Maximum likelihood

Class Notes

Manuel Arellano

Revised: January 17, 2018

1 Likelihood models

Given some data $\{w_1, ..., w_n\}$, a likelihood model is a family of density functions $f_n(w_1, ..., w_n; \theta)$, $\theta \in \Theta \subset R^p$, that is specified as a probability model for the observed data. If the data are a random sample from some population with density function $g(w_i)$ then $f_n(w_1, ..., w_n; \theta)$ is a model for $\prod_{i=1}^n g(w_i)$. If the model is correctly specified $f_n(w_1, ..., w_n; \theta) = \prod_{i=1}^n f(w_i, \theta)$ and $g(w_i) = f(w_i, \theta_0)$ for some value θ_0 in Θ. The function f may be a conditional or an unconditional density. It may be flexible enough to be unrestricted or it may place restrictions in the form of the population density.[1]

In the likelihood function $\mathcal{L}(\theta) = \prod_{i=1}^n f(w_i, \theta)$, the data are given and θ is the argument. Usually we work with the log likelihood function:

$$L(\theta) = \sum_{i=1}^n \ln f(w_i, \theta). \tag{1}$$

$L(\theta)$ measures the ability of different densities within the pre-specified class to generate the data. The maximum likelihood estimator (MLE) is the value of θ associated with the largest likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta). \tag{2}$$

Example 1: Linear regression under normality The model is $y \mid X \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$ so that

$$f_n(y_1, ..., y_n \mid x_1, ..., x_n; \theta) = \prod_{i=1}^n f(y_i \mid x_i; \theta)$$

where $\theta = \left(\beta', \sigma^2\right)'$ and

$$f(y_i \mid x_i; \theta) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2\right] \tag{3}$$

$$L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - x_i'\beta\right)^2 \tag{4}$$

The MLE $\widehat{\theta} = \left(\widehat{\beta}', \widehat{\sigma}^2\right)'$ consists of the OLS estimator $\widehat{\beta}$ and the residual variance $\widehat{\sigma}^2$ without degrees of freedom adjustment. Letting $\widehat{u} = y - X\widehat{\beta}$ we have:

$$\widehat{\beta} = \left(X'X\right)^{-1} X'y \qquad \widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{n}. \tag{5}$$

[1] This note follows the practice of using the term *density* both for continuous random variables and for the probability function of discrete random variables; as for example in David Cox, *Principles of Statistical Inference*, 2006.

Example 2: Logit regression There is a binary dependent variable y_i that takes only two values, 0 and 1. Therefore, in this case $f(y_i \mid x_i) = p_i^{y_i}(1-p_i)^{(1-y_i)}$ where $p_i = \Pr(y_i = 1 \mid x_i)$.

In the logit model the log odds ratio depends linearly on x_i:

$$\ln\left(\frac{p_i}{1-p_i}\right) = x_i'\theta,$$

so that $p_i = \Lambda(x_i'\theta)$ where Λ is the logistic cdf $\Lambda(r) = 1/(1+\exp(-r))$.

Assuming that $f_n(y_1, ..., y_n \mid x_1, ..., x_n; \theta) = \prod_{i=1}^n f(y_i \mid x_i; \theta)$, the log likelihood function is

$$L(\theta) = \sum_{i=1}^n \left\{ y_i \ln \Lambda\left(x_i'\theta\right) + (1-y_i) \ln\left[1 - \Lambda\left(x_i'\theta\right)\right] \right\}.$$

The first and second partial derivatives of $L(\theta)$ are:

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n x_i \left[y_i - \Lambda\left(x_i'\theta\right)\right]$$

$$\frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'} = -\sum_{i=1}^n x_i x_i' \Lambda_i (1-\Lambda_i).$$

Since the Hessian matrix is negative semidefinite, as long as it is nonsingular, there exists a single maximum to the likelihood function.[2] The first order conditions $\partial L(\theta)/\partial\theta = 0$ are nonlinear but we can find their root $\widehat{\theta}$ using the Newton-Raphson method of successive approximations. Namely, we begin by finding the root θ_1 to a linear approximation of $\partial L(\theta)/\partial\theta$ around some initial value θ_0 and iterate the procedure until convergence:

$$\theta_{j+1} = \theta_j - \left(\frac{\partial^2 L(\theta_j)}{\partial\theta\partial\theta'}\right)^{-1} \frac{\partial L(\theta_j)}{\partial\theta} \qquad (j = 0, 1, 2, ...).$$

What does $\widehat{\theta}$ estimate? The population counterpart of the sample calculation (2) is

$$\theta_0 = \arg\max_{\theta\in\Theta} E[\ln f(W, \theta)] \tag{6}$$

where W is a random variable with density $g(w)$ so that $E[\ln f(W,\theta)] = \int \ln f(w,\theta)\, g(w)\, dw$.

