ML Testing (Likelihood Ratio Testing) for non-Gaussian models

Surya Tokdar

ML test in a slightly different form

- Model $X \sim f(x|\theta)$, $\theta \in \Theta$. Hypothesist $H_0 : \theta \in \Theta_0$
- Good set: $B_c(x) = \{\theta : \ell_x(\theta) \geq \max_{\theta \in \Theta} \ell_x(\theta) - \frac{c^2}{2}\}$
- ML test = reject H_0 if $B_c(x) \cap \Theta_0 = \emptyset$
- Same as: reject H_0 if

$$\max_{\theta \in \Theta} \ell_x(\theta) - \max_{\theta \in \Theta_0} \ell_x(\theta) \geq c^2/2$$

- Or equivalently: reject H_0 if

$$\Lambda(x) = \frac{\max_{\theta \in \Theta} L_x(\theta)}{\max_{\theta \in \Theta_0} L_x(\theta)} \geq k$$

Size calculation

- Need to calculate $\max_{\theta \in \Theta_0} P_{[X|\theta]}(\Lambda(X) > k)$
- Will work for models of the form: $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$
- And hypothesis: $H_0 : A^T\theta = a_0$ for a given $p \times r$ matrix A and given scalar a_0 (typically $a_0 = 0$)
- Under regularity conditions: $2 \log \Lambda(X) \overset{d}{\to} \chi^2(r)$ when $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ with $\theta_0 \in \Theta_0$.
- Follows from asymptotic normality of $\hat{\theta}_{\text{MLE}}(X)$

Asymptotic normality of MLE

Theorem. Assume the following (Cramér) conditions

1. $\theta \mapsto \frac{\partial}{\partial \theta} \log g(x|\theta)$ is twice continuously differentiable.
2. $\theta \mapsto g(\cdot|\theta)$ is identifiable.
3. $\|\frac{\partial^3}{\partial \theta^3} \log g(x_i|\theta)\| < h(x)$, $\forall \theta \in \text{nbhd}(\theta_0)$, $\mathbb{E}_{[X|\theta_0]} h(X_1) < \infty$
4. $\hat{\theta}_{\text{MLE}}(x)$ is the unique solution of $\dot{\ell}_x(\theta) = 0$ for every x.

If $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ then

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \overset{d}{\to} N\left(0, \{I_1^F(\theta_0)\}^{-1}\right)$$

where $I_1^F(\theta) = -\mathbb{E}_{[X_i|\theta]} \frac{\partial^2}{\partial \theta^2} \log g(X_1|\theta)$ = Fisher information at θ.

A sketch of a proof

- First order Taylor approximation:

$$0 = \frac{1}{\sqrt{n}} \dot{\ell}_X(\hat{\theta}_{\text{MLE}}) \approx \frac{1}{\sqrt{n}} \dot{\ell}_X(\theta_0) + \sqrt{n} \frac{\ddot{\ell}_X(\theta_0)^T}{n} (\hat{\theta}_{\text{MLE}} - \theta_0)$$

- Write $T_i = \frac{\partial}{\partial \theta} \log g(X_i|\theta)$.
  - $\mathbb{E} T_i = 0$, $\mathbb{V}\text{ar}\, T_i = I_1^F(\theta_0)$
  - $\frac{1}{\sqrt{n}} \dot{\ell}_X(\theta_0) = \sqrt{n}(\bar{T} - 0) \overset{d}{\to} N(0, I_1^F(\theta_0))$ [by CLT]
- Write $W_i = -\frac{\partial^2}{\partial \theta^2} \log g(X_i|\theta)$ at $\theta = \theta_0$
  - $\mathbb{E} W_i = I_1^F(\theta_0)$
  - $\frac{\ddot{\ell}_X(\theta_0)}{n} = \bar{W} \overset{p}{\to} I_1^F(\theta_0)$
- Rest is Slutksy's theorem

Three large n approximations to $\Lambda(x)$

- Assume $\theta_0 \in \Theta_0$, i.e., $A^T \theta_0 = a_0$
- Notation: $\hat{\theta}_{H_0} = \text{argmax}_{\theta \in \Theta_0} \ell_x(\theta)$ (must have $A^T \hat{\theta}_{H_0} = a_0$)
- Quadratic, Wald, Rao approximations

$$2 \log \Lambda(X) \approx \begin{cases} F(X) = (A^T \hat{\theta}_{\text{MLE}} - a_0)^T \{A^T I_x^{-1} A\}^{-1} (A^T \hat{\theta}_{\text{MLE}} - a_0) \\ W(X) = n(A^T \hat{\theta}_{\text{MLE}} - a_0)^T \{A^T I_1^F(\hat{\theta}_{\text{MLE}})^{-1} A\}^{-1} (A^T \hat{\theta}_{\text{MLE}} - a_0) \\ S(X) = \frac{1}{n} \dot{\ell}_x(\hat{\theta}_{H_0})^T \{I_1^F(\hat{\theta}_{H_0})\}^{-1} \dot{\ell}_x(\hat{\theta}_{H_0}) \end{cases}$$

