EC771: Econometrics, Spring 2004

Greene, Econometric Analysis (5th ed, 2003)

Chapter 17: Maximum Likelihood Estimation

The preferred estimator in a wide variety of econometric settings is that derived from the principle of maximum likelihood (MLE). The pdf for a random variable y, conditioned on a set of parameters θ, is denoted $f(y|\theta)$. This function identifies the data generating process (DGP) that underlies an observed sample of data, and provides a mathematical description of the data that the DGP will produce. That is, for a well–specified DGP, we could generate any desired quantity of artificial data whose properties correspond to that DGP. The joint

density of n independently and identically distributed ($i.i.d.$) observations from that process is the product of the individual densities:

$$f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y).$$

This joint density, the likelihood function $L(\theta|y)$, is a function of the unknown parameter vector θ and the collection of sample data y. It is usually easier to work with its logarithm,

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta).$$

Of course, we need not consider a set of y_i that are $i.i.d.$ to use the MLE framework. In a simple regression framework, $y_i = x_i'\beta + \epsilon_i$. Assume that ϵ is normally distributed. Then conditioned on a specific x_i, y_i is distributed with $\mu_i = x_i'\beta$ and variance σ_ϵ^2. Nevertheless, normality of the observations implies that they are independently distributed, as we may write

the loglikelihood function of θ conditioned on y, X:

$$\ln L(\theta|y, X) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} [\ln \sigma^2 + \ln(2\pi) + (y_i - x_i'\beta)^2/\sigma^2],$$

where X is the $n \times K$ matrix of data with its i^{th} row equal to x_i'. Since the first-order conditions for maximization of this function with respect to β will not be affected by the first two terms in this expression, they are often ignored in the computation.

Before we consider the mechanics of MLE, we must discuss the *identifiability* of a particular model. Suppose that we had an infinitely large sample. Could we uniquely determine the values of the parameters θ from the information in that sample? We will not always be able to do so, in which case the underlying model is said

to be *unidentified*. We say that a parameter vector θ is *identified*, or *estimable*, if for any other parameter vector $\theta^* \neq \theta$ and some data y, $L(\theta^*|y) \neq L(\theta|y)$. For instance, in a probit equation, where we observe a binary outcome and hypothesize that the probability of that event (e.g. purchase of an automobile) is related to a causal factor x:

$$Prob(purchase|\beta_1, \beta_2, \sigma, x_i) = Prob(y_i > 0|\beta_1, \beta_2, \sigma, x_i)$$
$$= Prob(\epsilon_i/\sigma > -(\beta_1 + \beta_2 x_i)/\sigma|\beta_1, \beta_2, \sigma, x_i).$$

In this case y_i may be considered the difference between the amount a buyer is willing to pay and the price of the car: a latent variable. If that difference is positive, we observe purchase, and vice versa. Since multiplying $(\beta_1, \beta_2, \sigma)$ by the same nonzero constant leaves the left-hand-side of this expression unchanged, the model's parameter vector is not identified. We must apply a normalization: in

this case, we conventionally set $\sigma = \sigma^2 = 1$ in order to "tie down" the model's parameters, so that we may then uniquely identify the βs.

The principle of maximum likelihood provides a means of choosing an asymptotically efficient estimator for a set of parameters. Let us consider first how this might be done for a discrete pdf such as the Poisson distribution:

$$f(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}.$$

This is the density for each observation. The corresponding loglikelihood function (LLF) for the sample of n observations will be:

$$\ln L(\theta|y) = -n\theta + \ln\theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!)$$

To perform maximum likelihood estimation in Stata, you must code the likelihood function as a program, or `ado-file`. In the simplest form

of such a program, in a setting where the observations are considered to be independently distributed, you need only express the LLF for a single observation, and Stata will add it up over the sample. This is known in Stata terms as an `lf` (linear–form) estimator, and it is applicable to a number of estimation problems that satisfy the so–called linear form restrictions.