If the population density $g(w)$ belongs to the $f(w,\theta)$ family, then θ_0 as defined in (6) is the true parameter value. In effect, due to Jensen's inequality:

$$\begin{aligned} &\int \ln f(w,\theta)\, f(w,\theta_0)\, dw - \int \ln f(w,\theta_0)\, f(w,\theta_0)\, dw \\ = &\int \ln\left(\frac{f(w,\theta)}{f(w,\theta_0)}\right) f(w,\theta_0)\, dw \le \ln \int \left(\frac{f(w,\theta)}{f(w,\theta_0)}\right) f(w,\theta_0)\, dw = \ln \int f(w,\theta)\, dw = \ln 1 = 0. \end{aligned}$$

Thus, when $g(w) = f(w,\theta_0)$ we have

$$E[\ln f(W,\theta)] - E[\ln f(W,\theta_0)] \le 0 \qquad \text{for all } \theta.$$

[2] To see that the Hessian is negative semidefinite, note that using the decomposition $\kappa_i'\kappa_i = \Lambda_i(1-\Lambda_i)$ with $\kappa_i' = \left[(1-\Lambda_i)\Lambda_i^{1/2}, \Lambda_i(1-\Lambda_i)^{1/2}\right]$, the Hessian can be written as $\partial^2 L(\theta)/\partial\theta\partial\theta' = -\sum_{i=1}^n x_i \kappa_i' \kappa_i x_i'$.

The value θ_0 can be interpreted more generally as follows:[3]

$$\theta_0 = \arg\min_{\theta\in\Theta} E\left[\ln \frac{g(W)}{f(W,\theta)}\right]. \tag{7}$$

The quantity $E\left[\ln \frac{g(W)}{f(W,\theta)}\right] \equiv \int \ln\left(\frac{g(w)}{f(w,\theta)}\right) g(w)\, dw$ is the Kullback-Leibler divergence (KLD) from $f(w,\theta)$ to $g(w)$. The KLD is the expected log difference between $g(w)$ and $f(w,\theta)$ when the expectation is taken using $g(w)$. Thus, $f(w,\theta_0)$ can be regarded as the best approximation to $g(w)$ in the class $f(w,\theta)$ when the approximation is understood in the KLD sense.

If $g(w) = f(w,\theta_0)$ then θ_0 is called the "true value". If $g(w)$ does not belong to the $f(w,\theta)$ class and $f(w,\theta_0)$ is just the best approximation to $g(w)$ in the KLD sense, then θ_0 is called a "pseudo-true value".

The extent to which a pseudo-true value remains an interesting quantity is model specific. For example, (3) is a restrictive model of the conditional distribution of y_i given x_i, first because it assumes that the dependence of y_i on x_i occurs exclusively through the conditional mean $E(y_i \mid x_i)$ and secondly because this conditional mean is assumed to be a linear function of x_i. However, if y_i depends on x_i in other ways, for example through the conditional variance, the parameter values $\theta_0 = \left(\beta_0', \sigma_0^2\right)'$ remain interpretable quantities: β_0 as $\partial E(y_i \mid x_i)/\partial x_i$ and σ_0^2 as the unconditional variance of the errors $u_i = y_i - x_i'\beta_0$. If $E(y_i \mid x_i)$ is a nonlinear function, β_0 and σ_0^2 can only be characterized as the linear projection regression coefficient vector and the linear projection error variance, respectively.

Pseudo maximum likelihood estimation The statistic $\widehat{\theta}$ is the maximum likelihood estimator under the assumption that $g(w)$ belongs to the $f(w,\theta)$ class. In the absence of this assumption, $\widehat{\theta}$ is a pseudo-maximum likelihood estimator (PML) based on the $f(w,\theta)$ family of densities. Sometimes $\widehat{\theta}$ is called a quasi-maximum likelihood estimator.

2 Consistency and asymptotic normality of PML estimators

Under regularity and identification conditions a PML estimator $\widehat{\theta}$ is a consistent estimator of the (pseudo) true value θ_0. Since $\widehat{\theta}$ may not have a closed form expression we need a method for establishing the consistency of an estimator that maximizes an objective function. The following theorem taken from Newey and McFadden (1994) provides such a method. The requirements are boundedness of the parameter space, uniform convergence of the objective function to some nonstochastic continuous limit, and that the limiting objective function is uniquely maximized at the truth (identification).

Consistency Theorem Suppose that $\widehat{\theta}$ maximizes the objective function $S_n(\theta)$ in the parameter space Θ. Assume the following:

[3] Since $E\left[\ln \frac{g(W)}{f(W,\theta)}\right] = E[\ln g(W)] - [\ln f(W,\theta)]$ and $E[\ln g(W)]$ does not depend on θ, the arg min in (7) is the same as the arg max in (6).

(a) Θ is a compact set.

(b) The function $S_n(\theta)$ converges uniformly in probability to $S_0(\theta)$.

(c) $S_0(\theta)$ is continuous.

(d) $S_0(\theta)$ is uniquely maximized at θ_0.

Then $\widehat{\theta} \xrightarrow{p} \theta_0$.

In the PML context $S_n(\theta) = (1/n)\sum_{i=1}^{n} \ln f(w_i, \theta)$ and $S_0(\theta) = E[\ln f(W, \theta)]$. In particular, in the regression example, by the law of large numbers:[4]

$$S_0(\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}E\left[\left(y_i - x_i'\beta\right)^2\right] \tag{8}$$

and, noting that $y_i - x_i'\beta \equiv u_i - x_i'(\beta - \beta_0)$, also

$$S_0(\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left[\sigma_0^2 + (\beta - \beta_0)' E\left(x_i x_i'\right)(\beta - \beta_0)\right]. \tag{9}$$

In this example, $S_0(\theta)$ is uniquely maximized at θ_0 as long as $E(x_i x_i')$ has full rank.

