Econometrics

MLE Basics

Conditional and Joint Distributions

For conditional probability it is known that:

$$f_Y(y|X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

This means that the probability distribution for Y given that X = x is equal to the joint probability distribution for X and Y divided by the probability distribution of X.

It is also the case that the reverse is true:

$$f_X(x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

We can therefore derive the following relation:

$$f_Y(y|X = x)f_X(x) = f_{X,Y}(x,y) = f_X(x|Y = y)f_Y(y)$$

If events X and Y are independent, then this equation simplifies to:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Applying this to a set of random variables y_t each with identical and independent distributions conditional upon the unknown (but constant for all ys) population parameters θ, we can derive the following joint density function:

$$f(y_1, \dots, y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

The Likelihood Function

The probability density function $f(y_i; \theta)$ depends only upon the values of y and the values of θ. Thus, we can alter the value of the function by changing either of these variables/parameters. In the real world, however, we actually observe the y random variables in our sample, and want to use that information to infer the population parameters θ. Thus, we can think of the above equation (the joint density function) as being the likelihood function, a measure of how likely it is that the parameters θ prevail given the observed values of $y_1, \dots, y_n$.

The likelihood function is thus defined as:

$$L(\theta; y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

The interpretation of the likelihood function is always in terms of ratios. If, for example, $L(\theta_1)/L(\theta_2) > 1$, then θ_1 explains the data better than θ_2.

Note that we are only dealing with the value of the probability distribution function, not the actual probability; that is why we do not need to introduce any integrals.

Converting to Log-Likelihood

It is often more convenient to work with the log-likelihood. This is derived as follows:

$$L(\theta; y_i) = \prod_{i=1}^{n} f(y_i; \theta)$$

$$\log L(\theta; y_i) = \log \prod_{i=1}^{n} f(y_i; \theta)$$

$$\log L(\theta) = \sum \log f(y_i)$$

If there is more than one parameter to the probability distribution of Y, then θ will be a vector

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_k \end{bmatrix}$$

Finding the Maximum Likelihood Estimator

To get the maximum likelihood estimator, the likelihood function must be maximised with respect to each one of these parameters. Thus we must solve for:

$$G(\hat{\theta}) = \frac{\partial \log L(\hat{\theta})}{\partial \hat{\theta}} = \begin{bmatrix} \frac{\partial \log L(\hat{\theta})}{\partial \theta_1} \\ \frac{\partial \log L(\hat{\theta})}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

With this information, we can solve these equations to find the maximum likelihood estimates for each of the parameters of the distribution.

$$G(\hat{\theta}) = \begin{bmatrix} \frac{\partial \log L(\hat{\theta})}{\partial \hat{\theta}_1} \\ \frac{\partial \log L(\hat{\theta})}{\partial \hat{\theta}_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \therefore \hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2\}$$

Checking for Negative Definiteness

We can check these maximum likelihood estimators by calculating the Hessian, and checking to see whether the determinant of the Hessian is negative definite. If so, then our values are correct.

Calculating the Hessian for a multi-parameter function is complicated by the fact that we must include not only all own-second derivatives, but also the cross-second derivatives of each term.

For two parameters, this is expressed in matrix form as follows:

$$H(\hat{\theta}) = \begin{bmatrix} \frac{\partial^2 \log L(\hat{\theta})}{\partial \hat{\theta}_1^2} & \frac{\partial^2 \log L(\hat{\theta})}{\partial \hat{\theta}_2 \hat{\theta}_1} \\ \frac{\partial^2 \log L(\hat{\theta})}{\partial \hat{\theta}_1 \hat{\theta}_2} & \frac{\partial^2 \log L(\hat{\theta})}{\partial \hat{\theta}_2^2} \end{bmatrix}$$

Now that we have the Hessian, we must check that both the following criteria for negative definiteness are met:

$$a_{1,1} < 0$$
$$|H| = a_{1,1}a_{2,2} - a_{1,2}a_{2,1} > 0$$

Note that a computer is used to check the negative definiteness of the Hessian for more than two parameters.

Finding the MLE Variance

The Information Matrix

The information matrix is useful for calculating the variances of our maximum likelihood estimators. It is calculated as follows:

$$\begin{aligned} I(\theta) &= -E\big(H(\theta)\big) \\ &= -\sum H(\theta)\frac{1}{T} \end{aligned}$$

If there is only one unknown parameter, the information matrix will actually be a scalar.

