LECTURE 12 NOTES

1. The likelihood ratio test. We begin by studying the asymptotic distribution of the log-LR statistic

$$\log \lambda_n(\mathbf{x}) = \sup_{\theta\in\Theta} \textstyle\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta) - \sup_{\theta\in\Theta_0} \sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta).$$

Once we know the asymptotic distribution of $2\log\lambda_n(\mathbf{x})$, setting the critical value t equal to its $1-\alpha$ quantile ensures

$$\mathbf{P}_0(2\log\lambda_n(\mathbf{x}) > t) \to 1 - F_0(t) \le \alpha,$$

where $F_0(t)$ is the CDF of the asymptotic distribution of $2\log\lambda_n(\mathbf{x})$. Equivalently,

$$p(\mathbf{x}) = 1 - F_0(2\log\lambda_n(\mathbf{x}))$$

is an asymptotically valid p-value: its asymptotic distribution is $\mathrm{unif}(0,1)$.

THEOREM 1.1. *Consider testing $H_0 : \theta \in \Theta_0$, where Θ_0 is a q-dimensional subspace of $\mathbf{R}^p$. Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} f_{\theta^*}(x)$ for some $\theta^* \in \mathrm{int}(\Theta_0)$. Assume*

1. *ℓ_x is twice-continuously differentiable for any $x \in \mathcal{X}$,*
2. *$\frac{1}{\sqrt{n}} \sum_{i\in[n]} \nabla\ell_{\mathbf{x}_i}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, I(\theta^*)\big)$, where $I(\theta^*) = -\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]$.*

If the (unrestricted) MLE and restricted MLE are consistent,

$$2\log\lambda_n(\mathbf{x}) \xrightarrow{d} \chi^2_{p-q}.$$

PROOF. To keep notation manageable, let

1. $\ell_n(\theta) = \sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta)$ be the log-likelihood of $\{\mathbf{x}_i\}$,
2. $I_n(\theta) = -\nabla^2\ell_n(\theta^*)$.

The proof consists of three steps:

1. show that the restricted MLE $\{\tilde{\theta}_n\}$ is asymptotically normal:

$$\sqrt{n}\big(\tilde{\theta}_n - \theta^*\big) = \big(\tfrac{1}{n} P\nabla^2 I_n(\theta^*) P\big)^\dagger \tfrac{1}{\sqrt{n}} \nabla\ell_n(\theta^*) + o_P(1), \tag{1.1}$$

   where P is the orthogonal projector onto $\mathrm{span}(\Theta_0)$.
2. show that $2\log\lambda(\mathbf{x})$ is essentially a quadratic function of the score.
3. show that the quadratic function of the score is asymptotically χ^2.

We first show (1.1). Let $I_n(\theta^*) = -\nabla^2 \ell_n(\theta^*)$ be the observed Fisher information. By the optimality of $\tilde{\theta}_n$,

$$\begin{aligned} 0 &= P\nabla \ell_n(\tilde{\theta}_n) \\ &= P\big(\nabla \ell_n(\theta^*) - I_n(\theta^*)\big(\tilde{\theta}_n - \theta^*\big) + r_n\big), \end{aligned}$$

which, since $\tilde{\theta}_n - \theta^* \in \operatorname{span}(\Theta_0)$,

$$= P\big(\nabla \ell_n(\theta^*) - I_n(\theta^*) P\big(\tilde{\theta}_n - \theta^*\big) + r_n\big).$$

Rearranging,

$$\sqrt{n}\big(\tilde{\theta}_n - \theta^*\big) = \big(\tfrac{1}{n} P I_n(\theta^*) P\big)^\dagger \tfrac{1}{\sqrt{n}}\big(\nabla \ell_n(\theta^*) + r_n\big),$$

It is possible to show the remainder term is asymptotically negligible by appealing to the consistency of $\tilde{\theta}_n$. Thus the restricted MLE is asymptotically normal:

$$\sqrt{n}\big(\tilde{\theta}_n - \theta^*\big) = \big(\tfrac{1}{n} P I_n(\theta^*) P\big)^\dagger \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1).$$