This is a different use of 'linear' than that of 'linear regression' or 'linear model', in that a linear–form MLE problem can involve nonlinearities (such as a binomial probit model). The linear–form restrictions require that the log–likelihood contributions can be calculated separately for each observation, and that the sum of the individual contributions equals the overall log likelihood (Gould, Pitblado and Sribney, 2003, p.30). More complicated problems require use of Stata's `d0` form, in which the entire LLF is coded (or its relatives, `d1` and `d2`, in

which first or first and second analytic derivatives are also provided). For the `lf` model, no analytic derivatives are required, and the optimization is more accurate and computationally efficient than that derived from the `d0` form, since differentiation of the loglikelihood function may be done with respect to the linear form rather than with respect to each of its elements.

Appendix 1 contains the logfile for a simple MLE of the Poisson problem described on p. 471 of the text. The Stata program `fishy1_ml.ado` expresses the likelihood of a single observation y_i given the parameter θ. Note that to avoid numerical overflow with the factorial function, we use Stata's `lnfact` function. The `ml model` statement sets up the maximum likelihood problem; we use `ml check` to see whether any obvious errors have been made in the routine. The `ml maximize` statement instructs Stata to

perform the estimation. Following estimation, just as with any estimation command, we may use the `test` command to perform a Wald test of any linear hypothesis, or the `testnl` command to perform a Wald–type test of a nonlinear hypothesis via the delta method.

The probability of observing the given sample of data is not exact when we switch to a continuous distribution, since a particular sample has probability zero; nevertheless, the principle is the same. The values of the parameters that maximize the LF or LLF are the maximum likelihood estimates, $\widehat{\theta}$, and the necessary conditions for maximizing the LLF are

$$\frac{\partial \ln L(\theta|data)}{\partial \theta} = 0.$$

That is, each element of the gradient or score vector must be approximately zero at the optimum (or the norm of that vector must be appropriately close to zero).

Let us consider how the MLE for the mean and variance of a normally distributed random variable may be computed. The LLF for the univariate normal distribution is

$$\ln L(\mu, \sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}[\frac{(y_i - \mu)^2}{\sigma^2}].$$

To maximize this function, we differentiate with respect to both μ and σ and solve the resulting FOCs:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0.$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0.$$

Solution of these two equations will lead to the familiar formula for the mean, and the MLE of σ^2, which is the biased estimator

$$\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

Appendix 2 illustrates MLE of the parameters of a Normal distribution, where we have generated the artificial data with μ=5 and σ=2, and want to recover those parameters. The Stata `ado-file` is a `lf`-form function here as well, expressing the likelihood of a single observation in terms of the two unknown parameters. Note that the results of estimation give us confidence intervals for both μ and σ.

The beauty of the linear–form model is that a whole variety of other DGPs may be expressed in this same context without altering the program. The example in Appendix 2 considers a univariate estimation problem. What if, instead, we consider a model:

$$y_t = \alpha + \beta t + \epsilon_t,$$

that is, a linear trend model? We then want to estimate (by means of MLE, although we could surely use a least squares regression to solve

this problem) those two parameters as well as σ, the standard error of the ϵ process. In this case, the LLF above is modified to include μ_i, the conditional mean of $y|x$, rather than the scalar parameter μ. Appendix 3 presents the results of that estimation. Since the "explanatory variable" t enters the conditional mean equation in linear form, we may merely change the `ml model` statement to express y as a linear function of x (by default with a constant term, α above). Notice that since we are now working with a bivariate relationship, Stata computes a Wald χ^2 statistic for this model which compares it with the "naive model" in which the coefficient of x is zero (that is, a model in which the time trend plays no role).

Finally, we may also model the standard error, rather than assuming it to be common across observations. We may express ϵ_t in this model as not being uniformly distributed, but

rather possessing a standard error, $\gamma = \sigma x_t$: an instance where the relationship is becoming less precise over time (in the implementation, we demean x so that the multiplicative factor is mean zero). This assumption on the error process does not violate the independence of the errors (their second moments are systematically related to t). In regression terms, it is a heteroskedastic regression model, in which we explicitly model the form of the heteroskedasticity. Appendix 4 presents the estimates of that model: again, we need not modify the program, but merely change the `ml model` statement so that we may indicate that the second "equation" being estimated (that for the standard error of ϵ) is non–trivial.