The Asymptotic Distribution

It turns out that as the sample size increases ad infinitum, that is $T \to \infty$, the maximum likelihood estimator $\hat{\theta}$ has a normal distribution with mean θ_0 (the actual population parameter) and variance-covariance matrix $I(\theta_0)^{-1}$. This is compactly written as:

$$\hat{\theta} \to N(\theta_0, I(\theta_0)^{-1})$$

Note that θ_0 refers to the true population parameter (which is unknown). The second value will be the variance of $\hat{\theta}$ (if it is scalar), or if it is a matrix then the diagonal terms will be the variances for each of the different parameters. Also note that $\hat{\theta}$ is normally distributed regardless of the underlying distribution of y (central limit theorem).

A typical case can be rewritten as a standard normal distribution as shown below:

$$\hat{\sigma} \to N\left(\mu, \frac{\sigma^2}{T}\right)$$
$$\left(\frac{\hat{\sigma} - \mu}{\sigma}\right)\sqrt{T} \to N(0,1)$$

The main purpose of the asymptotic distribution is that it allows us to derive the variances of the estimated parameters, which in turn can be used when conducting hypothesis tests.

Finding the MLE Variance

To obtain the variances of the MLE estimates, we simply calculate the Hessian matrix for the maximum likelihood estimator of the population parameters, then take the negative expected value of this matrix (expectation is done element by element), and then simply read off the variances from the resulting matrix.

For example, if the two unknown parameters have been found to be given by $\hat{\theta}_1 = \bar{y}; \ \ \hat{\theta}_2 = \hat{\sigma}^2$ and the population is normally distributed then asymptotically:

$$I(\theta_0)^{-1} = \begin{bmatrix} \frac{\sigma^2}{T} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}$$

In reality, we evaluate the estimates of the standard errors using the MLE:

$$I\left(\hat{\theta}\right)^{-1} = \begin{bmatrix} \frac{\hat{\sigma}^2}{T} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{T} \end{bmatrix}$$

$$\bar{y} \to N\left(\mu_0, \frac{2{\sigma_0}^4}{T}\right)$$

$$\hat{\sigma}^2 \to N\left(\sigma_0^2, \frac{{\sigma_0}^2}{T}\right)$$

Terminology and Standard Errors

Standard errors can be computed using either:

1. The square root of the diagonal elements of the inverse of the information matrix evaluated at the maximum likelihood estimator $I\left(\hat{\theta}\right)^{-1} = -E\left(H\left(\hat{\theta}\right)\right)^{-1} = \Omega\left(\hat{\theta}\right)$. This is the best theoretical quantity to use, but may be difficult to compute in some cases.
2. The square root of the diagonal elements of the inverse of the negative Hessian matrix evaluated at the maximum likelihood estimator $-H\left(\hat{\theta}\right)^{-1}$. This is an approximation of the information matrix, and tends to be the quantity reported as it is easier to use. For some models this approach yields identical standard errors to the information matrix approach.
3. Note that the BHHH algorithm uses the negative of the outer product of the gradients as an approximation for the Hessian:

$$H(\theta_k) \approx -J(\theta_k) = -\left[\sum_{t=1} g_t(\theta_k) g_t(\theta_k)'\right]$$

Numerical Methods

Taylor's Theorem

Taylor's theorem asserts that any sufficiently smooth function can locally be approximated by polynomials. It can be written as follows:

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^n(a)}{n!}(x-n)^n$$

Where a is a point close to point x.

Applying this to the maximum likelihood estimator and taking only the first order results we can derive the following:

$$\begin{aligned}
G(\theta) &\approx G(\theta_0) + H(\theta_0)(\theta - \theta_0) \\
G(\hat{\theta}) &\approx G(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) \\
&At\ MLE\ G(\hat{\theta}) = 0 \\
\therefore 0 &\approx G(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) \\
&\approx G(\theta_0) + \hat{\theta}H(\theta_0) - \theta_0 H(\theta_0) \\
-\hat{\theta}H(\theta_0) &\approx G(\theta_0) - \theta_0 H(\theta_0) \\
\hat{\theta} &\approx -\frac{G(\theta_0)}{H(\theta_0)} + \theta_0 \\
\hat{\theta} &\approx \theta_0 - \frac{G(\theta_0)}{H(\theta_0)}
\end{aligned}$$

This result is very useful for numerical methods.