Asymptotic normality To discuss asymptotic normality, in addition to the conditions required for consistency, we assume that $f(w_i, \theta)$ has first and second derivatives in a neighborhood of θ_0, and θ_0 is an interior point of Θ.[5] For simplicity, use the notation $\ell_i(\theta) = \ln f(w_i, \theta)$ and $q_i(\theta) = \partial \ell_i(\theta)/\partial\theta$. Note that if the data are iid the score $q_i(\theta_0)$ is also iid with zero mean vector and covariance matrix

$$V = E\left(\frac{\partial \ell_i(\theta_0)}{\partial\theta}\frac{\partial \ell_i(\theta_0)}{\partial\theta'}\right). \tag{10}$$

Next, because of the central limit theorem we have

$$\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial\theta} \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell_i(\theta_0)}{\partial\theta} \xrightarrow{d} \mathcal{N}(0, V). \tag{11}$$

As for the Hessian matrix, its convergence follows from the law of large numbers:

$$\frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'} \equiv H_n(\theta_0) \xrightarrow{p} E\left[\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'}\right] \equiv H. \tag{12}$$

We assume that H is a non-singular matrix and that $H_n\left(\widetilde{\theta}\right) \xrightarrow{p} H$ for any $\widetilde{\theta}$ such that $\widetilde{\theta} \xrightarrow{p} \theta_0$.

Now, using the mean value theorem:

$$0 = \frac{\partial L\left(\widehat{\theta}\right)}{\partial\theta_j} = \frac{\partial L(\theta_0)}{\partial\theta_j} + \sum_{\ell=1}^{p}\frac{\partial^2 L\left(\widetilde{\theta}_{[j]}\right)}{\partial\theta_j\partial\theta_\ell}\left(\widehat{\theta}_\ell - \theta_{0\ell}\right) \quad (j = 1, ..., p) \tag{13}$$

[4] Given the equivalence in this case between pointwise and uniform convergence.

[5] We can proceed as if $\widehat{\theta}$ were an interior point of Θ since consistency of $\widehat{\theta}$ for θ_0 and the assumption that θ_0 is interior to Θ implies that the probability that $\widehat{\theta}$ is not interior goes to zero as $n \to \infty$.

where $\widehat{\theta}_\ell$ is the ℓ-th element of $\widehat{\theta}$, and $\widetilde{\theta}_{[j]}$ denotes a $p \times 1$ random vector such that $\left\|\widetilde{\theta}_{[j]} - \theta_0\right\| \leq \left\|\widehat{\theta} - \theta_0\right\|$.[6]

Note that $\widehat{\theta} \xrightarrow{p} \theta_0$ implies $\widetilde{\theta}_{[j]} \xrightarrow{p} \theta_0$ and also

$$\frac{1}{n}\frac{\partial^2 L\left(\widetilde{\theta}_{[j]}\right)}{\partial\theta_j \partial\theta'_\ell} \xrightarrow{p} (j,\ell) \text{ element of } H,$$

which leads to the asymptotic linear representation of the estimation error:[7]

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -H^{-1}\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial\theta} + o_p(1). \tag{14}$$

Finally, using (11) and Cramér's theorem we obtain:

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, W) \tag{15}$$

where $W = H^{-1}VH^{-1}$ or at length:

$$W = \left[E\left(\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'}\right)\right]^{-1} E\left(\frac{\partial \ell_i(\theta_0)}{\partial\theta}\frac{\partial \ell_i(\theta_0)}{\partial\theta'}\right)\left[E\left(\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'}\right)\right]^{-1}. \tag{16}$$

Asymptotic standard errors A consistent estimator of the asymptotic variance matrix W is:

$$\widehat{W} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i\left(\widehat{\theta}\right)}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell_i\left(\widehat{\theta}\right)}{\partial\theta}\frac{\partial \ell_i\left(\widehat{\theta}\right)}{\partial\theta'}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i\left(\widehat{\theta}\right)}{\partial\theta\partial\theta'}\right)^{-1}. \tag{17}$$

3 The information matrix identity

As long as $f(w,\theta)$ is a density function it integrates to one:

$$\int f(w,\theta)\, dw = 1. \tag{18}$$

Taking partial derivatives in (18) with respect to θ we get the zero mean property of the score:

$$\int \frac{\partial \ln f(w,\theta)}{\partial\theta} f(w,\theta)\, dw = 0. \tag{19}$$

Next, taking partial derivatives again:

$$\int \frac{\partial^2 \ln f(w,\theta)}{\partial\theta\partial\theta'} f(w,\theta)\, dw + \int \frac{\partial \ln f(w,\theta)}{\partial\theta}\frac{\partial \ln f(w,\theta)}{\partial\theta'} f(w,\theta)\, dw = 0. \tag{20}$$

Therefore, if $g(w) = f(w,\theta_0)$ we have

$$E\left(\frac{\partial \ell_i(\theta_0)}{\partial\theta}\frac{\partial \ell_i(\theta_0)}{\partial\theta'}\right) = -E\left(\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'}\right). \tag{21}$$

[6] The expansion has to be made element by element since $\widetilde{\theta}_{[j]}$ may be different for each j.

