Likelihood

Nathaniel Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093
beck@ucsd.edu
http://weber.ucsd.edu/∼nbeck

April, 2001

Prelims

Let $\mathbf{y}$ be observations on y_i for individuals $i = 1, \ldots, N$. ($\mathbf{y}$ is stochastic.) Let $\hat{\theta}$ be a vector of parameters of interest.

Note: this is a much more general setup than we had previously. Here the $\mathbf{X}$'s are just "stuff" and only the stochastic dependent variable is modelled. This will be clearer when we do OLS below.

By Bayes Theorem, the conditional density of $\hat{\theta}|\mathbf{y}$,

$$\Lambda(\hat{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\hat{\theta})g(\hat{\theta})}{h(\mathbf{y})} \tag{1}$$

Note that the denominator is just a function of the data. It only makes sense to compare these conditional densities for the same data, so we can ignore the denominator.

The second term in the numerator is the "prior" density of $\hat{\theta}$ which is fixed before our observations and hence is invariant over our problem.

What is left to compare is called the "likelihood", that is,

$$L(\hat{\theta}|\mathbf{y}) = f(\mathbf{y}|\hat{\theta}) \tag{2}$$

This is the fundamental likelihood equation.

It only makes sense to compare likelihoods for the same set of data and the same prior. In words, the likelihood is the sample information that transforms a "prior" into a "posterior" density for θ.

We then think of L as a function of $\hat{\theta}$, ($\mathbf{y}$ is already observed, so everything is conditioned on that observation). The best estimator, $\hat{\theta}$ is then whatever value of $\hat{\theta}$ maximizes

$$L(\hat{\theta} = f(\mathbf{y}|\hat{\theta}) \tag{3}$$

Likelihood for Independent Observations

IF the y_i are all independent (or conditionally independent, given x_i)

$$L = f(\mathbf{y}) = \prod_{i=1}^{N} f(y_i|\hat{\theta}) \tag{4}$$

or

$$\log L = \sum_{i=1}^{N} \log f(y_i|\hat{\theta}) \tag{5}$$

Since likelihoods are all positive, the likelihood and its log have their maxima at the same place (simple theorem from calculus). It is almost always much simpler to work with the log of the likelihood since get to sum things up

For this course, all logs are natural, NOT base 10. Sometimes you will see $\ln$ used, $\ln$ and $\log$ are interchangeable.

Information matrix and se's

The information matrix, I(θ), is used to compute variances.

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

$$= -E\left[\frac{\partial \log L(\vec{\theta})}{\partial\vec{\theta}}\frac{\partial \log L(\vec{\theta})}{\partial\vec{\theta}'}\right] \tag{7}$$

The variance–covariance matrix of the maximum likelihood estimator, $\vec{\theta}$ is $[I(\vec{\theta})]^{-1}$. Note: All derivatives and expectations are evaluated at the maximum likelihood point. The expectation is over the probability distribution of $\mathbf{y}$. With data this expectation is calculated by just computing the matrix of second derivatives at the observed data. With enough data, the this computation is very close to the theoretical expectation.

(The matrix of second derivatives is called the *Hessian*.)

The inverse of the information matrix is also the Rao-Cramer bound, the smallest variance-covari ance matrix that an unbiased estimator can attain.

The se's of the estimator are just the square roots of the diagonal terms of the VCV matrix

Why - easy and hard

The easy way to see this is to look at the curvature of the likelihood and note that the second derivative is a measure of curvature.

The hard way is correct, but is hard and may not be helpful. If it isn't, nothing lost.

Where does this come from? Take a linearization of the derivative of the likelihood at the maximum likelihood point, $\hat{\beta}$ around the true value, β.

$$0 = |\frac{\partial L}{\partial \beta}\Big|_{\hat{\beta}} \tag{8}$$

$$= |\frac{\partial L}{\partial \beta}\Big|_{\beta} + \tag{9}$$

$$\frac{\partial^2 L}{\partial \beta \partial \beta'}(\hat{\beta} - \beta) \tag{10}$$

so

$$\hat{\beta} - \beta = \left\{ \frac{-\partial^2 L}{\partial \beta \partial \beta'} \right\}^{-1} \frac{\partial L}{\partial \beta} \tag{11}$$

But the variance of $\hat{\beta}$ is just the expectation of the outer product of the above, so

$$V(\hat{\beta}) = E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\} \tag{12}$$

$$= E\left[\left\{ \frac{-\partial^2 L}{\partial \beta \partial \beta'} \right\}^{-1} \frac{\partial L}{\partial \beta} \frac{\partial L'}{\partial \beta} \left\{ \frac{-\partial^2 L}{\partial \beta \partial \beta'} \right\}^{-1} \right] \tag{13}$$

The *score* is the gradient of the likelihood ($\frac{\partial L}{\partial \beta}$).