The Newton-Raphson Algorithm

Many functional forms cannot be solved analytically to produce a simple solution for the MLE. In such circumstances, numerical methods must be used to find the MLE iteratively.

Let θ_k be the value of the scalar parameter at the kth iteration. The Newton-Raphson algorithm yields the updated parameter θ_{k+1} according to the following rule:

$$\theta_{k+1} = \theta_k - \frac{G(\theta_k)}{H(\theta_k)}$$

Note that this algorithm begins with an initial parameter value θ_0, which becomes the first θ_k, and effectively is chosen arbitrarily. The algorithm will work regardless of what this initial parameter value is.

The algorithm is continually iterated until $\theta_{k+1} \approx \theta_k$, at which point we say that $\theta_k = \theta_{k+1} = \hat{\theta}$. This is effective at finding the maximum likelihood estimator because if $\theta_{k+1} = \theta_k$ then:

$$\begin{aligned}
\theta_{k+1} - \theta_k = 0 &= -\frac{G(\theta_k)}{H(\theta_k)} \\
\frac{G(\theta_k)}{H(\theta_k)} &= 0 \\
G(\theta_k) &= 0
\end{aligned}$$

This is the condition of the maximum likelihood estimator, and thus $\theta_{k+1} = \hat{\theta}$ at the final iteration.

The vector case of the Newton-Raphson algorithm is the same as the scalar case except that all calculations are performed using vectors and matrices.

The BHHH Algorithm

One potential problem with the Newton-Raphson algorithm is that the Hessian is not guaranteed to be negative definite. If this happens then the algorithm will not find global maxima of the log-likelihood function.

The BHHH algorithm avoids this problem by approximating the Hessian as follows:

$$H(\theta_k) \approx -\left[\sum_{t=1} g_t(\theta_k) g_t(\theta_k)'\right]$$

Where $\sum_{t=1} g_t(\theta_k) g_t(\theta_k)'$ is known as the outer product of the gradients, and $g_t(\theta_k)$ is the vector of gradients, each element of the vector being evaluated for parameter θ_k. Such a vector is calculated for each value of t, multiplied by its transpose matrix, and then the resulting vectors for each t are all summed together. This works because the squared gradient term forces the hessian approximation to be positive, and then placing a negative out the front ensures that it is negative definite.

The updating scheme of the BHHH algorithm is then given by:

$$\theta_{k+1} = \theta_k + \left[\sum_{t=1} g_t(\theta_k) g_t(\theta_k)'\right]^{-1} \sum g_t(\theta_k)$$

The Marquardt Algorithm

The Marquardt algorithm is an extension of the BHHH algorithm by changing the Hessian approximation to:

$$H(\theta_k) \approx -\left[\sum_{t=1} g_t(\theta_k) g_t(\theta_k)' + kI\right]$$

Where I is the unit matrix and k > 0 is a scalar. The parameter k is chosen to ensure that the Hessian approximation is negative definite as opposed to negative semidefinite in the BHHH algorithm.

Starting Values

All of the algorithms require some initial value to start the algorithm. Choices are:

- 1 Arbitrary: Choose q(0) based on random numbers
- 2 Simple model: Estimate a simpler model say the linear regression model to provide starting values for at least some of the unknown parameters
- 3 Vary values: if the model doesn't work for one set of values, choose another set

Note that if log L is globally concave (i.e. $H(\theta)$ is negative definite for all values of θ), then an arbitrary choice will always result in the algorithm converging to the maximum likelihood estimator. If log L is not globally concave then an arbitrary choice may result in the algorithm converging to a local maximum and not the global maximum.

Hypothesis Testing

Likelihood Ratio Test

The Likelihood ratio (LR) test requires estimating the model under both the null and alternative hypotheses. For this purpose, $\hat{\theta}_0$ is the MLE estimator under the null, and $\hat{\theta}_1$ is the MLE under the alternative.

Note that it is always the case that $L(\hat{\theta}_0) < L(\hat{\theta}_1)$, as the restrictions that we place under the model under the null will always cause it to fit the data less well than the maximum possible fit under $\hat{\theta}_1$. The question, then, is how important is this difference? This can be measured by the ratio of the likelihoods:

$$\lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)}$$

Unfortunately, we do not know the distribution of λ, so it is more convenient to use:

$$\begin{aligned} LR &= -2\log\lambda \\ &= -2\big(\log L(\hat{\theta}_0) - \log L(\hat{\theta}_1)\big) \end{aligned}$$

This log-ratio is distributed asymptotically as chi-square with K degrees of freedom where K is the number of restrictions under the null hypothesis. After calculating the LR (either by hand or by estimating both the null and alternative models in Eviews), a hypothesis test can be conducted by selecting an appropriate critical value from the chi-square distribution.