Properties of MLEs

MLEs are most attractive due to their large–sample or asymptotic properties; their finite

sample properties may be suboptimal. For example, the MLE of σ^2 in a standard regression problem involves a divisor of N, rather than $N-1$, so that the MLE is biased downward. In large–sample terms, of course, this does not matter, since $plim(\frac{N-1}{N}) = 1$. To discuss their properties, let us define $\widehat{\theta}$ as the MLE, θ_0 as the true parameter vector, and θ as an arbitrary parameter vector (not necessarily either of the above). Under regularity conditions to be discussed, the MLE has the following asymptotic properties:

- **Consistency**: $plim\,\widehat{\theta} = \theta_0$.

- **Asymptotic normality**: $\widehat{\theta} \xrightarrow{a} N[\theta_0, \{I(\theta_0)\}^{-1}]$, where $I(\theta_0) = -E_0[\partial^2 \ln L/\partial\theta_0\partial\theta_0']$

- **Asymptotic effficiency**: $\widehat{\theta}$ is asy. efficient and achieves the Cramér–Rao Lower Bound (CRLB) for consistent estimators.

- **Invariance**: The MLE of $\gamma_0 = c(\theta_0)$ is $c(\widehat{\theta})$ if $c(\theta_0)$ is a continuous and continuously differentiable function.

The regularity conditions, which we will not further discuss, require that the first three derivatives of $\ln f(y_i|\theta)$ with respect to θ are continuous and finite for almost all y_i and all θ. This condition guarantees the existence of a certain Taylor series approximation. Furthermore, the conditions necessary to obtain the expectations of the first and second derivatives of $\ln f(y_i|\theta)$ must be met, and the third derivative must be bounded, permitting that Taylor series to be truncated. With these conditions met, we may define the moments of the derivatives of the log–likelihood function:

- 1. $\ln f(y_i|\theta)$, $g_i = \partial \ln f(y_i|\theta)/\partial\theta$ and $H_i = \partial^2 \ln f(y_i|\theta)/\partial\theta\partial\theta'$, $i = 1 \ldots n$ are all random samples of random variables.

- 2. $E_0[g_i(\theta_0)] = 0$.

- 3. $Var[g_i(\theta_0)] = -E[H_i(\theta_0)]$.

The first definition follows from the definition of the likelihood function. The second defines the moment condition, or **score vector**, by which the MLE may locate the optimum, and indicates that the MLE is one of a more general class of Generalized Method of Moments (GMM) estimators. Intuitively, this condition indicates that the gradient of the likelihood function must be zero at the optimum, requiring that the first–order conditions for a maximum be satisfied. The third condition provides the Information Matrix Equality, which defines the asymptotic covariance matrix of the MLE as being related to the expectation of the Hessian of the log-likelihood. The Hessian

must be negative (semi-)definite for a maximum, corresponding to the notion that the asymptotic covariance matrix must be positive (semi-)definite.

The expected value of the log–likelihood is maximized at the true value of the parameters, which implies that the MLE will be consistent. To demonstrate the asymptotic normality of the MLE, expand the first–order conditions $g(\widehat{\theta}) = 0$ in a second–order Taylor series around the true parameters:

$$g(\widehat{\theta}) = g(\theta_0) + H(\bar{\theta})(\widehat{\theta} - \theta_0) = 0.$$

The Hessian is evaluated at a point $\bar{\theta}$ that is between $\widehat{\theta}$ and θ_0. Rearrange this function and multiply by $\sqrt{n}$ to obtain:

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \left[-H(\bar{\theta})\right]^{-1} \left[\sqrt{n}\, g(\theta_0)\right].$$

Since $plim(\widehat{\theta} - \theta_0) = 0$, $plim(\widehat{\theta} - \bar{\theta}) = 0$ as well, and (dividing the derivatives by n):