[7] The notation $o_p(1)$ denotes a term that converges to zero in probability.

This result is known as the information matrix identity. It says that when evaluated at θ_0 the covariance matrix of the score coincides with minus the expected Hessian of the log-likelihood function for observation i. It is an identity in the sense of (20), but in general it need not hold if the expectations in (21) are taken with respect to $g(w)$ and $g(w) \neq f(w, \theta_0)$.

The implication is that under correct specification $V = -H$ in the sandwich formula (16) and therefore:

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[I(\theta_0)\right]^{-1}\right) \tag{22}$$

where

$$I(\theta_0) = -E\left(\frac{\partial^2 \ell_i(\theta_0)}{\partial\theta\partial\theta'}\right) \equiv -\operatorname*{plim}_{n\to\infty} \frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'}. \tag{23}$$

The matrix $I(\theta_0)$ is known as the information matrix or the Fisher information, after the work of Ronald Fisher. It is called information because it can be regarded as a measure of the amount of information that the random variable W with density $f(w, \theta)$ contains about the unknown parameter θ. Intuitively, the greater the expected curvature of the log likelihood at $\theta = \theta_0$ the greater the information and the smaller the asymptotic variance of the maximum likelihood estimator, which is given by the inverse of the information matrix.

The asymptotic Cramér-Rao inequality Under suitable regularity conditions, the matrix $\left[I(\theta_0)\right]^{-1}$ is a lower bound for the asymptotic covariance matrix of any consistent estimator of θ_0. Furthermore, this lower bound is attained by the maximum likelihood estimator.

Estimating the information matrix There are a variety of consistent estimators of $I(\theta_0)$. One possibility is to use the observed Hessian evaluated at $\widehat{\theta}$:

$$\widehat{I} = -\frac{1}{n}\frac{\partial^2 L\left(\widehat{\theta}\right)}{\partial\theta\partial\theta'}. \tag{24}$$

Another possibility is the expected Hessian evaluated at $\widehat{\theta}$, as long as its functional form is known:

$$\widetilde{I} = I\left(\widehat{\theta}\right), \tag{25}$$

Yet another possibility in a conditional likelihood model $f(y \mid x; \theta)$ is to use a sample average of the expected Hessian conditioned on x_i and evaluated at $\widehat{\theta}$:

$$\widetilde{\widetilde{I}} = -\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial^2 \ln f\left(y \mid x_i; \widehat{\theta}\right)}{\partial\theta\partial\theta'} \mid x_i\right). \tag{26}$$

Finally, one can use the variance of the score-form of the information matrix to obtain an estimate:

$$\widehat{\widehat{I}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ell_i\left(\widehat{\theta}\right)}{\partial\theta}\frac{\partial \ell_i\left(\widehat{\theta}\right)}{\partial\theta'}\right). \tag{27}$$

4 Example: Normal linear regression

Letting $u_i = y_i - x_i'\beta_0$, the score at the true value for the log-likelihood function in (4) is

$$\frac{\partial \ell_i(\theta_0)}{\partial \theta} = \frac{1}{\sigma_0^2}\begin{pmatrix} x_i u_i \\ \frac{1}{2\sigma_0^2}\left(u_i^2 - \sigma_0^2\right) \end{pmatrix}.$$

The covariance matrix of the score is

$$V = E\begin{pmatrix} \frac{1}{\sigma_0^4}u_i^2 x_i x_i' & \frac{1}{2\sigma_0^6}x_i u_i\left(u_i^2-\sigma_0^2\right) \\ \frac{1}{2\sigma_0^6}x_i' u_i\left(u_i^2-\sigma_0^2\right) & \frac{1}{4\sigma_0^8}\left(u_i^2-\sigma_0^2\right)^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_0^4}E\left(u_i^2 x_i x_i'\right) & \frac{1}{2\sigma_0^6}E\left(u_i^3 x_i\right) \\ \frac{1}{2\sigma_0^6}E\left(u_i^3 x_i'\right) & \frac{1}{4\sigma_0^8}\left[E\left(u_i^4\right)-\sigma_0^4\right] \end{pmatrix}.$$

The expected Hessian is

$$H = E\begin{pmatrix} -\frac{1}{\sigma_0^2}x_i x_i' & -\frac{1}{\sigma_0^4}x_i u_i \\ -\frac{1}{\sigma_0^4}x_i' u_i & -\frac{1}{2\sigma_0^4} - \left(u_i^2-\sigma_0^2\right)\frac{1}{\sigma_0^6} \end{pmatrix} = -\begin{pmatrix} \frac{1}{\sigma_0^2}E\left(x_i x_i'\right) & 0 \\ 0' & \frac{1}{2\sigma_0^4} \end{pmatrix}.$$