The decision rule will be: Reject H_0 if $LR \geq \chi^2_c$

Wald Test

The Wald test (WD) test requires estimating the model just under the alternative hypothesis. It thus has a computational advantage over the likelihood ratio test which requires estimating the model under both the null and the alternative hypotheses.

This works because in the Wald test we effectively calculate the MLE under the alternative (i.e. just the unrestricted MLE), and then see how much it differs from the estimate assumed under the null. If there is a significant difference between the two, we reject the null. This contrasts to the likelihood ratio test, where we actually compare the log-likelihood under the null to that under the alternative.

From the asymptotic distribution we know that:

$$\hat{\theta} \to N(\theta_0, I(\theta_0)^{-1})$$

$$Z = \left(\frac{\hat{\theta} - \theta_0}{\sqrt{I(\theta_0)^{-1}}}\right) \rightarrow N(0,1)$$

This further implies that:

$$Z^2 = \left(\frac{\hat{\theta} - \theta_0}{\sqrt{I(\theta_0)^{-1}}}\right)^2 \rightarrow \chi_k^2$$

Where k is the number of unknown parameters to be estimated.

The Wald test is conducted using the quantity WD, defined as shown below. Note that the middle term is also written as $\Omega(\hat{\theta}_1)$ or $J(\hat{\theta}_1)^{-1}$. Terms $\hat{\theta}_1$ refer to the MLE parameter estimates under the alternative hypothesis, while the θ_0 are the parameter values as specified under the null.

$$WD = (\hat{\theta}_1 - \theta_0)'(\sum g_t(\hat{\theta}_1) g_t(\hat{\theta}_1)')^{-1}(\hat{\theta}_1 - \theta_0)$$

The decision rule will be: Reject H_0 if $WD \geq \chi_c^2$

Wald Test with Restrictions

In many cases it is necessary to compute tests of linear restrictions. Such restrictions can be expressed in matrix form as shown in the examples below.

$$H_0: \beta_2 + \beta_3 = 1$$
$$H_1: \beta_2 + \beta_3 \neq 1$$

This can be written in the form $R\theta = Q$, where these are matrices as given below:

$$R = [0 \quad 1 \quad 1], \qquad \theta = \begin{bmatrix}\beta_1 \\ \beta_2 \\ \beta_3\end{bmatrix}, \qquad \mathrm{Q} = [1]$$

If two of the coefficients are hypothesised to be equal then:

$$R = [0 \quad 1 \quad -1], \qquad \theta = \begin{bmatrix}\beta_1 \\ \beta_2 \\ \beta_3\end{bmatrix}, \qquad \mathrm{Q} = [0]$$

For linear hypotheses of this form, the Wald statistic is given by:

$$WD = (R\hat{\theta}_1 - Q)' \left(R\left[\sum g_t(\hat{\theta}_1) g_t(\hat{\theta}_1)'\right]^{-1} R'\right)^{-1} (R\hat{\theta}_1 - Q)$$

Note that this works because $\mathrm{var}(R\hat{\theta}_1) = R\ \mathrm{var}(\hat{\theta}_1)R' = R\left[\sum g_t(\hat{\theta}_1) g_t(\hat{\theta}_1)'\right]^{-1} R'$. The test statistic is asymptotically distributed under the null hypothesis as χ_k^2, K is the number of restrictions.

The decision rule will be: Reject H_0 if $WD \geq \chi_c^2$

Lagrange Multiplier Test

The Lagrange multiplier test (LM) requires estimating the model just under the null hypotheses. This is sometimes convenient of the alternative is particularly hard to estimate. This can be done by

evaluating the MLE under the assumption that the null is true, and then checking to see if the resulting gradient vector is significantly different to zero. If it is, then we can reject the null, as the gradient of the MLE should equal zero.

The Lagrange multiplier statistic is given by:

$$LM = g'(\hat{\theta}_0)\left[\sum g_t(\hat{\theta}_0)g_t(\hat{\theta}_0)'\right]^{-1} g(\hat{\theta}_0)$$

Which is asymptotically distributed under the null hypothesis as χ^2_k, K is the number of restrictions.