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \left[-\frac{1}{n} H(\theta_0)\right]^{-1} \left[\sqrt{n}\, \bar{g}(\theta_0)\right].$$

Since $[\sqrt{n}\,\bar{g}(\theta_0)]$ is $\sqrt{n}$ times the mean of a random sample, we find the limiting variance of that expression, and can express its distribution as:

$$\left[\sqrt{n}\,\bar{g}(\theta_0)\right] \xrightarrow{d} N\left[0, -E_0\left[\frac{1}{n}H(\theta_0)\right]\right].$$

Combining results, we may derive the expression above for asymptotic normality of the MLE, with a covariance matrix equal to the inverse of the **information matrix**: minus the expectation of the mean of the Hessian, evaluated at the true parameter vector.

For the normal distribution, the second derivatives of the LLF are:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{-n}{\sigma^2},$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2,$$

$$\frac{\partial \ln L}{\partial \mu \partial \sigma^2} = -\frac{-1}{\sigma^4} \sum_{i=1}^{n} (y_i - \mu).$$

The expectation of the cross–partial is zero, since $E[x_i] = \mu$. The first expression is non-stochastic, while the second has expectation $\frac{-n}{2\sigma^4}$ since each of the n terms has expected value σ^2. Collecting these expectations in the (diagonal) information matrix, reversing the sign and inverting, we derive the asymptotic covariance matrix for the MLE:

$$\left(-E_0 \left[\frac{\partial^2 \ln L}{\partial \theta_0 \partial \theta_0'} \right] \right)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}.$$

The elements of this covariance matrix may then be calculated, following the successful maximization of the LLF, replacing the unknown parameter σ^2 with its point estimate. Generally, though, the form of the expected values of the second derivatives of the LLF may be unknown, as it will be a complicated nonlinear

function of the data. We can derive an estimate of the information matrix by evaluating the actual (not expected) second derivatives of the LLF of the ML estimates:

$$[I(\theta_0)]^{-1} = \left(-\frac{\partial^2 \ln L(\widehat{\theta})}{\partial \widehat{\theta}\, \partial \widehat{\theta}'} \right)^{-1}.$$

This computation will require that the second derivatives are evaluated, however, and expressing them analytically may be burdensome. As an alternative, we may use the **outer product of gradients (OPG)** estimator,

$$[I(\theta_0)]^{-1} = \left[\sum_{i=1}^{n} \widehat{g}_i \widehat{g}_i' \right]^{-1} = \left[\widehat{G}' \widehat{G} \right]^{-1},$$

where

$$\widehat{g}_i = \frac{\partial \ln f(x_i, \widehat{\theta})}{\partial \widehat{\theta}},$$

$$\widehat{G} = [\widehat{g}_1, \widehat{g}_2, \ldots \widehat{g}_n]'.$$

$\widehat{G}$ is a $n \times K$ matrix with i^{th} row equal to the transpose of the i^{th} vector of derivatives with respect to the estimated parameters (the gradient vector for the i^{th} observation, given the current guess for θ). This expression is very convenient, since it does not require evaluation of the second derivatives (i.e. computation of the Hessian). This estimate of the covariance matrix is also known as the BHHH (Berndt, Hall, Hall and Hausman) estimator. Nevertheless, if the estimator based on the second derivatives is available, it will often be preferable, especially in small to moderate sized samples.

Many econometric software packages provide the option to calculate ML estimate by alternative methods, or by combinations of algorithms. For instance, Stata's `ml` command will default to a Newton–Raphson method, `technique(nr)`, but will also allow specification of the BHHH

method (`technique(bhhh)`, the Davidon–Fletcher–Powell method (`technique(dfp)` or the Broyden–Fletcher–Goldfarb–Shanno method, `technique(bfgs)`. These techniques are described in Gould, Pitblado and Sribney, *Maximum Likelihood Estimation with Stata*, 2d ed. (2003). All four of these methods may be employed, by themselves or in combination, with Stata's `lf` (linear form) method. For all but the `bhhh` method, Stata computes estimates of the covariance matrix via `vce(oim)`, utilizing the observed information matrix (the inverse of the negative Hessian). Alternatively, OPG standard errors may be computed via `vce(opg)`, which is the default method when `technique(bhhh)` is used. Most software that produces maximum likelihood estimates affords similar capabilities.