The sandwich formula in (16) is:

$$W = \begin{pmatrix} \sigma_0^2\left[E\left(x_i x_i'\right)\right]^{-1} & 0 \\ 0' & 2\sigma_0^4 \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_0^4}E\left(u_i^2 x_i x_i'\right) & \frac{1}{2\sigma_0^6}E\left(u_i^3 x_i\right) \\ \frac{1}{2\sigma_0^6}E\left(u_i^3 x_i'\right) & \frac{1}{4\sigma_0^8}\left[E\left(u_i^2\right)-\sigma_0^4\right] \end{pmatrix}\begin{pmatrix} \sigma_0^2\left[E\left(x_i x_i'\right)\right]^{-1} & 0 \\ 0' & 2\sigma_0^4 \end{pmatrix}.$$

Note that under misspecification the information matrix identity does not hold since $V \neq -H$. The first block-diagonal components of V and $-H$ will coincide under conditional homoskedasticity, that is, if $E\left(u_i^2 \mid x\right) = \sigma_0^2$. The off-diagonal block of V is zero under conditional symmetry, that is, if $E\left(u_i^3 \mid x\right) = 0$. Lastly, the second block-diagonal terms of V and $-H$ will coincide under the normal kurtosis condition $E\left(u_i^4\right) = 3\sigma_0^4$. These conditions are satisfied when model (4) is correctly specified but not in general.

Under correct specification:

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left[0, \begin{pmatrix} \sigma_0^2\left[E\left(x_i x_i'\right)\right]^{-1} & 0 \\ 0' & 2\sigma_0^4 \end{pmatrix}\right]. \tag{28}$$

5 Estimation subject to constraints

We may wish to estimate parameters subject to constraints. Sometimes one seeks to ensure internal consistency in a model that is required to answer a question of interest such as a welfare calculation, for example enforcing symmetry of cross-price elasticities in a demand system. Another common situation is an interest in the value or adequacy of economic restrictions, such as constant returns to scale in a production function. Finally, one may be simply willing to consider a restricted version of a model as a way of producing a simpler or tighter summary of data.

Constraints on parameters may be expressed as equation restrictions

$$h(\theta_0) = 0 \tag{29}$$

where $h(\theta)$ is a vector of r restrictions: $h(\theta) = (h_1(\theta), ..., h_r(\theta))'$. Alternatively, constrains may be expressed in parametric form

$$\theta_0 = \theta(\alpha_0) \tag{30}$$

whereby θ is functionally related to a free parameter vector α of smaller dimension than θ such that the number of restrictions is $r = \dim(\theta) - \dim(\alpha)$. Depending on the problem it may be more convenient to express restrictions in one way or the other (or in mixed form).

One way of obtaining a constrained estimator of θ_0 is to maximize the log-likelihood function $L(\theta)$ subject to the constraints $h(\theta) = 0$:

$$\left(\widetilde{\theta}, \widetilde{\lambda}\right) = \arg\max_{\theta,\lambda} \left[\frac{1}{n} L(\theta) - \lambda' h(\theta)\right] \tag{31}$$

where λ is an $r \times 1$ vector of Lagrange multipliers.

Alternatively, if the restrictions have been parameterized as in (30), the log-likelihood function can be maximized with respect to α as in an unrestricted problem:

$$\widetilde{\alpha} = \arg\max_{\alpha} L[\theta(\alpha)]. \tag{32}$$

Restricted estimates of θ_0 are then given by $\widetilde{\theta} = \theta(\widetilde{\alpha})$.

Asymptotic normality of constrained estimators The asymptotic variance of $\sqrt{n}(\widetilde{\alpha} - \alpha_0)$ can be obtained as an application of the result in (15)-(17) to the log-likelihood $L^*(\alpha) = L[\theta(\alpha)]$. Letting $G = G(\alpha_0)$ where $G(\alpha) = \partial\theta(\alpha)/\partial\alpha'$, using the chain rule we have

$$\frac{\partial L^*(\alpha_0)}{\partial \alpha} = G' \frac{\partial L(\theta_0)}{\partial \theta}, \tag{33}$$

and[8]

$$E\left(\frac{\partial^2 L^*(\alpha_0)}{\partial \alpha \partial \alpha'}\right) = G'HG. \tag{34}$$

Moreover,

$$\frac{1}{\sqrt{n}} \frac{\partial L^*(\alpha_0)}{\partial \alpha} \xrightarrow{d} \mathcal{N}\left(0, G'VG\right). \tag{35}$$

Therefore,

$$\sqrt{n}(\widetilde{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}\left(0, \left(G'HG\right)^{-1} G'VG \left(G'HG\right)^{-1}\right). \tag{36}$$

[8] The matrix $\frac{\partial^2 L^*(\alpha_0)}{\partial \alpha \partial \alpha'}$ contains an additional term, which is equal to zero in expectation.