The decision rule will be: Reject H_0 if $WD \geq \chi^2_c$

Microeconometric Models

Probit Regression Model

The probit regression model is useful when the dependent variable y_t is binary in that it takes on one of two values. For example, it is common to let the two values be 0 and 1 to signify the two choices, like vote choice. A good way to model this situation is using the Bernoulli distribution.

$$f(y_t;\theta) = \mu^{y_t}(1-\mu)^{1-y_t}$$

Where $\theta = \mu$ is the probability that y = 1. We can find the MLE for this just like any other pdf.

This simple model assumes that the probability is the same across the entire sample. However, we can allow the distribution to change across the sample by assuming that μ is a function of the explanatory variable(s) x_t. Thus we end up with:

$$\mu_t = \beta_1 + \beta_2 x_t$$

As the variable μ_t is a probability, however, it is necessary to impose the restriction on the above function that $0 < \mu < 1$. Although there are numerous ways of doing this, one of the most common is to multiply the basic μ value by the cumulative density function of the normal distribution:

$$\mu_t = \Phi(\beta_1 + \beta_2 x_t) = \int_{-\infty}^{\beta_1+\beta_2 x_t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Where $z = \beta_1 + \beta_2 x_t$.

The MLE can then be calculated by substituting this equation into the original Bernoulli distribution specification, such that we have:

$$f(y_t;\theta) = \mathcal{L} = \Phi(\beta_1 + \beta_2 x_t)^{y_t}\left(1 - \Phi(\beta_1 + \beta_2 x_t)\right)^{1-y_t}$$

Maximising the log-likelihood will then give the values of β_1 and β_2 that are most likely given the model specification and provided input values of x and y.

Poisson Regression Model

These are useful when we want to investigate dependent variables that take on integer values, especially when these refer to counts of number of events. The Poisson distribution is given as:

$$f(y_t) = \frac{\mu^{y_t} \exp(-\mu)}{y_t!}$$

Where $\theta = \mu$ is the parameter representing the mean of the distribution. We can then calculate the MLE for μ as shown above.

Often, however, we will be interested in how μ changes across the sample in accordance with changes in different explanatory variables x_t. Hence:

$$\mu_t = \beta_1 + \beta_2 x_t$$

Similar to above, we must restrict the values of μ so that it cannot be negative. A common way to do this is to apply the following formula:

$$\mu_t = \exp(\beta_1 + \beta_2 x_t)$$

To interpret the results, we must take the derivative of μ with respect to the desired explanatory variable. To interpret dummy or binary variables, we can calculate the percentage change between the μ under the 'on' and 'off' versions of the variable.

Heteroskedasticity

In the case of the simple linear model $y_t = \beta_1 + \beta_2 x_t + u_t$, (where $u_t \sim N(0, \sigma^2)$) the mean of this distribution changes with each different value of x, however the variance of the distribution σ^2 does not change. This means that the model is homoskedastic.

Heteroskedasticity is about allowing the variance to vary over the sample. This is a natural extension of the linear regression model which allows the mean of y to vary over the sample. Why not allow the variance to change over the sample as well?

A common way of doing this is to set up the variance as its own separate function, scaled with the exponential function in order to force the variance to always be positive.

$$\sigma_t^2 = \exp(\alpha_1 + \alpha_2 z_t)$$

Where z is an additional random variable. This introduces two additional parameters into the model: α_1 and α_2.

The LM test of heteroskedasticity involves the following steps:

1. Regress y_t on a constant and x_t and extract the OLS residuals $\hat{u}_t$
2. Regress $\hat{u}_t^2$ on a constant and z_t
3. Compute $LM = TR^2$ where T is the sample size and R^2 is the coefficient of determination from the second stage regression and compare with χ_1^2

Intuition: $\hat{u}_t^2$ is an estimate of the variance at each t. If the population variance is constant, regressing $\hat{u}_t^2$ on z_t does not help to explain movements in the variance, as there are none ie $R^2 = 0$.

Autocorrelation

Any patterns in the residual of a regression are indicative of a misspecification of the model, as the residuals are picking up that excluded pattern information. In particular, any 'trends' of positive or negative residuals indicate that the residual in one period is related to/partly determined by the residual in the previous period. This is autocorrelation; residuals are correlated with previous residuals in the same series.