- $2 \log \Lambda(X) - F(X) \overset{p}{\to} 0$, etc.
- Each of $2 \log \Lambda(X), F(X), W(X), S(X) \overset{d}{\to} \chi^2(r)$

Sketches of proof if you cared - part I

- $2\log\Lambda(X) \approx F(X)$ follows from
  - Quadratic approximation:
    $$\ell_x(\theta) \approx \ell_x(\hat\theta_{\text{MLE}}) - \frac{1}{2}(\theta - \hat\theta_{\text{MLE}})^T I_X(\theta - \hat\theta_{\text{MLE}})$$
  - and the corresponding profile log-likelihood of $\eta = A^T\theta$:
    $$\ell_x^*(\eta) \approx \ell_x(\hat\theta_{\text{MLE}}) - \frac{1}{2}(\eta - A^T\hat\theta_{\text{MLE}})^T\{A^T I_X^{-1}A\}^{-1}(\eta - A^T\hat\theta_{\text{MLE}})$$
  - and by noting $\Lambda(x) = \ell_x^*(a^T\hat\theta_{H_0}) - \ell_x^*(a^T\hat\theta_{\text{MLE}})$
- $2\log\Lambda(X) \approx W(X)$ follows similarly, use $I_X \approx nI_1^F(\hat\theta_{\text{MLE}})$.

Sketches of proof if you cared - part II

- $2\log\Lambda(X) \approx S(X)$ follows because
  - $0 = \dot\ell_x(\hat\theta_{\text{MLE}}) \approx \dot\ell_x(\hat\theta_{H_0}) + \ddot\ell_x(\hat\theta_{H_0})^T(\hat\theta_{\text{MLE}} - \hat\theta_{H_0})$
  - $\ddot\ell_x(\hat\theta_{H_0}) \approx nI_1^F(\hat\theta_{H_0})$
  - And so $\hat\theta_{\text{MLE}} - \hat\theta_{H_0} \approx I_1^F(\hat\theta_{H_0})^{-1}\dot\ell_x(\hat\theta_{H_0})$
  - Plug this into the original quadratic approximation and use $I_X \approx nI_1^F(\theta_0) \approx nI_1^F(\hat\theta_{H_0})$ because $\hat\theta_{\text{MLE}} \approx \theta_0$

Sketches of proof if you cared - part III

- It is easiest to prove $W(X) \xrightarrow{d} \chi^2(r)$
- By asymptotic normality of MLE and Slutsky's theorem
  $$W(X) \xrightarrow{d} Z^T\{A^T I_1^F(\theta_0)^{-1}A\}^{-1}Z$$
  where $Z \sim N_r(0, \{A^T I_1^F(\theta_0)^{-1}A\})$
- But $Z \sim N_r(0, \Sigma)$ means $Z^T\Sigma^{-1}Z \sim \chi^2(r)$.

Back to size calculation

- Recall: Each of $2\log\Lambda(X), F(X), W(X), S(X) \xrightarrow{d} \chi^2(r)$
- So each of the following test procedures has size α
  - ML/LRT: Reject H_0 if $2\log\Lambda(x) > q_{\chi^2}(1-\alpha, r)$
  - Quadratic: Reject H_0 if $F(x) > q_{\chi^2}(1-\alpha, r)$
  - Wald: Reject H_0 if $W(x) > q_{\chi^2}(1-\alpha, r)$
  - Rao: Reject H_0 if $S(x) > q_{\chi^2}(1-\alpha, r)$
- Here $q_{\chi^2}(u, r) = F_{\chi^2}^{-1}(u, r)$ where $F_{\chi^2}(x, r) =$ CDF of $\chi^2(r)$
- Also, one can calculate p-values as $1 - F_{\chi^2}(2\log\Lambda(x), r)$ etc.

Which one to use?