Combining the asymptotic normality of the (unrestricted) MLE and restricted MLE, we show that $2 \log \lambda(\mathbf{x})$ is approximately quadratic. By a second-order Taylor expansion of ℓ_n at θ^*,

$$\begin{aligned} &\ell_n(\hat{\theta}_n) - \ell_n(\theta^*) \\ &\quad = \nabla \ell_n(\theta^*)^T \big(\hat{\theta}_n - \theta^*\big) - \tfrac{1}{2}\big(\hat{\theta}_n - \theta^*\big)^T I_n(\theta^*)\big(\hat{\theta}_n - \theta^*\big) + r_n, \end{aligned} \tag{1.2}$$

for any $\theta \in \Theta$. Once again, it is possible to show that the remainder term is asymptotically negligible by appealing to the consistency of $\hat{\theta}_n$. The first term in (1.2) is

$$\begin{aligned} &\nabla \ell_n(\theta^*)^T \big(\hat{\theta}_n - \theta^*\big) \\ &\quad = \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*)^T \sqrt{n}\big(\hat{\theta}_n - \theta^*\big), \end{aligned}$$

which, by the asymptotic normality of the MLE,

$$\begin{aligned} &= \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*)^T \big(\tfrac{1}{n} I_n(\theta^*)\big)^{-1} \big(\tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1)\big) \\ &= \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*)^T \big(\tfrac{1}{n} I_n(\theta^*)\big)^{-1} \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1). \end{aligned}$$

Similarly, the second term is

$$\begin{aligned} &-\tfrac{1}{2}\big(\hat{\theta}_n - \theta^*\big)^T I_n(\theta^*)\big(\hat{\theta}_n - \theta^*\big) \\ &\quad = -\tfrac{1}{2}\sqrt{n}\big(\hat{\theta}_n - \theta^*\big)^T \big(\tfrac{1}{n} I_n(\theta^*)\big)\sqrt{n}\big(\hat{\theta}_n - \theta^*\big) \\ &\quad = -\tfrac{1}{2}\big(\tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1)\big)^T \big(\tfrac{1}{n} I_n(\theta^*)\big)^{-1} \big(\tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1)\big) \\ &\quad = -\tfrac{1}{2} \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*)^T \big(\tfrac{1}{n} I_n(\theta^*)\big)^{-1} \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta^*) + o_P(1). \end{aligned}$$

We plug the expressions into (1.2) to obtain

$$\ell_n(\hat{\theta}_n) - \ell_n(\theta^*) = \tfrac{1}{2}\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*)^T\big(\tfrac{1}{n}I_n(\theta^*)\big)^{-1}\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*) + o_P(1). \tag{1.3}$$

There is a similar expression for $\ell_n(\tilde{\theta}_n) - \ell_n(\theta^*)$, and we skip to the conclusion:

$$\begin{aligned}&\ell_n(\tilde{\theta}_n) - \ell_n(\theta^*)\\ &\quad= \tfrac{1}{2}\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*)^T\big(\tfrac{1}{n}PI_n(\theta^*)P\big)^\dagger\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*) + o_P(1).\end{aligned} \tag{1.4}$$

Finally, we subtract (1.4) from (1.3) to deduce $2\log\lambda_n(\mathbf{x})$ is approximately quadratic:

$$\begin{aligned}2\log\lambda_n(\mathbf{x}) &= 2(\ell_n(\hat{\theta}_n) - \ell_n(\tilde{\theta}_n))\\ &= \tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*)^T\big(\big(\tfrac{1}{n}I_n(\theta^*)\big)^{-1} - \big(\tfrac{1}{n}PI_n(\theta^*)P\big)^\dagger\big)\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta^*) + o_P(1).\end{aligned}$$

Under the conditions of the Theorem, $\frac{1}{\sqrt{n}}\nabla\ell_n(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, I(\theta^*)\big)$. By Slutsky's theorem,

$$\begin{aligned}2\log\lambda_n(\mathbf{x}) &\xrightarrow{d} \big(I(\theta^*)^{\frac{1}{2}}\mathbf{z}\big)^T\big(I(\theta^*)^{-1} - (PI(\theta^*)P)^\dagger\big)\big(I(\theta^*)^{\frac{1}{2}}\mathbf{z}\big)\\ &= \mathbf{z}^T\big(I_p - \underbrace{I(\theta^*)^{\frac{1}{2}}(PI(\theta^*)P)^\dagger I(\theta^*)^{\frac{1}{2}}}_{J}\big)\mathbf{z},\end{aligned}$$

where $\mathbf{z} \sim \mathcal{N}(0, I_p)$.