This suggests that we should model the disturbance term as:

$$u_t = \rho_1 u_{t-1} + v_t$$

We can rearrange the basic regression model to be written as:

$$u_t = y_t - \beta_1 - \beta_2 x_t$$
$$u_{t-1} = y_{t-1} - \beta_1 - \beta_2 x_{t-1}$$

Substituting these equations into the original disturbance term equation we have:

$$(y_t - \beta_1 - \beta_2 x_t) = \rho_1 (y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + v_t$$
$$y_t = \beta_1 (1 - \rho_1) + \beta_2 (x_t - \rho_1 x_{t-1}) + \rho_1 y_{t-1} + v_t$$

This probability model is still normally distributed (as v is normally distributed), but it is no longer an idd distribution, as each value y is conditional upon y_{t-1}. To find the log-likelihood of this model we must therefore use the conditional normal distribution function:

$$f(y_t | y_{t-1}; \beta_1, \beta_2, \sigma^2, \rho_1)$$

$$f(y_2 | y_1) = \frac{f(y_2, y_1)}{f(y_1)}$$
$$f(y_2 | y_1) f(y_1) = f(y_1, y_2)$$

This can be generalised for T observations (utilising the fact that each observation is only conditional upon the previous observation and not any before that), we derive the following:

$$f(y_1, y_2 \dots y_T) = f(y_T | y_{T-1}) f(y_{T-1} | y_{T-2}) \dots f(y_2 | y_1) f(y_1)$$
$$\log \mathrm{L} = \log f(y_T | y_{T-1}) f(y_{T-1} | y_{T-2}) \dots f(y_2 | y_1) f(y_1)$$
$$= \log f(y_1) + \sum \log f(y_t | y_{t-1})$$
$$\approx \sum \log f(y_t | y_{t-1})$$

Note that this approximation is generally used because as t increases, $\log f(y_1)$ becomes increasingly less important to the overall value of logL.

To test whether this more complicated model is necessary, we can simply introduce a restricted model where $\rho = 0$, and then test to see of this model is valid.

The LM test of autocorrelation involves the following steps:

1. Regress y_t on a constant and x_t and extract the OLS residuals $\hat{u}_t$
2. Regress $\hat{u}_t$ on a constant, x_t and $\hat{u}_{t-1}$

The ARCH Model

This stands for AutoRegressive Conditional Heteroskedasticity Model. Basically this model permits us to have the variance as a function of the lagged disturbance term, instead of having the mean as lagged. The form of the model is shown below:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$
$$\sigma_t^2 = \alpha_1 + \alpha_2 u_{t-1}^2$$
$$u_t \sim N(0, \sigma_t^2)$$

We can then test for heteroskedasticity/autocorrelation by conducting any of the usual tests with the null that $\alpha_2 = 0$.

Specification Analysis

Model Misspecification

One or more of the following things may be incorrect about our model:

- 1 Distribution: The distribution is not normal, say exponential.
- 2 Mean: The disturbance is not independent, but is related to previous disturbances (autocorrelated). This has the implication that the mean should also include x_{t-1} and y_{t-1}
- 3 Variance: The variance is not homoskedastic, but is time-varying (heteroskedastic). This has the implication that the variance should have some variables which vary over the sample

Calculating the Variance

- The variance-covariance matrix of the MLE is $\Omega(\hat{\theta})$. It can be calculated in three ways:
- Outer product of gradients: $J(\hat{\theta})$, $\Omega(\hat{\theta}) = \left[\sum g_t(\hat{\theta}) g_t(\hat{\theta})'\right]^{-1} = J(\hat{\theta})^{-1}$
- Main method used so far as this is what the LogL option computes
- Hessian: $H(\hat{\theta})$, $\Omega(\hat{\theta}) = \left[-H(\hat{\theta})\right]^{-1}$
- Information matrix: $I(\hat{\theta})$, $\Omega(\hat{\theta}) = I(\hat{\theta})^{-1}$
- Most appropriate from a theoretical point of view, but can be difficult to compute

Variance with Misspecified Model

- If the model is misspecified, the variance-covariance matrix requires using both $H(\hat{\theta})$ and $J(\hat{\theta})$, not just one of them
- This method is used to calculate the White Heteroskedasticity consistent standard errors

$$
\begin{aligned}
G(\theta) &\approx G(\theta_0) + H(\theta_0)(\theta - \theta_0) \\
G(\hat{\theta}) &\approx G(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) \\
0 &\approx G(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) \\
-\frac{G(\theta_0)}{H(\theta_0)} &\approx (\hat{\theta} - \theta_0) \\
(\hat{\theta} - \theta_0) &\approx -G(\theta_0)H(\theta_0)^{-1} \\
\Omega(\hat{\theta}) &= E\left[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'\right] \\
&= E[(-G(\theta_0)H(\theta_0)^{-1})(-G(\theta_0)H(\theta_0)^{-1})'] \\
&= E\left[G(\theta_0)H(\theta_0)^{-1}G(\theta_0)'H(\theta_0)^{-1'}\right]
\end{aligned}
$$