The asymptotic distribution of $\widetilde{\theta}$ then follows from the delta method:

$$\sqrt{n}\left(\widetilde{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, W_R\right) \tag{37}$$

where W_R is a matrix of reduced rank given by

$$W_R = G\left(G'HG\right)^{-1} G'VG\left(G'HG\right)^{-1} G'. \tag{38}$$

Under correct specification $V=-H$, so that the previous results become

$$\sqrt{n}\left(\widetilde{\alpha}-\alpha_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G'I\left(\theta_0\right)G\right)^{-1}\right). \tag{39}$$

and

$$\sqrt{n}\left(\widetilde{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, G\left(G'I\left(\theta_0\right)G\right)^{-1}G'\right). \tag{40}$$

Under correct specification, $\widetilde{\theta}$ is asymptotically more efficient than $\widehat{\theta}$ since the difference between their asymptotic variance matrices is positive-semidefinite:

$$\left[I\left(\theta_0\right)\right]^{-1} - G\left(G'I\left(\theta_0\right)G\right)^{-1}G' = A^{-1}\left[I - G^*\left(G^{*\prime}G^*\right)^{-1}G^{*\prime}\right]A^{-1\prime} \geq 0 \tag{41}$$

where $I\left(\theta_0\right) = A'A$ and $G^* = AG$. However, under misspecification the sandwich matrix $W = H^{-1}VH^{-1}$ and W_R in (38) cannot be ordered.

The likelihood ratio statistic By construction an unrestricted maximum is greater than a restricted one: $L\left(\widehat{\theta}\right) \geq L\left(\widetilde{\theta}\right)$. It seems therefore natural to look at the log-likelihood change $L\left(\widehat{\theta}\right) - L\left(\widetilde{\theta}\right)$ as a measure of the cost of imposing the restrictions. It can be shown that under correct specification (or if the information matrix identity holds) and the r restrictions on θ_0 also hold then

$$LR = 2\left[L\left(\widehat{\theta}\right) - L\left(\widetilde{\theta}\right)\right] \xrightarrow{d} \chi^2_r. \tag{42}$$

This result is used to construct a large-sample test of the restrictions: the rule is to reject the restrictions if $LR > \kappa$ where κ is the $(1-\alpha)$-quantile of the χ^2_r distribution and α is the chosen size of the test.

The Wald statistic We can also learn about the adequacy of the restrictions by examining how far the unrestricted estimates are from satisfying the constraints. We are thus led to look at unrestricted estimates of the constraints given by $h\left(\widehat{\theta}\right)$.

Letting $D = D\left(\theta_0\right)$ and $\widehat{D} = D\left(\widehat{\theta}\right)$ where $D\left(\theta\right) = \partial h\left(\theta\right)/\partial\theta'$, under the assumption that $h\left(\theta_0\right) = 0$, from the delta method we have

$$\sqrt{n}h\left(\widehat{\theta}\right) \xrightarrow{d} \mathcal{N}\left(0, DWD'\right) \tag{43}$$

and

$$\mathcal{W}_R = n\left[h\left(\widehat{\theta}\right)'\left(\widehat{D}\widehat{W}\widehat{D}'\right)^{-1}h\left(\widehat{\theta}\right)\right] \xrightarrow{d} \chi^2_r. \tag{44}$$

The quantity $\mathcal{W}_R$ is a Wald statistic. Like the LR statistic it can be used to construct a large-sample test of the restrictions with a similar rejection region. However, while the calculation of LR requires both $\widehat{\theta}$ and $\widetilde{\theta}$, the calculation of $\mathcal{W}_R$ only requires the unrestricted estimate $\widehat{\theta}$.

Another difference is that, contrary to LR, $\mathcal{W}_R$ still has a large-sample chi-square distribution under misspecification if it relies on a robust estimate of the variance of $\widehat{\theta}$ as in (44). A Wald statistic that is directly comparable to the LR statistic would be a non-robust version of the form:

$$\mathcal{W} = n\left[h\left(\widehat{\theta}\right)'\left(\widehat{D}\left[I\left(\widehat{\theta}\right)\right]^{-1}\widehat{D}'\right)^{-1}h\left(\widehat{\theta}\right)\right]. \tag{45}$$

The Lagrange Multiplier statistic Another angle on the cost of imposing the restrictions is to examine how far the estimated Lagrange multiplier vector $\widetilde{\lambda}$ is from zero. The first-order conditions from the optimization problem in (31) are:

$$\frac{1}{n}\frac{\partial L\left(\widetilde{\theta}\right)}{\partial\theta} = D\left(\widetilde{\theta}\right)'\widetilde{\lambda} \tag{46}$$

$$h\left(\widetilde{\theta}\right) = 0, \tag{47}$$

which lead to the asymptotic linear representation of $\widetilde{\lambda}$:[9]

$$\sqrt{n}\widetilde{\lambda} = \left(DH^{-1}D'\right)^{-1}DH^{-1}\frac{1}{\sqrt{n}}\frac{\partial L\left(\theta_0\right)}{\partial\theta} + o_p\left(1\right) \tag{48}$$

and

$$\sqrt{n}\widetilde{\lambda} \xrightarrow{d} \mathcal{N}\left[0, \left(DH^{-1}D'\right)^{-1}DH^{-1}VH^{-1}D'\left(DH^{-1}D'\right)^{-1}\right], \tag{49}$$

and also

$$LM_R = n\widetilde{\lambda}'\widetilde{D}\widetilde{H}^{-1}\widetilde{D}'\left(\widetilde{D}\widetilde{H}^{-1}\widetilde{V}\widetilde{H}^{-1}\widetilde{D}'\right)^{-1}\widetilde{D}\widetilde{H}^{-1}\widetilde{D}'\widetilde{\lambda} \xrightarrow{d} \chi^2_r \tag{50}$$