Asymptotically equivalent test procedures

We consider MLE of a parameter vector θ and a test of the hypothesis $c(\theta) = q$. There are three approaches that we might use to test the hypothesis:

- **Likelihood ratio (LR) test.** If the restrictions $c(\theta) = 0$ are valid, then imposing them will not lead to a large reduction in the LLF. The test is thus based on the ratio $\lambda = \frac{\hat{L}_R}{\hat{L}_U}$, or in terms of log–likelihood the difference $(\ln L_R - \ln L_U)$, where L_U is the likelihood function value of the unrestricted estimate and L_R is the likelihood function value of the restricted estimate. Under the null hypothesis, $-2(\ln L_R - \ln L_U) \sim \chi^2$, with degrees of freedom equal to the number of restrictions imposed in $c()$. This is the natural test to apply in a MLE setting, but it requires that we solve two MLE

problems since both the unrestricted and restricted estimates must be computed.

- **Wald test.** If the restrictions are valid, then $c(\widehat{\theta}_{MLE})$ should be close to zero, since the MLE is consistent. The Wald test statistic is then a quadratic form in the difference

$$\left[c(\widehat{\theta}) - q\right]' \left(Asy.Var.\left[c(\widehat{\theta}) - q\right]\right)^{-1} \left[c(\widehat{\theta}) - q\right].$$

Under the null hypothesis, this quadratic form is distributed as χ^2, with degrees of freedom equal to the number of restrictions imposed in $c()$. The Wald test only requires computation of the unrestricted model.

- **Lagrange multiplier (LM, or score) test.** If the restrictions are valid, the restricted estimates should be near the point that

maximizes the LLF, and the slope of the LLF should be near zero at the restricted estimator. The test is based on the slope of the LLF at the point where the function is maximized subject to the restrictions, and is a quadratic form in the scores (first derivatives) of the restricted LLF:

$$\left(\frac{\partial \ln L(\widehat{\theta}_R)}{\partial \widehat{\theta}_R}\right)' \left[I(\widehat{\theta}_R)\right]^{-1} \left(\frac{\partial \ln L(\widehat{\theta}_R)}{\partial \widehat{\theta}_R}\right)$$