Note that $H(\theta_0)^{-1} = H(\theta_0)^{-1'}$ and for normal errors $E[H(\theta_0)^{-1}] = H(\theta_0)^{-1}$

$$
= H(\theta_0)^{-1}E[G(\theta_0)G(\theta_0)']H(\theta_0)^{-1}
$$

As $J(\hat{\theta}) = E\left[G(\hat{\theta})G(\hat{\theta})'\right]$ and replacing θ_0 with $\hat{\theta}$ we have:

$$
\Omega(\hat{\theta}) = H(\hat{\theta})^{-1}J(\hat{\theta})H(\hat{\theta})^{-1}
$$

Degrees of Freedom

The $\Omega(\hat{\theta})$ matrix (and hence standard errors) can be adjusted for the degrees of freedom by multiplying the matrix by $\frac{T}{T-K'}$, where T is the sample size and K is the number of parameters.

Autocorrelation Version

Consider the following expression for the outer product of the gradients:

$$
\begin{aligned}
E[G(\theta_0)G(\theta_0)'] &= E\begin{bmatrix}\sum g_{1,t} \\ \sum g_{2,t}\end{bmatrix}[\sum g_{1,t} \quad \sum g_{2,t}] \\
&= E\begin{bmatrix}(\sum g_{1,t})^2 & (\sum g_{1,t})(\sum g_{2,t}) \\ (\sum g_{1,t})(\sum g_{2,t}) & (\sum g_{2,t})^2\end{bmatrix}
\end{aligned}
$$

If it is assumed that errors are uncorrelated with each other, then $E(e_t e_{t+1}) = 0$, in which case cross-products can be ignored, and the above expression simplifies to:

$$
= E\begin{bmatrix}\sum (g_{1,t})^2 & \sum (g_{1,t})(g_{2,t}) \\ \sum (g_{1,t})(g_{2,t}) & \sum (g_{2,t})^2\end{bmatrix}
$$

This in turn can be estimated using the MLE:

$$
\begin{aligned}
J(\hat{\theta}) &= \begin{bmatrix}\sum (g_{1,t})^2 & \sum (g_{1,t})(g_{2,t}) \\ \sum (g_{1,t})(g_{2,t}) & \sum (g_{2,t})^2\end{bmatrix} \\
&= [\sum g_t(\hat{\theta})g_t'(\hat{\theta})]
\end{aligned}
$$

However, in the event that errors are not independent (e.g. autocorrelation), we need to re-estimate the expression using the Newey-West outer product of the gradients:

$$
J_{NW}(\hat{\theta}) = [\sum g_t g_t'] + w_1[\sum g_t g_{t-1}' + \sum g_{t-1} g_t'] + w_2[\sum g_t g_{t-2}' + \sum g_{t-2} g_t']
$$

$$= \sum \begin{bmatrix} (\sum g_{1,t})^2 & (\sum g_{1,t})(\sum g_{2,t}) \\ (\sum g_{1,t})(\sum g_{2,t}) & (\sum g_{2,t})^2 \end{bmatrix} + w_1 \begin{bmatrix} \sum(g_{1,t})(g_{1,t-1}) & \sum(g_{1,t})(g_{2,t-1}) \\ \sum(g_{2,t})(g_{1,t-1}) & \sum(g_{2,t})(g_{2,t-1}) \end{bmatrix}$$
$$+ w_1 \begin{bmatrix} \sum(g_{1,t-1})(g_{1,t}) & \sum(g_{1,t-1})(g_{2,t}) \\ \sum(g_{2,t-1})(g_{1,t}) & \sum(g_{2,t-1})(g_{2,t}) \end{bmatrix} + w_1 \begin{bmatrix} \sum(g_{1,t})(g_{1,t-2}) & \sum(g_{1,t})(g_{2,t-2}) \\ \sum(g_{2,t})(g_{1,t-2}) & \sum(g_{2,t})(g_{2,t-2}) \end{bmatrix}$$
$$+ w_1 \begin{bmatrix} \sum(g_{1,t-2})(g_{1,t}) & \sum(g_{1,t-2})(g_{2,t}) \\ \sum(g_{2,t-2})(g_{1,t}) & \sum(g_{2,t-2})(g_{2,t}) \end{bmatrix}$$