[9] Using the mean value theorem for each component of the first-order conditions

$$\begin{aligned} D\left(\widetilde{\theta}\right)'\widetilde{\lambda} &= \frac{1}{n}\frac{\partial L\left(\theta_0\right)}{\partial\theta} + \frac{1}{n}\frac{\partial^2 L\left(\theta^*\right)}{\partial\theta\partial\theta'}\left(\widetilde{\theta}-\theta_0\right) \\ 0 &= h\left(\widetilde{\theta}\right) = h\left(\theta_0\right) + D\left(\theta^{**}\right)\left(\widetilde{\theta}-\theta_0\right) \end{aligned}$$

and combining the two expressions using that $h\left(\theta_0\right)=0$ we get:

$$D\left(\theta^{**}\right)\left(\frac{1}{n}\frac{\partial^2 L\left(\theta^*\right)}{\partial\theta\partial\theta'}\right)^{-1}D\left(\widetilde{\theta}\right)'\widetilde{\lambda} = D\left(\theta^{**}\right)\left(\frac{1}{n}\frac{\partial^2 L\left(\theta^*\right)}{\partial\theta\partial\theta'}\right)^{-1}\frac{1}{n}\frac{\partial L\left(\theta_0\right)}{\partial\theta}.$$

where

$$\widetilde{D} = D\left(\widetilde{\theta}\right) \quad \widetilde{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i\left(\widetilde{\theta}\right)}{\partial\theta\partial\theta'} \quad \widetilde{V} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell_i\left(\widetilde{\theta}\right)}{\partial\theta}\frac{\partial \ell_i\left(\widetilde{\theta}\right)}{\partial\theta'}.$$

The quantity LM_R is a Lagrange Multiplier statistic. Like the Wald statistic in (44) it can be used to construct a large-sample test of the restrictions, which remains valid if the objective function is only a pseudo likelihood function and it does not satisfy the information matrix identity.

In view of (46) we can replace $\widetilde{D}'\widetilde{\lambda}$ in (50) with $n^{-1}\partial L\left(\widetilde{\theta}\right)/\partial\theta$, which produces the score form of the statistic. The score function is exactly equal to zero when evaluated at $\widehat{\theta}$, but not when evaluated at $\widetilde{\theta}$. If the constraints are true, we would expect both $n^{-1}\partial L\left(\widetilde{\theta}\right)/\partial\theta$ and $\widetilde{\lambda}$ to be small quantities, so that the rejection region of the null hypothesis $h(\theta_0) = 0$ is associated with large values of LM_R.

Under correct specification $V = -H$, we get

$$\sqrt{n}\widetilde{\lambda} \xrightarrow{d} \mathcal{N}\left[0, \left(D\left[I\left(\theta_0\right)\right]^{-1}D'\right)^{-1}\right] \tag{51}$$

and

$$LM = n\widetilde{\lambda}'\widetilde{D}\left[I\left(\widetilde{\theta}\right)\right]^{-1}\widetilde{D}'\widetilde{\lambda} \equiv \frac{1}{n}\frac{\partial L\left(\widetilde{\theta}\right)}{\partial\theta'}\left[I\left(\widetilde{\theta}\right)\right]^{-1}\frac{\partial L\left(\widetilde{\theta}\right)}{\partial\theta} \xrightarrow{d} \chi^2_r. \tag{52}$$

The statistic LM is a non-robust version of the Lagrange Multiplier statistic that is directly comparable to the LR statistic.

6 Example: LR, Wald and LM in the normal linear regression model

The model is the same as in (4). We consider the partitions $X = (X_1, X_2)$ and $\beta = \left(\beta_1', \beta_2'\right)'$ where X_1 is of order $n \times r$, X_2 is $n \times (k-r)$, β_1 is $r \times 1$ and β_2 is $(k-r) \times 1$, and the r restrictions

$$\beta_1 = 0. \tag{53}$$

The unrestricted estimates are $\widehat{\theta} = (\widehat{\beta}', \widehat{\sigma}^2)'$ as in (5), whereas the restricted estimates are

$$\widetilde{\beta}_1 = 0, \quad \widetilde{\beta}_2 = \left(X_2'X_2\right)^{-1}X_2'y, \quad \widetilde{\sigma}^2 = \frac{\widetilde{u}'\widetilde{u}}{n} \tag{54}$$

where $\widetilde{u} = y - X_2\widetilde{\beta}_2$.

The LR statistic is given by

$$LR = n\ln\left(\frac{\widetilde{u}'\widetilde{u}}{\widehat{u}'\widehat{u}}\right). \tag{55}$$

If we modify the example to assume that σ^2 is known, $\widehat{\beta}$ and $\widetilde{\beta}$ remain unchanged but in this case

$$LR_\sigma = \frac{\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u}}{\sigma^2}, \tag{56}$$

which is exactly distributed as χ^2_r under normality.