To complete the proof, we check that $I_p - J$ is a projector onto a q-dimensional subspace of $\mathbf{R}^p$. Indeed, J is symmetric, and, since $\mathcal{R}(PI(\theta^*)P)$ is $\mathrm{span}(\Theta_0)$,

1. J is equivalently $I(\theta^*)^{\frac{1}{2}}P(PI(\theta^*)P)^\dagger PI(\theta^*)^{\frac{1}{2}}$, which begets $J^2 = J$.
2. $$\begin{aligned}\mathrm{tr}(J) &= \mathrm{tr}\big(I(\theta^*)^{\frac{1}{2}}P(PI(\theta^*)P)^\dagger PI(\theta^*)^{\frac{1}{2}}\big)\\ &= \mathrm{tr}\big((PI(\theta^*)P)(PI(\theta^*)P)^\dagger\big)\\ &= \dim(\mathrm{span}(\Theta_0)),\end{aligned}$$
   which implies $\mathrm{tr}(I_p - J) = p - q$.

□

Example 1.2. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{bin}(n, p)$. *Consider testing* $H_0 : p = \frac{1}{2}$ *versus* $H_1 : p \neq \frac{1}{2}$. *We showed that the likelihood ratio is*

$$\lambda_n(\mathbf{t}) = \big(\tfrac{1}{2}\big)^{-n}\textstyle\prod_{i\in[n]}\big(\frac{\mathbf{t}}{n}\big)^{\mathbf{t}}\big(1 - \big(\frac{\mathbf{t}}{n}\big)\big)^{n-\mathbf{t}},$$

By Theorem 1.1, $2\log\lambda_n(\mathbf{t}) \xrightarrow{d} \chi^2_{1-0=1}$ *under* H_0, *and the test rejects* H_0 *if* $2\log\lambda_n(\mathbf{x})$ *exceeds* $\chi^2_{1,\alpha}$.

2. The Wald and Rao tests. To evaluate the LR, the investigator must evaluate both the (unrestricted) MLE and the restricted MLE. Sometimes, one is easier to obtain than the other. The Wald and score test statistics only depend on one of the MLE's and are appropriate in such cases.

2.1. *The Wald test.*

THEOREM 2.1. *Under the conditions of Theorem 1.1, the Wald statistic is*

$$\hat{\theta}_n^T\big((I_p - P)I_n\big(\hat{\theta}_n\big)(I_p - P)\big)\hat{\theta}_n, \tag{2.1}$$

where $\hat{\theta}_n$ is the MLE on Θ and P is the projector onto $\operatorname{span}(\Theta_0)$*, is asymptotically χ^2_{p-q}.*

PROOF. The Wald statistic is equivalently

$$\begin{aligned}&\big(\hat{\theta}_n - \theta_0\big)^T\big((I_p - P)I_n\big(\hat{\theta}_n\big)(I_p - P)\big)\big(\hat{\theta}_n - \theta_0\big)\\ &\quad= \sqrt{n}\big(\hat{\theta}_n - \theta_0\big)^T\big(\tfrac{1}{n}(I_p - P)I_n\big(\hat{\theta}_n\big)(I_p - P)\big)\sqrt{n}\big(\hat{\theta}_n - \theta_0\big),\end{aligned}$$

which, by the asymptotic normality of $\hat{\theta}_n$, the LLN, and Slutsky's theorem,

$$\xrightarrow{d} \mathbf{z}^T I(\theta^*)^{-\frac{1}{2}}\big((I_p - P)I(\theta^*)(I_p - P)\big)I(\theta^*)^{-\frac{1}{2}}\mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(0, I_p)$. It is possible to adapt the third part of the proof of Theorem 1.1 to show that

$$I(\theta^*)^{-\frac{1}{2}}\big((I_p - P)I(\theta^*)(I_p - P)\big)I(\theta^*)^{-\frac{1}{2}}$$

is a projector onto a $p-q$-dimensional subspace of $\mathbf{R}^p$, which implies the Wald statistic has the stated asymptotic distribution. □