If the model is correctly specified the expectation of the outer product of the scores is equal to the information matrix, so the variance (Equation 13) reduces to the inverse of the information matrix.

We can read off the standard errors of $\hat{\beta}$ from the square roots of the diagonal elements of this matrix. Note that these are only correct asymptotically and are hard to calculate in finite samples.

Properties of MLE's

If a minimum variance unbiased estimator (MVUE) exists, the MLE estimator will be it

Invariance: If $\hat{\vec{\theta}}$ is an MLE of $\vec{\theta}$, then $g(\hat{\vec{\theta}})$ is the MLE of $g(\vec{\theta})$.

(Query - how do we compute standard errors of transformed variables? Not by transforming se of original! Easiest is to take the 95% confidence interval and transform that. Can also use simulation. Can use analytic delta method, though this rests heavily on quadratic approximation. - To understand how to proceed, read King, Tomz and Wittenberg)

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

Large sample properties of MLE's

Consistent

Asymptotically normal

Variance-Covariance is the Rao-Cramer bound (if the model is well specified)

Thus ML is best asymptotically normal (BAN)

Small sample properties of MLE's

Know little. Can do by Monte Carlo. But, in general, if a good small sample estimator exists, it will look like maximum likelihood. Don't have many, if any, examples of good small sample estimators which are not maximum likelihood.

Note that ml may not be unbiased in small samples. Thus ml estimator of σ divides the sum of squares by N, not N-1. Not very big difference.

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

$$V(\hat{\beta}) = I^{-1} \frac{\partial L}{\partial \beta} \frac{\partial L}{\partial \beta}' I^{-1} \tag{14}$$

which we call the robust variances (from Equation 13. This is the maximum likelihood analogue of White's consistent standard errors.

Note: This is just a restatement of our previous slide which computed variances.

Tests of hypotheses in ML

We can use the asymptotic normality of MLE to construct tests; tests like the common t or F tests are of this nature. For historical reasons, these are called *Wald* tests for Abraham

Wald who did the unifying theory (the various tests predate a precise mathematical understanding of them).

We can also do *likelihood ratio* tests. These are based on the large sample property that twice 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, as in asking if we should use the $\mathbf{X}_2$ variables in a regression) 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 $2(\log(\text{likelihood of Model 1}) - \log(\text{likelihood of Model 2})$ which has the appropriate χ^2 distribution.

To make your lives easier as to which to subtract from which, remember that χ^2's are always positive.

A third method, *Lagrange multiplier* tests, are based on the "cost" of imposing the constraint implied by the simpler model.

While LM tests look complicated, they are often the easiest to do, since they only require estimation under the null hypothesis that the constraints hold.

They are invaluable in time series analysis

The three tests are asymptotically equivalent, but may differ in small samples. No one knows which is better in small samples, so may choose on grounds of convenience.

A simple picture (not in notes) should help.

MLE and OLS

The standard setup of OLS looks different than what we have just seen, but we can easily renotate linear models to fit what we have just done. This will also highlight some differences.

The new notation is

$$\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{\Omega}) \tag{15}$$

Assume that $\mathbf{\Omega}$ meets the Gauss-Markov assumptions previously discussed. Note that we then have $E(y_i) = \mathbf{x}_i\beta + \epsilon_i$ where the errors have all the nice properties.

Note that we are here assuming that the $\mathbf{y}$ are normally distributed, which we didn't do previously. But as we shall see, either normality is benign or OLS isn't right.

But, one issue with MLE is that you need distributional assumptions. On the other hand, it is hard to think of an interesting case where this is a weakness of MLE as compared to other methods.

The ugly calculus

The derivation of the ML estimation of the β is easy. The observations are independent; for a univariate normal (y_i), we have that

$$y_i \sim N(\mathbf{x}_i\beta, \sigma^2) \tag{16}$$

(Note we use variance, not sd in normal notation, we are no longer children)

or, for each observation,

$$L_i = f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\beta)^2}{2\sigma^2}} \tag{17}$$

$$\log(L_i) = \log(f(y_i) =$$

$$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\beta)^2 \tag{18}$$

so for all observations

$$\log(L) = \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\beta)^2 \tag{19}$$

or, putting it together in one big matrix

$$\log(L) =$$

$$-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{20}$$

The scalar setup

We then take derivatives. Working with the scalar setup (Equation 19), we have

$$\frac{\partial \log(L)}{\partial \beta} =$$

$$\frac{1}{2\sigma^2} \sum_{i=1}^{N} -(y_i - \mathbf{x}_i\beta)\mathbf{x}_i \quad (21)$$

This derivative is zero when

$$\sum y_i \mathbf{x}_i = \hat{\beta} \sum \mathbf{x}_i \mathbf{x}_i \quad (22)$$

which are the K OLS normal equations.