Turning to the Wald statistic, recall that

$$\sqrt{n}\left(\widehat{\beta}-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\operatorname{plim}\left(X'X/n\right)^{-1}\right)$$

and introduce the partition

$$\left(X'X\right)^{-1}=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}. \tag{57}$$

Thus, if the restrictions hold

$$\sqrt{n}\widehat{\beta}_1\xrightarrow{d}\mathcal{N}\left[0,\sigma^2\operatorname{plim}\left(nA_{11}\right)\right] \tag{58}$$

and therefore the (non-robust) Wald statistic is given by:

$$W=\frac{\widehat{\beta}_1'A_{11}^{-1}\widehat{\beta}_1}{\widehat{\sigma}^2}. \tag{59}$$

It can be shown that $\widehat{\beta}_1'A_{11}^{-1}\widehat{\beta}_1=\widetilde{u}'\widetilde{u}-\widehat{u}'\widehat{u}$,[10] so that also

$$W=\frac{\widetilde{u}'\widetilde{u}-\widehat{u}'\widehat{u}}{\widehat{\sigma}^2}. \tag{60}$$

Moreover, if σ^2 is known

$$W_\sigma=\frac{\widetilde{u}'\widetilde{u}-\widehat{u}'\widehat{u}}{\sigma^2},$$

so that $W_\sigma=LR_\sigma$. With unknown σ^2 it can be shown that $W\geq LR$ even if both statistics have the same asymptotic distribution.

Finally, turning to the LM statistic, the components of the score are:

$$\begin{array}{l}\frac{\partial L(\theta)}{\partial\beta_1}=\frac{1}{\sigma^2}X_1'\left(y-X\beta\right)\\ \frac{\partial L(\theta)}{\partial\beta_2}=\frac{1}{\sigma^2}X_2'\left(y-X\beta\right)\\ \frac{\partial L(\theta)}{\partial\sigma^2}=\frac{n}{2\sigma^4}\left[\frac{1}{n}\left(y-X\beta\right)'\left(y-X\beta\right)-\sigma^2\right],\end{array} \tag{61}$$

which once evaluated at restricted estimates are

$$\begin{array}{l}\frac{\partial L(\widetilde{\theta})}{\partial\beta_1}=\frac{1}{\widetilde{\sigma}^2}X_1'\widetilde{u}\neq 0\\ \frac{\partial L(\widetilde{\theta})}{\partial\beta_2}=\frac{1}{\widetilde{\sigma}^2}X_2'\widetilde{u}=0\\ \frac{\partial L(\widetilde{\theta})}{\partial\sigma^2}=\frac{n}{2\widetilde{\sigma}^4}\left(\frac{1}{n}\widetilde{u}'\widetilde{u}-\widetilde{\sigma}^2\right)=0.\end{array} \tag{62}$$

[10] Using the partitioned inverse matrix result $A_{11}^{-1}=X_1'\left(I-M_2\right)X_1$ and $MM_2=M_2$ where $M_2=X_2\left(X_2'X_2\right)^{-1}X_2'$ and $M=X\left(X'X\right)^{-1}X'$. Premultiplying $y=X\widehat{\beta}+\widehat{u}$ by $\left(I-M_2\right)$ and taking squares we get $y'\left(I-M_2\right)y=\widehat{\beta}_1'X_1'\left(I-M_2\right)X_1\widehat{\beta}_1+\widehat{u}'\left(I-M_2\right)\widehat{u}$, which equals $\widetilde{u}'\widetilde{u}=\widehat{\beta}_1'A_{11}^{-1}\widehat{\beta}_1+\widehat{u}'\widehat{u}$.

Moreover, the information matrix in this case is

$$I(\theta_0) = \begin{pmatrix} \frac{1}{\sigma_0^2} \operatorname{plim}\left(\frac{X'X}{n}\right) & 0 \\ 0' & \frac{1}{2\sigma_0^4} \end{pmatrix}, \tag{63}$$

so that using

$$I\left(\widetilde{\theta}\right) = \begin{pmatrix} \frac{1}{\widetilde{\sigma}^2}\left(\frac{X'X}{n}\right) & 0 \\ 0' & \frac{1}{2\widetilde{\sigma}^4} \end{pmatrix}, \tag{64}$$

the LM statistic becomes

$$LM = \begin{pmatrix} \frac{\widetilde{u}'X_1}{\widetilde{\sigma}^2} & 0' & 0 \end{pmatrix} \begin{pmatrix} \widetilde{\sigma}^2 \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} & 0 \\ 0' \quad 0 & 2\widetilde{\sigma}^4/n \end{pmatrix} \begin{pmatrix} \frac{X_1'\widetilde{u}}{\widetilde{\sigma}^2} \\ 0 \\ 0 \end{pmatrix} \tag{65}$$

and

$$LM = \frac{\widetilde{u}'X_1 A_{11} X_1'\widetilde{u}}{\widetilde{\sigma}^2}. \tag{66}$$

It can be shown that

$$\widetilde{u}'X_1 A_{11} X_1'\widetilde{u} = \widehat{\beta}_1' A_{11}^{-1}\widehat{\beta}_1 = \widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u} \tag{67}$$

so that also

$$LR = \frac{\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u}}{\widetilde{\sigma}^2}. \tag{68}$$

Moreover, if σ^2 is known

$$LM_\sigma = \frac{\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u}}{\sigma^2}$$

so that $LM_\sigma = W_\sigma = LR_\sigma$. With unknown σ^2 it is easy to show that $W \geq LR \geq LM$.