If $\operatorname{span}(\Theta_0)$ is the kernel of some $R \in \mathbf{R}^{r\times p}$ (i.e. $H_0 : R\theta = 0$) that has full row rank,

$$I_p - P = R^T\big(R^T\big)^\dagger = R^T\big(RR^T\big)^{-1}R,$$

and the Wald test statistic is

$$\begin{aligned}&\big(R\hat{\theta}_n\big)^T\big(\big(R^T\big)^\dagger I_n\big(\hat{\theta}_n\big)R^\dagger\big)^{-1}\big(R\hat{\theta}_n\big)\\ &\quad= \big(R\hat{\theta}_n\big)^T\big(R^T I_n\big(\hat{\theta}_n\big)^{-1}R\big)^{-1}\big(R\hat{\theta}_n\big)\\ &\quad= \big\|R\hat{\theta}_n\big\|^2_{\left(R^T I_n\left(\hat{\theta}_n\right)^{-1}R\right)^{-1}}.\end{aligned}$$

We see that the Wald test statistic is a measure of the size of the infeasibility of $\hat{\theta}_n$. Under H_1, the infeasibility is likely large. Thus we reject H_0 if the Wald test statistic is large.

2.2. *The Rao test.* The Rao test is the counterpart to the Wald test: it only depends on the restricted MLE.

THEOREM 2.2. *Under the conditions of Theorem 1.1, the Rao statistic is*

$$\nabla \ell_n(\tilde{\theta}_n)^T I_n(\tilde{\theta}_n)^{-1} \nabla \ell_n(\tilde{\theta}_n),$$

where $\tilde{\theta}_n$ is the MLE on Θ_0, is asymptotically χ^2_{p-q}.

PROOF. By the optimality of $\tilde{\theta}_n$, $\nabla \ell_n(\tilde{\theta}_n) \in \operatorname{span}(\Theta_0)^\perp$. Thus the Rao statistic is equivalently

$$\tfrac{1}{\sqrt{n}} \nabla \ell_n(\tilde{\theta}_n)^T (I_p - P) \big(\tfrac{1}{n} I_n(\tilde{\theta}_n)\big)^{-1} (I_p - P) \tfrac{1}{\sqrt{n}} \nabla \ell_n(\tilde{\theta}_n),$$

which, by the CLT, the LLN, and Slutsky's theorem,

$$\xrightarrow{d} \mathbf{z}^T I(\theta^*)^{\frac{1}{2}} (I_p - P) I(\theta^*)^{-1} (I_p - P) I(\theta^*)^{\frac{1}{2}} \mathbf{z}.$$

where $\mathbf{z} \sim \mathcal{N}(0, I_p)$. It is possible to adapt the third part of the proof of Theorem 1.1 to show that

$$I(\theta^*)^{\frac{1}{2}} (I_p - P) I(\theta^*)^{-1} (I_p - P) I(\theta^*)^{\frac{1}{2}}$$

is a projector onto a $p - q$-dimensional subspace of $\mathbf{R}^p$, which implies the Rao statistic has the stated asymptotic distribution. □

Under the alternative, the restricted MLE $\tilde{\theta}_n$ is far from the true parameter. Thus $\nabla \ell_n(\tilde{\theta}_n)$ is likely large, and we should reject the null when the Rao statistic is large.

The geometric intuition behind Theorems 1.1, 2.1, 2.2 is worth mentioning. It is most obvious in the form of the score test statistic. By the optimality of $\tilde{\theta}_n$, $\nabla \ell_n(\tilde{\theta}_n)$ falls in the orthocomplement of $\operatorname{span}(\Theta_0)$. Thus it is supported on a $p - q$-dimensional subspace of $\mathbf{R}^p$. By the CLT, $\nabla \ell_n(\tilde{\theta}_n)$ is asymptotically normal and quadratic forms of normals are χ^2, which is the form of the score test statistic.

We remark that the conclusions of Theorems 1.1, 2.1, 2.2 are valid under more general conditions. A practical generalization allows Θ_0 to be a submanifold of $\mathbf{R}^p$, which includes sets of the form

$$\Theta_0 = \{\theta \in \Theta : h(\theta) = 0\},$$

where $h : \Theta \to \mathbf{R}^r$ is continuously-differentiable and $\nabla h(\theta)$ is surjective for any $\theta \in \text{int}(\Theta)$. The LR and Rao statistics for testing $H_0 : h(\theta) = 0$ are unchanged; the Wald statistic is

$$h(\hat{\theta}_n)^T \big(\nabla h(\hat{\theta}_n)^T I_n\big(\hat{\theta}_n\big)^{-1} \nabla h(\hat{\theta}_n)\big)^{-1} h(\hat{\theta}_n).$$

Under conditions that ensure the MLE is asymptotically normal, the three statistics are asymptotically χ^2_r under H_0.