- Depends on which is easier to compute
  - To get $W(x)$ you don't need $\hat\theta_{H_0}$, only need $\hat\theta_{\text{MLE}}$.
  - To get $S(x)$ you only need $\hat\theta_{H_0}$, don't need $\hat\theta_{\text{MLE}}$.
- In many cases $F(X) = W(X)$ because $I_X = nI_1^F(\hat\theta_{\text{MLE}})$
- This happens in most generalized linear models:
  $$Y_i \overset{\text{IND}}{\sim} g(y_i|\beta, z_i) = h(z_i, y_i)e^{\xi(\beta)y_i z_i^T\beta - A(\beta, z_i)}$$
  and Wald's tests are popular choices
- Multinomial model of category counts, $S(X) =$ Pearson's-χ^2

Example: low birthweight

- Natality data $n = 500$ records on US births in June 1997.
- $Y_i = 1$ if i-th birth record has birthweight < 2500g, $Y_i = 0$ otherwise.
- $z_i = (\text{Cigarettes}_i, \text{Black}_i)$ records daily number of cigarettes and the race of the mother.
- Model
  $$\log\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \beta_1 + \beta_2\text{Cigarettes}_i + \beta_3\text{Black}_i + \beta_4\text{Cigarettes}_i \times \text{Black}_i.$$
- Can compute $\hat\beta_{\text{MLE}}$ and I_X (numerically).

MLE and curvature

- From `glm()` function in `R`

$$\hat{\beta}_{\text{MLE}} = \begin{pmatrix} -3.170 \\ 0.079 \\ 1.064 \\ -0.003 \end{pmatrix}$$

$$I_X^{-1} = \begin{pmatrix} 0.065 & -0.003 & -0.065 & 0.003 \\ -0.003 & 0.001 & 0.003 & -0.001 \\ -0.065 & 0.003 & 0.192 & -0.011 \\ 0.003 & -0.001 & -0.011 & 0.005 \end{pmatrix}$$

- How would you test the null hypothesis that for a black mother, probability of low birthweight depends on the number of cigarettes? For a non-black mother?
- How would you test that this dependence is different for black and non-black mothers?

Categorical data & Multinomial models

- Data: Counts of categories formed by one or more attributes
- Tables could be one-way, two-way, etc., depending on how many attributes are used to decide categories.

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20	15	18	14	67
	Red	11	4	24	2	41
	Brown	9	11	36	18	74
	Black	8	17	20	4	49
	Total	48	47	98	38	231

Multinomial model

- Data are k category counts of n objects $X = (X_1, \cdots, X_k)$,
- $X_j \geq 0$, $X_1 + X_2 + \cdots + X_k = n$.
- Model $X \sim \text{Multinomial}(n, p)$ where $p = (p_1, \cdots, p_k) \in \Delta_k$
- Δ_k is k-dim simplex, contains k-dim prob vectors $b = (b_1, \cdots, b_k)$, i.e., $b_j \geq 0\ \forall j$ and $\sum_j b_j = 1$)
- $\text{Multinomial}(n, p)$ has pmf

$$f(x|p) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

for $x = (x_1, \cdots, x_k)$ with $x_i \geq 0$ and $\sum_i x_i = n$, $f(x|p) = 0$ otherwise.

Multinomial model (contd.)

- This is an extension of the binomial distributions. In X we record the counts from n independent trials with k outcomes with probabilities given by p.
- Consequently $X_j \sim Bin(n, p_j)$ for any single category $1 \leq j \leq k$.
- Similarly for two categories $j_1 \neq j_2$, $X_{j_1} + X_{j_2} \sim Bin(n, p_{j_1} + p_{j_2})$, and so on.

Multinomial model (contd.)

- For two-way tables, with k_1 rows and k_2 columns, we can write X and p either as $k_1 k_2$-dim vectors, or more commonly as $k_1 \times k_2$ matrices.
- Even when they're written as matrices, we'll write $X \sim \text{Multinomial}(n, p)$ to mean that the multinomial distribution is placed on the vector forms of X and p.
- For multi-way table we'll use arrays of appropriate dimensions to represent X and p.

MLE

- To obtain the MLE of p based on data $X = x$, we can maximize the log-likelihood function in p with an additional Lagrange component to account for the constraint $\sum_{j=1}^k p_j = 1$:

$$\tilde{\ell}_x(p, \lambda) = \text{const} + \sum_{j=1}^k x_j \log p_j + \lambda(\sum_{j=1}^k p_j - 1)$$

- The solution in p equals:

$$\hat{p}_{\text{MLE}} = \left(\frac{x_1}{n}, \cdots, \frac{x_k}{n}\right)$$

Hypothesis testing: Mendel's peas

Mendel, the founder of modern genetics, studied how physical characteristics are inherited in plants. His studies led him to propose the laws of segregation and independent assortment. We'll test this in a simple context. Under Mendel's laws, when pure round-yellow and pure green-wrinkled pea plants are cross-bred, the next generation of plant seeds should exhibit a 9:3:3:1 ratio of round-yellow, round-green, wrinkled-yellow and wrinkled-green combinations of shape and color. In a sample of 556 plants from the next generation the observed counts for these combinations are (315, 108, 101, 32). Does the data support Mendel's laws?

