.

Example (cont.) Let us calculate the Wald statistic in our example with a random sample from the Poisson(λ) distribution. The log-likelihood is

$$\ell(\lambda|X) = \sum_{i=1}^{n} X_i \log\lambda - n\lambda - \log\prod_{i=1}^{n} X_i!$$

So,

$$\partial\ell(\lambda|X)/\partial\lambda = \sum_{i=1}^{n} X_i/\lambda - n$$

and

$$\partial^2\ell(\lambda|X)/\partial\lambda^2 = -\sum_{i=1}^{n} X_i/\lambda^2.$$

Thus,

$$I_1(\lambda) = -E\left[\frac{1}{n}\frac{\partial^2\ell(\lambda)}{\partial\lambda^2}\right] = \frac{1}{\lambda}.$$

So, the Wald statistic is

$$W = n(\hat{\lambda} - \lambda_0)^2/\hat{\lambda} = 100 \cdot (5-6)^2 \cdot (1/5) = 20.$$

So, the test based on the Wald statistic rejects the null hypothesis with an even smaller p-value than the test based on the LR statistic.

3 Score Test

3.1 1-dimensional case

Recall that the score is defined by

$$S(\theta) = \frac{\partial\ell}{\partial\theta}(\theta|X) = \frac{\partial\log\mathcal{L}}{\partial\theta}(\theta|X) = \sum_{i=1}^{n}\frac{\partial\log f(X_i|\theta)}{\partial\theta}.$$

By the first order condition for the ML estimator, $S(\hat{\theta}_{ML}) = 0$. By the first information equality,

$$E[S(\theta_0)] = \sum_{i=1}^{n} E\left[\frac{\partial \log f(X_i|\theta_0)}{\partial \theta}\right] = 0.$$

By definition of Fisher information,

$$E\left[\left(\frac{\partial \log f(X_i|\theta)}{\partial \theta}\right)^2\right] = I_1(\theta_0).$$

So, by the Central limit theorem, under the null hypothesis (if θ is 1-dimensional)

$$\frac{1}{\sqrt{n}} \frac{S(\theta_0)}{\sqrt{I_1(\theta_0)}} \Rightarrow N(0,1).$$

By the continuous mapping theorem,

$$LM = S(\theta_0)^2/(nI_1(\theta_0)) \Rightarrow \chi_1^2.$$

The LM is called Lagrange Multiplier (LM) statistic. Let us show where the name comes from. Consider the constrained optimization problem $\log \mathcal{L}(\theta|x) \to \max$ s.t. $\theta = \theta_0$. The lagrangian is

$$H = \log \mathcal{L}(\theta|x) - \lambda(\theta - \theta_0).$$

The FOC is

$$S(\theta_0) = \lambda.$$

So, indeed, the score is connected to the lagrange multiplier.

Comparison to LR and Wald Let us show that $LM - LR \to_p 0$. By the Taylor's expansion,

$$S(\theta_0) = S(\theta_0) - S(\hat{\theta}_{ML}) = \frac{\partial^2 \ell}{\partial \theta^2}(\theta^\star|X)(\theta_0 - \hat{\theta}_{ML}),$$

where $\theta^\star$ is between θ_0 and $\hat{\theta}_{ML}$. As before, $-(1/n)\partial^2\ell(\theta^\star|X)/\partial\theta^2 \to_p I_1(\theta_0)$. By the Slutsky theorem,

$$LM = n^2 \left(\frac{1}{n}\frac{\partial^2 \ell}{\partial \theta^2}(\theta^\star|X)\right)^2 (\theta_0 - \hat{\theta}_{ML})^2/(nI_1(\theta_0)) = nI_1(\theta_0)(\theta_0 - \hat{\theta}_{ML})^2(1 + o_p(1))$$

Thus, we have shown that LR, Wald, and LM statistics are all asymptotically equivalent under the null hypothesis. However, they differ in finite samples. For example, in the case of normal likelihood, we have $LM \leq LR \leq W$.

3.2 Multi-dimensional case

Assume that the unknown parameter θ is k-dimensional, while the null hypothesis is imposing p-dimensional restriction $\Theta_0 = \{\theta \in \Theta : g_1(\theta) = 0, ..., g_p(\theta) = 0\}$. Score function is $k \times 1$- vector function $S(\theta) = \frac{\partial \ell}{\partial \theta}(\theta|X)$. Denote $\hat{\theta}_0$ to be restricted estimator: $\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \ell(\theta|X)$. Then

$$LM = \frac{1}{n} S(\hat{\theta}_0) I_1(\hat{\theta}_0)^{-1} S(\hat{\theta}_0) \Rightarrow \chi^2_p$$

if the null holds.

An advantage of the LM statistic is that it only includes calculations based on the restricted estimator θ_0. On the other hand, in order to find the LM statistic, we have to estimate Fisher information.

Example (cont.) Let us calculate the LM statistic in our example with a random sample from Poisson(λ) distribution. We have

$$S(\lambda_0) = \sum_{i=1}^n X_i/\lambda_0 - n = 500/6 - 100 = -100/6$$

and $I(\lambda_0) = 1/\lambda_0 = 1/6$. So,

$$LM = \frac{S(\lambda_0)^2}{nI(\lambda_0)} = \frac{1}{100} \cdot \left(\frac{100}{6}\right)^2 \cdot 6 = \frac{100}{6} \approx 17$$

4 Generalizations and Summary

Let $x = (X_1, ..., X_n)$ be a random sample from distribution $f(X|\theta)$ with $\theta \in \Theta$. Suppose we want to test the null hypothesis, H_0, that $\theta \in \Theta_0$ against the alternative hypothesis, H_a, that $\theta \notin \Theta_0$. Let $\hat{\theta}_0$ be a restricted estimator, i.e. $\hat{\theta}_0$ solves $\max_{\theta \in \Theta_0} \mathcal{L}(\theta|x)$, and $\hat{\theta}_{ML}$ an unrestricted (ML) estimator, i.e. $\hat{\theta}_{ML}$ solves $\max_{\theta \in \Theta} \mathcal{L}(\theta|x)$. Assume for simplicity that the null can be formulated as $g(\theta) = 0$, where g is p-dimensional function. Then, under the null hypothesis,

$$LR = 2(\ell(\hat{\theta}_{ML}|X) - \ell(\hat{\theta}_0|X)) \Rightarrow \chi^2_p$$

$$W = (g(\hat{\theta}_{ML}) - 0)\hat{\Sigma}^{-1}(g(\hat{\theta}_{ML}) - 0) \Rightarrow \chi^2_p$$

$$LM = S(\hat{\theta}_0) I_n^{-1}(\hat{\theta}_0) S(\hat{\theta}_0) \Rightarrow \chi^2_p$$

where $\hat{\Sigma} = \left(\frac{\partial g}{\partial \theta}(\hat{\theta}_{ML})\right)' I_1^{-1}(\hat{\theta}_{ML}) \left(\frac{\partial g}{\partial \theta}(\hat{\theta}_{ML})\right)$ is a natural delta-method inspired estimate of asymptotic variance. Under a proper regularity conditions all of these tests are asymptotically equivalent to each other.

Notice, that LR and LM are invariant to formulation of the null hypothesis, while Wald is not.