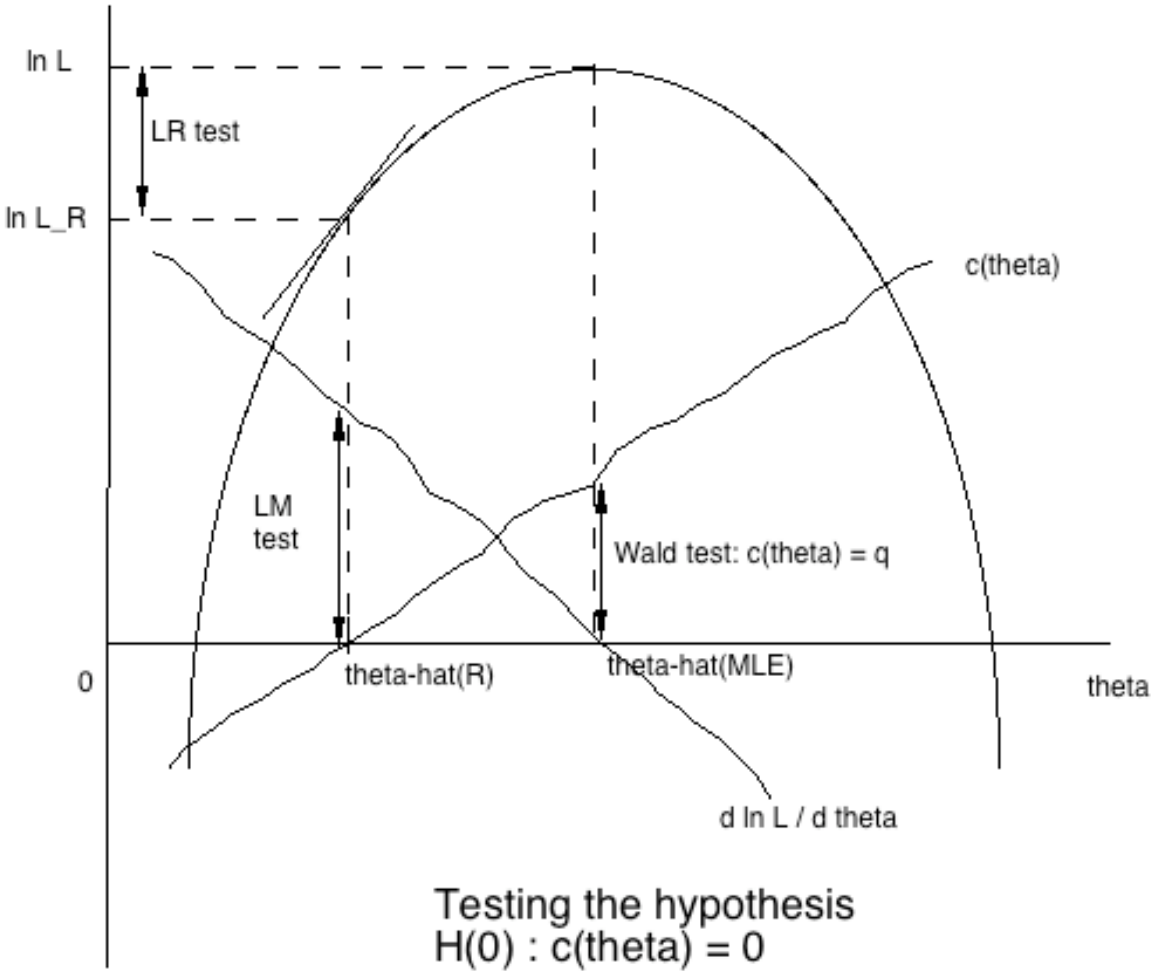
Under the null hypothesis, this quadratic form is distributed as χ^2, with degrees of freedom equal to the number of restrictions imposed in $c()$. The LM test only requires computation of the restricted model.

All three of these tests have the same asymptotic distribution, and are thus equivalent. However, in finite samples, they will yield different results, and may even lead to different inferences. Ease of computation is sometimes

an issue; for instance, estimating a restricted model in the context of linear regression is often difficult if the constraints are nonlinear. In that case, the Wald test (which does not require explicit estimation of the restricted model) may be preferable. On the other hand, if the model is nonlinear in the parameters and the restrictions render it linear, the LM test may be easily computed. For an explicit MLE problem, the LR test is usually straightforward. For a linear model, it has been shown that the χ^2 statistics from the three tests follow the ordering

$$W \geq LR \geq LM,$$

so that the LM test will be the most conservative (if it rejects, the others will). A graphical representation of the tests:



Testing the hypothesis
H(0) : c(theta) = 0

In Stata, the `test` command performs Wald tests of linear constraints. The `testnl` command performs Wald–type tests via the "delta method," an approximation appropriate in large samples. The `lrtest` command performs likelihood ratio tests after maximum likelihood estimation. Many commonly employed tests are

actually LM tests: in particular, any test which may be described as an "$n\,R^2$" test is an LM test. Note also that many of these test procedures will generate F–statistics, rather than χ^2 statistics. Stata uses the rule that if small–sample inferences are presented for the parameter estimates–e.g. if t statistics are computed from the parameters' estimated standard errors–then tests based on those estimates will be reported as F. If large–sample inferences are presented (so that z statistics are computed from the parameters' estimated standard errors), those estimates are reported as χ^2. Since the F statistic is the ratio of two independent χ^2 variables, and in most large–sample problems the denominator $\chi^2 \rightarrow j$, where j is the number of restrictions, it follows that $F = j\,\chi^2$. For a single restriction ($j = 1$), the reported $F_{(j,\infty)}$ and $\chi^2(j)$ will be identical.