Formalization

- Data $X = (X_1, X_2, X_3, X_4)$ giving the category counts of the four types of plants
- Model $X \sim \text{Multinomial}(n = 556, p)$, $p \in \Delta_4$
- Test $H_0 : p = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$

Hardy-Weinberg equilibrium

The spotting on the wings of Scarlet tiger moths are controlled by a gene that comes in two varieties (alleles) whose combinations (moths have pairs of chromosomes) produce three varieties of spotting pattern: "white spotted", "little spotted" and "intermediate". If the moth population is in Hardy-Weinberg equilibrium (no current selection drift), then these varieties should be in the ratio $a^2 : (1-a)^2 : 2a(1-a)$, where $a \in (0,1)$ denotes the abundance of the dominant white spotting allele. In a sample of 1612 moths, the three varieties were counted to be 1469, 5 and 138. Is the moth population in HW equilibrium?

Formalization

- Data $X = (X_1, X_2, X_3)$ the category counts of the three spotting patterns,
- model $X \sim \text{Multinomial}(n = 1612, p)$, $p \in \Delta_3$,
- Test whether $H_0 : p \in \Delta_3^{HW}$,
- Δ_3^{HW} is a subset of Δ_3 containing all vectors of the form $(a^2, (1-a)^2, 2a(1-a))$ for some $a \in (0,1)$.

Hair and eye color

- Are hair color and eye color two independent attributes?
- In this case, writing p in the matrix form $p = ((p_{ij}))$, $1 \le i \le k_1$ and $1 \le j \le k_2$, we want to test if $p_{ij} = p_i^{(1)} p_j^{(2)}$ for some $p^{(1)} \in \Delta_{k_1}$ and $p^{(2)} \in \Delta_{k_2}$.
- It's elementary that if p_{ij} factors as above, then $p_i^{(1)} = \sum_{j=1}^{k_2} p_{ij}$ and $p_j^{(2)} = \sum_{i=1}^{k_1} p_{ij}$.
- Writing the row and column totals as $p_{i\cdot}$ and $p_{\cdot j}$, the test of independence is often represented by

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \forall i, j$$

- We'll use $\Delta^I_{k_1,k_2}$ to denote this set.

ML testing

- $\dim(\Delta_k) = k - 1$, as vector $p \in \Delta_k$ satisfy $\sum_{i=1}^k p_i = 1$.
- Hypotheses $H_0 : p \in \Theta_0$, Θ_0 is a sub-simplex of Δ_k of dimension $q < k$.
  - Mendel's peas: $\Theta_0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$, $q = 0$
  - HW equilibrium: $\Theta_0 = \Delta_3^{HW}$, $q = 1$
  - Independence: $\Theta_0 = \Delta^I_{k_1,k_2}$, $q = k_1 + k_2 - 2$.
- Could use any of ML, quadratic, Wald or Rao tests

Pearson's χ^2 tests

- Rao's test statistics boils down to a very convenient form:

$$S(X) = \sum_{j=1}^{k} \frac{(X_j - n\hat{p}_{H_0,j})^2}{n\hat{p}_{H_0,j}}$$
$$= \sum \frac{(Observed - Expected)^2}{Expected}$$

- For categorical data, this was earlier discovered by Pearson who also ascertained its approximate χ^2 distribution.
- Because this is Rao statistic, we have $S(X) \xrightarrow{d} \chi^2(k-1-q)$

Example 1: point null

- For testing $H_0 : p = p_0$ against $p \neq p_0$, where $p_0 = (p_{0,1}, \cdots, p_{0,k}) \in \Delta_k$ is a fixed probability vector of interest (as in Mendel's peas example), then

$$S(X) = \sum_{j=1}^{k} \frac{(X_j - np_{0,j})^2}{np_{0,j}}$$

which is asymptotically $\chi^2(k-1-0) = \chi^2(k-1)$.

- Size-α Pearson's test rejects H_0 if $S(X) > q_{\chi^2}(1-\alpha, k-1)$.