This shows the intimate tie between normality and least squares, and either least squares justifies ML or vice versa. But if you worry about normality, then you have to worry about least squares.

The calculus to find the value of $\hat{\sigma^2}$ which sets the appropriate derivative to zero is slightly more tedious calculus,

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} e_i^2 \quad (23)$$

which yields when the derivative is set to zero

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{N} e_i^2}{N} \quad (24)$$

where the e_i are the usual OLS residuals.

Note this differs from OLS formula by dividing by N not $N - K$. MLE is NOT unbiased (but who cares????). Asymptotically the two converge, N and $N - K$ are quite similar for large N and fixed K.

The matrix setup

Working with Equation 20, we get

$$\frac{\partial \log(L)}{\partial \beta} = \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \quad (25)$$

which when set to zero yields the normal equations.

For estimating σ^2, we use

$$\frac{\partial \log(L)}{\partial \sigma^2} = \frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{26}$$

which when set to zero yields the same estimator as in the scalar case.

We can get the VCV matrix by taking the $K+1 \times K+1$ Hessian (second derivative) matrix, and then taking the negative of the expectation and inverting (Equation 7).

The Hessian is

$$\begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & -\frac{1}{\sigma^4}\mathbf{X}'\epsilon \\ -\frac{1}{\sigma^4}\epsilon'\mathbf{X} & \frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6} \end{bmatrix} \tag{27}$$

On expectation, the two off diagonal terms are zero ($\mathbf{X}$ and ϵ are uncorrelated) and we have already have the relationship of the expectation of $\epsilon'\epsilon$ and σ^2, so we are inverting a diagonal matrix.

Thus the VCV matrix of the estimates is

$$\begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{N} \end{bmatrix} \tag{28}$$

Note what this shows: the estimate $\hat{\beta}$is independent of the estimate $\hat{\sigma^2}$ (which is why in OLS we can first estimate β without worrying about σ^2 and then use those estimates to estimate σ^2, the independence of the mean and variance is a peculiar (but very handy) feature of normals). Note also we get standard errors on our estimate of σ^2 (though these are seldom reported).

So maximum likelihood produces all the OLS produces and more; where they both produce, they more or less produce the same thing. OLS has no basic theoretical justification; MLE does.

Binary dv models

Suppose the dependent variable, y_i takes on the values 0 and 1 (success-failure, vote-abstain, Republican-Democrat), etc. Let π_i be the i'th person's probability of success (1). Then the likelihood is

$$L = \prod_{i=1}^{N} (\pi_i)^{y_1}(1-\pi_i)^{(1-y_i)} \tag{29}$$

and the log likelihood is

$$\log L = \sum_{i=1}^{N} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \tag{30}$$

To make this work we need to parameterize π_i as a function of independent variables, x_i and a parameter vector, β. Suppose that we expect π to increase as a linear function of x, that is, π increases as $x'\beta$ increases. So

$$\pi_i = f(x'\beta) \tag{31}$$

where f is monotonically increasing (or at least non-decreasing) Now π is limited to being between 0 and 1.

Probit/Logit

Any probability distribution function will do for f. One common model is *probit*

$$f(x'\beta) = \Phi(x'\beta) \tag{32}$$

while another common model is *logit*

$$f(x'\beta) = \frac{1}{1 + e^{-x'\beta}} \tag{33}$$

Interpreting coefficients

Elasticities are useful (a 1% change in x_i leads to what % change in dependent variable)
Partial derivatives ($\frac{\partial y}{\partial x_i}$) are also useful. ($x_i$ is i'th independent variable.)
Note that for a linear equation $\hat{\beta}_i$ is exactly $\frac{\partial y}{\partial x_i}$ while the elasticity is $\hat{\beta}_i$ in a log-log form.
For logit, sometimes calculate elasticities. But since function is non-linear, where calculate?
Often at mean, sometimes at other interesting values.
For logit,

$$\frac{\partial \pi}{\partial x_i} = \frac{\partial [1 + e^{-x'\hat{\beta}}]^{-1}}{\partial x_i} \tag{34}$$

$$= \hat{\beta}_i \hat{\pi}(1 - \hat{\pi}) \tag{35}$$

Can use the simulation methods of KTW to both calculate "quantities of interest" (probabilities) and also se's of those QOI's.

How? Estsimp, setx, simqi

Note that this is based on MLE coef estimates being (asymptotically) multivariate normal. May be a problem for small samples.