

Likelihood - maths

Nathaniel Beck

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

QII - Week 3

1

Some preliminary stuff

- Define STATISTIC: s any function of the data
- (sample mean, sample quartile, three times first observation plus square root of 7th observation, or whatever)
- Define SUFFICIENT STATISTIC (FOR SOME PARAMETER ESTIMATE)
- The estimate, given the statistic, is identical to the estimate given all the data
- that is, all the information in the data (about the parameter) is contained in the statistic.
- Sample mean is sufficient statistic for estimating λ in Poisson
- Thus any sample data that produces the same mean will produce the same estimate (no difference if -100000,0,100000 or 0,0,0)

2

- For function of K variables $y = f(\mathbf{x})$, y is scalar
- the gradient (\mathbf{g}) is the K vector of first partial derivatives
- The Hessian (\mathbf{H}) is the $K \times K$ matrix of second partials (cross-partial off diagonal)
- If x is scalar, these are usual first and second derivatives
- Computer can compute these numerically
- First derivative, just have computer take δ to be some really tiny positive number, and compute
- $f'(x) = \frac{f(x+\delta) - f(x)}{\delta}$ at any given x
- BUT ALWAYS EXAMINE THE FIRST DERIVATIVE ANALYTICALLY
- IT TELLS YOU A LOT

3

Taylor series

- Any continuous function smooth enough to have a bunch of derivatives
- can be approximated (locally) by a polynomial of sufficient order (p)
- Scalar x - approximate around x_0 (so $f(x_0)$ means the function evaluated at x_0 and likewise for derivatives)

$$f(x) = f(x_0) + \sum_{i=1}^p \frac{1}{i!} \frac{d^i f(x_0)}{d(x_0)^i} (x - x_0)^i \quad (1)$$

- Late terms go to zero quickly if x is close to x_0 since small number raised to big power (and also divided by $i!$ which gets big)
 - Linear approximation $f(x) = f(x_0) + \frac{df(x_0)}{d(x_0)}(x - x_0)$
 - Quadratic approximation
- $$f(x) = f(x_0) + \frac{df(x_0)}{d(x_0)}(x - x_0) + \frac{1}{2} \frac{d^2 f(x_0)}{d(x_0)^2} (x - x_0)^2$$
- Note that higher order derivatives represent “wildness”
 - When divided by large N assumption they go to zero rules out “wildness”

4

- Similarly for function of vector \mathbf{x} (returning scalar y)
- Linear $f(\mathbf{x}) = f(\mathbf{x}_0) + \mathbf{g}(\mathbf{x}_0)'(\mathbf{x} - \mathbf{x}_0)$
- Quadratic $f(\mathbf{x}) = f(\mathbf{x}_0) + \mathbf{g}(\mathbf{x}_0)'(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)'\mathbf{H}(\mathbf{x} - \mathbf{x}_0)$

5

Basics for properties of ML

- Let Θ be the parameter space
- $\hat{\Theta}$ is the ml estimator and Θ_0 is true value
- Take a linearization of the derivative of the likelihood at $\hat{\Theta}$ around the Θ_0
- As before, a specific value in a function (or derivative) is evaluated at that point

$$\frac{\partial L}{\partial \hat{\Theta}} = \frac{\partial L}{\partial \Theta_0} + \frac{\partial^2 L}{\partial \Theta_0 \partial \Theta_0'} (\hat{\Theta} - \Theta_0) \quad (2)$$

- but since at the maximum the partial is zero, we get

$$\hat{\Theta} - \Theta_0 = \left\{ \frac{-\partial^2 L}{\partial \Theta_0 \partial \Theta_0'} \right\}^{-1} \frac{\partial L}{\partial \Theta_0} \quad (3)$$

6

Consistency

- To show consistency, we need to show that the plim of the last equation is zero
- Notice how similar this is to the proof for OLS
- Multiply the first term (inverse) by N and divide the second term by N
- ASSUME (regularity condition) that the first term (\mathbf{H}) converges to some finite value (not a hard condition)
- The second term in expectation (\mathbf{g}) is zero
- Is average of the partials of the log of the likelihood for each observation
- By law of large numbers, this just converges to the expectation of the partial of the log likelihood of each observation
- By iid assumption these are the same (can weaken)
- But the expectation of the likelihood has a maximum at the true value
- Follows by some easy but not enlightening theorems on integrals of densities
- So expectation of derivative is zero
- So ML is consistent

7

Asymptotic normality

- Now I look at the limiting distribution of $\frac{\Theta - \Theta_0}{\sqrt{N}}$
- The first term, by assumption converges to something positive definite
- The second term converges to a normal by the central limit theorem
- So $\hat{\Theta}$ is normally distributed, with mean Θ_0

8

- The variance of $\hat{\Theta}$ is just the expectation of the outer product of the above

$$V(\hat{\Theta}) = E\{(\hat{\Theta} - \Theta_0)(\hat{\Theta} - \Theta_0)'\} \quad (4)$$

$$= E \left[\left\{ \frac{-\partial^2 L}{\partial \Theta_0 \partial \Theta_0'} \right\}^{-1} \frac{\partial L}{\partial \Theta_0} \frac{\partial L}{\partial \Theta_0}' \left\{ \frac{-\partial^2 L}{\partial \Theta_0 \partial \Theta_0'} \right\}^{-1} \right] \quad (5)$$