EXAMPLE 2.3. *Let* $\mathbf{x} \overset{\text{i.i.d.}}{\sim} \text{Cat}(p)$*, where* p_1*,* p_2*,* p_3 *are positive and sum to one. We wish to test*

$$H_0 : 2\sqrt{p_1}\sqrt{p_2} - p_3 = 0.$$

If p_1*,* p_2*,* p_3 *are the frequencies of homozygous-dominant, homozygous-recessive, and heterozygous genotypes in a population,* H_0 *asserts the frequencies are in Hardy-Weinberg equilibrium.*

Since $\Theta_0 := \big\{p \in [\,0,1\,]^3 : 2\sqrt{p_1}\sqrt{p_2} - p_3 = 0\big\}$ *is defined by a non-linear constraint, it is hard to optimize on* Θ_0*. However, the (unrestricted) MLE of* p_1*,* p_2*,* p_3 *are easy to obtain:* $\hat{p}_j = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}\{\mathbf{x}_i = j\}$*. We are led to consider the Wald test.*

Since the parameter space

$$\big\{(p_1, p_2, p_3) \in [\,0,1\,]^3 : p_1 + p_2 + p_3 = 1\big\}$$

is an affine subset of $\mathbf{R}^3$*, we cannot directly appeal to Theorem 2.1 to deduce the asymptotic distribution of the Wald test statistic. However, if we reparametrize the model in terms of* p_1 *and* p_2*, the (reduced) parameter space is a full-dimensional set in* $\mathbf{R}^2$*, and Theorem 2.1 is applicable.*

After re-parametrization, the hypothesis is equivalently

$$H_0 : 2\sqrt{p_1}\sqrt{p_2} - (1 - p_1 - p_2) = 0$$

Let $h(p) = 2\sqrt{p_1}\sqrt{p_2} - (1 - p_1 - p_2)$*. It is possible to show the Wald test statistic is*

$$\mathbf{w}_n = h(\hat{p}_n)^T \big(\nabla h(\hat{p}_n)^T I_n\big(\hat{p}_n\big)^{-1} \nabla h(\hat{p}_n)\big)^{-1} h(\hat{p}_n),$$

where $\nabla h(p)^T = \begin{bmatrix} \frac{\sqrt{p_2}}{\sqrt{p_1}} + 1 & \frac{\sqrt{p_1}}{\sqrt{p_2}} + 1 \end{bmatrix}$ *and* $I_n(p) = n\big(\mathbf{diag}(p) - pp^T\big)^{-1}$*.*

Since Θ_0 *is a 1-dimensional curve in* $\mathbf{R}^2$*,* $\mathbf{w}_n \overset{d}{\to} \chi^2_{1=2-1}$ *by Theorem 2.1. Thus we reject* H_0 *if* $\mathbf{w}_n$ *exceeds* $\chi^2_{1,\alpha}$*.*

3. GMM analogues. To wrap up, we describe the GMM analogues of the LR, Wald, and Rao tests for testing $H_0 : h(\theta) = 0$.

1. The analogue of the LR statistic is the *distance metric statistic*

$$2(Q_n(\tilde{\theta}_n) - Q_n(\hat{\theta}_n)),$$

   where $\tilde{\theta}_n$ and $\hat{\theta}_n$ are the restricted and unrestricted GMM estimators.
2. The analogue of the Wald test statistic is

$$h(\hat{\theta}_n)^T \left(\nabla h(\hat{\theta}_n)^T \left(n\widehat{J}_n\right)^{-1} \nabla h(\hat{\theta}_n)\right)^{-1} h(\hat{\theta}_n),$$

   where $\widehat{J}_n$ is a consistent estimator of the asymptotic variance of

$$\sqrt{n}\nabla Q_n(\theta).$$

3. The analogue of the Rao test statistic is

$$n\nabla Q_n(\tilde{\theta}_n)^T \widehat{J}_n^{-1} \nabla Q_n(\tilde{\theta}_n),$$

   where $\widehat{J}_n$ is some consistent estimator of the asymptotic variance of $\sqrt{n}\nabla Q_n(\theta_0)$. In econometrics, where GMM estimators are more popular than the MLE, the Rao test is usually referred to as the *score test* or *Lagrange multiplier test*.

Under conditions that ensure the restrcted and unrestricted GMM estimators are asymptotically normal, the three test statistics are asymptotically χ^2_r.

Yuekai Sun
Berkeley, California
December 4, 2015