Example 2: parametric form

- HW test:

$H_0 : p_0 = (\eta^2, \eta(1-\eta), \eta(1-\eta), (1-\eta)^2)$, for some $0 < \eta < 1$

- To compute $\hat{p}_{H_0}$ it is equivalent to write the likelihood function in η and maximize:

$$L_X(\eta) = \text{const.} \times \{\eta^2\}^{x_1}\{\eta(1-\eta)\}^{x_2}\{\eta(1-\eta)\}^{x_3}\{(1-\eta)^2\}^{x_4}$$
$$= \text{const.} \times \eta^{2x_1+x_2+x_3}(1-\eta)^{x_2+x_3+2x_4}$$

and so $\hat{\eta}_{\text{MLE}} = \frac{2x_1+x_2+x_3}{2n}$.

Example 2: parametric form (contd.)

- Once we have $\hat{\eta}_{\text{MLE}}$, we can construct $\hat{p}_{H_0} = (\hat{\eta}^2_{\text{MLE}}, \hat{\eta}_{\text{MLE}}(1-\hat{\eta}_{\text{MLE}}), \hat{\eta}_{\text{MLE}}(1-\hat{\eta}_{\text{MLE}}), (1-\hat{\eta}_{\text{MLE}})^2)$ and evaluate $S(X)$.
- Because Θ_0 has dimension $q = 1$ (only a single number η needs to be known), the asymptotic distribution of $S(X)$ is $\chi^2(k-2)$.
- The same calculations carry through for a more general parametric form:

$$H_0 : p = (g_1(\eta), \cdots, g_k(\eta))$$

where $\eta \in \mathcal{E}$ is q-dim vector and $g_1(\eta), \cdots, g_k(\eta)$ are functions such that for every $\eta \in \mathcal{E}$, $(g_1(\eta), \cdots, g_k(\eta)) \in \Delta_k$.

Example 3: independence

- Consider testing $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \forall i, j$
- To get $\hat{p}_{H_0}$, we write the likelihood function in terms of $p_{i\cdot}$, $1 \leq i \leq k_1$ and $p_{\cdot j}$, $1 \leq j \leq k_2$:

$$L_x(p_{1\cdot}, \cdots, p_{k_1\cdot}, p_{\cdot 1}, \cdots, p_{\cdot k_2}) = \text{const.} \times \prod_{i=1}^{k_1}\prod_{j=1}^{k_2}(p_{i\cdot}p_{\cdot j})^{x_{ij}}$$
$$= \text{const.} \times \left\{\prod_{i=1}^{k_1} p_{i\cdot}^{x_{i\cdot}}\right\}\left\{\prod_{j=1}^{k_2} p_{\cdot j}^{x_{\cdot j}}\right\}$$

where $x_{i\cdot} = \sum_{j=1}^{k_2} x_{ij}$ and $x_{\cdot j} = \sum_{i=1}^{k_1} x_{ij}$ are the margin counts of our two-way table.

Example 3: independence (contd.)

- Because $(p_{1\cdot}, \cdots, p_{k_1\cdot}) \in \Delta_{k_1}$ and $(p_{\cdot 1}, \cdots, p_{\cdot k_2}) \in \Delta_{k_2}$, the maximizer is given by:

$$\hat{p}_{i\cdot} = \frac{x_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{x_{\cdot j}}{n}, \quad 1 \leq i \leq k_1, 1 \leq j \leq k_2$$

- And so $\hat{p}_{H_0}$ has coordinates: $\hat{p}_{H_0,ij} = \frac{x_{i\cdot}x_{\cdot j}}{n^2}$ giving

$$S(X) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(X_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{n})^2}{\frac{X_{i\cdot}X_{\cdot j}}{n}}$$

- Because $\dim(\Theta_0) = q = k_1 - 1 + k_2 - 1$, $S(X) \xrightarrow{d} \chi^2(k_1k_2 - 1 - k_1 + 1 - k_2 + 1) = \chi^2((k_1-1)(k_2-1))$.

Hair-eye color

		Eye color				
		Blue	Green	Brown	Black	Total
Hair color	Blonde	20 (13.9)	15 (13.6)	18 (28.4)	14 (11.0)	67
	Red	11(8.5)	4 (8.3)	24 (17.4)	2 (6.7)	41
	Brown	9 (15.4)	11 (15.1)	36 (31.4)	18 (12.2)	74
	Black	8 (10.2)	17 (10.0)	20 (20.8)	4 (8.1)	49
	Total	48	47	98	38	231

- $S(x) = 30.9$.
- So the p-value is $1 - F_{(4-1)(4-1)}(30.9) = 1 - F_9(30.9) \approx 0$.