9

Score vector and information matrix

- Some more terminology
- The *score* is the gradient of the likelihood ($\frac{\partial L}{\partial \Theta}$)
- The *information matrix*, $I(\Theta_0)$, is used to compute variances

$$I(\Theta_0) = -E \left[\frac{\partial^2 \log L(\Theta_0)}{\partial \Theta_0 \partial \Theta_0'} \right] \quad (6)$$

$$= -E \left[\frac{\partial \log L(\Theta_0)}{\partial \Theta_0} \frac{\partial \log L(\Theta_0)}{\partial \Theta_0'} \right] \quad (7)$$

- The latter is based on another regularity condition

- Regularity conditions on second partials give
- The expectation of the outer product of the scores is equal to the information matrix
- So the variance (Equation 5) reduces to the inverse of the information matrix
- We can read off the standard errors of $\hat{\Theta}$ from the square roots of the diagonal elements of this matrix
- Note that these are only correct asymptotically
- Off diagonal terms give asymptotic covariances of the elements of Θ

11

Rao-Cramer bound

- The famous theorem of Rao and Cramer says
- The minimum (in a matrix sense) variance of an unbiased estimator
- is given by the inverse of the information matrix
- ML achieves this bound
- Hence, if a MVUE exists, it is ML (but one may not exist)

12

- If a minimum variance unbiased estimator (MVUE) exists, the MLE estimator will be it
- VCV of ML is the inverse of the information matrix
- Invariance: If $\hat{\Theta}$ is an MLE of Θ , then $f(\hat{\Theta})$ is the MLE of $f(\Theta)$.
- Invariance to Sampling Plan: The data affect estimates only through the likelihood function; information about the sampling plan that does not affect the likelihood is irrelevant
- ML is consistent
- ML is BAN

13

Robust Errors

- If model is not well specified but the mean function is correctly specified
- and the variance function is not horribly specified than ML
- is asymptotically normal with variance-covariance matrix

$$V(\hat{\Theta}) = I^{-1} \frac{\partial L}{\partial \Theta_0} \frac{\partial L'}{\partial \Theta_0} I^{-1} \quad (8)$$

which we call the robust variances (from Equation 5.

- This is the maximum likelihood analogue of White's consistent standard errors.
- The middle term is call the OPG (outer product of the gradient)

14

Tests of hypotheses in ML

- We can use the asymptotic normality of MLE to construct tests
- tests like the common t (really z) or F tests are of this nature
- For historical reasons, these are called *Wald* tests for Abraham Wald
- We can also do *likelihood ratio* tests.
- These are based on the large sample property that
- twice the log of the ratio of the likelihood for two models
- if one is *nested* inside the other (that is, the simpler model is just the bigger model with some constraints imposed)
- has a χ^2 distribution with degrees of freedom equal to the number of constraints.
- Thus we can take the two models (assuming one is nested inside the other) and form the test statistic twice the difference of the logs of the likelihood functions which has the appropriate χ^2 distribution.
- A third method, *Lagrange multiplier* tests, are based on the “cost” of imposing the constraint implied by the simpler model.
- The three tests are asymptotically equivalent, but may differ in small samples.

15

Why is the LR chi sq?

- The likelihood ratio is

$$LR = -2 \log \lambda = -2 \log \left(\frac{L(\hat{\Theta}^R)}{L(\hat{\Theta}^U)} \right) = 2 \{ \log L(\hat{\Theta}^U) - \log L(\hat{\Theta}^R) \} \quad (9)$$

where $L(\hat{\Theta}^U)$ is the maximum likelihood estimate of the unrestricted model and $L(\hat{\Theta}^R)$ is the likelihood of the restricted model (so must be lower, because some parameters are restricted, typically assumed to be zero)

- Note that $L(\hat{\Theta}^R) < L(\hat{\Theta}^U)$ so $\lambda < 1$ and $LR > 0$
- This is good since χ^2 is non-negative (sum of squares).
- Using our usual (and only!) trick, taking a second order Taylor expansion around the maximum likelihood estimate $\hat{\Theta}^U$ (the estimate of the unrestricted model is from ml), we get
- Note change in notation - D stands for the derivative operator

16

-

$$\log L(\hat{\Theta}^R) = \log L(\hat{\Theta}^U) + (\hat{\Theta}^R - \hat{\Theta}^U)' D \log L(\hat{\Theta}^U) + \frac{(\hat{\Theta}^R - \hat{\Theta}^U)' D^2 \log L(\hat{\Theta}^U) (\hat{\Theta}^R - \hat{\Theta}^U)}{2} \quad (10)$$

- Since $\hat{\Theta}^U$ is the maximum likelihood estimator, the derivative of the log likelihood at $\hat{\Theta}^U$ must be zero and hence

$$-2 \log \lambda = (\hat{\Theta}^R - \hat{\Theta}^U)' D^2 \log L(\hat{\Theta}^U) (\hat{\Theta}^R - \hat{\Theta}^U) \quad (11)$$

- But the $(\hat{\Theta}^R - \hat{\Theta}^U)$ terms are normal deviates
- and quadratic forms of normal deviates are χ^2 .