This can then be used in the following expression to calculate the autocorrelation consistent standard errors. Note that adjustment for degrees of freedom is done last, after $\Omega(\hat{\theta})$ is calculated.

$$\Omega(\hat{\theta}) = H(\hat{\theta})^{-1} J_{NW}(\hat{\theta}) H(\hat{\theta})^{-1}$$

Concentrating out the Variance

To simplify the calculations the focus is on the mean parameters β_1, β_2. This is achieved by noting that the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t^2$$

To use this expression for **SIG2** in the **LogL** window the command is

$$SIG2 = @MEAN\,(u\hat{}2)$$

In which case the only parameters that are estimated are β_1 and β_2.

Simultaneous Equations

Introduction

Up until now it was assumed that only the dependent variable was stochastic, while all explanatory variables were deterministic. This means that $E(x_t) = x_t$. This is a useful assumption, but is not valid if x and y are drawn from a joint probability distribution $f(y_t, x_t)$.

Marginal and Conditional Distributions

The marginal distribution refers to the subset of a joint distribution function where the distribution of one variable is given in relation to the 'average' of the other. This is achieved by summing or integrating out the undesired variable. The marginal distribution for y and for x are:

$$f(y) = \int f(y,x)dx$$
$$f(x) = \int f(y,x)dy$$

The conditional distribution is similar to the marginal distribution, except that it gives the probability distribution for one variable given a fixed, particular value of the other variable, as opposed to the average of the other variable, as in marginal distributions. The conditional distribution of y given x is:

$$f(y|x) = \frac{f(y,x)}{f(x)} = \frac{f(y,x)}{\int f(y,x)dy}$$

The marginal distribution of (say) y will not be a function of x, as x has already been integrated out. However, the conditional distribution of $y|x$ will be a function of x, unless x and y are independent.

If we want to find the expectation of y, then we can simply integrate across all values of y multiplied by dy. The same goes for the expectation of x.

$$E(y) = \int yf(y)dy = \iint yf(y,x)dxdy$$
$$E(x) = \int xf(x)dx = \iint xf(y,x)dxdy$$
$$E(x) = \iint xyf(x,y)dxdy$$

We can use all these results to find the expected value of y given x:

$$E(y|x) = \int yf(y|x)dy = \int y\frac{f(y,x)}{f(x)}dy$$

Proving the Linear Conditional Expectation

$$E(y|x) = \int y\frac{f(y,x)}{f(x)}dy = \beta_1 + \beta_2 x$$
$$\int yf(y,x)dy = f(x)\beta_1 + f(x)\beta_2 x$$
$$\int \left[\int yf(y,x)dy\right]dx = \int [f(x)\beta_1 + f(x)\beta_2 x\,]dx$$
$$\int \left[\int yf(y,x)dx\right]dy = \beta_1\int f(x)dx + \beta_2\int xf(x)dx$$
$$\int yf(y)dy = \beta_1\int f(x)dx + \beta_2\int xf(x)dx$$
$$E(y) = \beta_1 + \beta_2 E(x)$$
$$\beta_1 = E(y) - \beta_2 E(x)$$

$$\int xyf(y,x)dy = f(x)x\beta_1 + f(x)\beta_2 x^2$$
$$\int x\left[\int yf(y,x)dy\right]dx = \int [f(x)x\beta_1 + f(x)\beta_2 x^2\,]dx$$
$$\iint yxf(y,x)dydx = \beta_1\int xf(x)dx + \beta_2\int x^2 f(x)\,dx$$
$$E(yx) = \beta_1 E(x) + \beta_2 E(x^2)$$

$$E(yx) - E(y)E(x) = \beta_1 E(x) + \beta_2 E(x^2) - [E(x)\beta_1 + \beta_2 E(x)^2]$$
$$E(yx) - E(y)E(x) = \beta_2 E(x^2) - \beta_2 E(x)^2$$
$$\beta_2 = \frac{E(yx) - E(y)E(x)}{E(x^2) - \beta_2 E(x)^2}$$
$$= \frac{cov(x,y)}{var(x)}$$

$$\therefore E(y|x) = \beta_1 + \beta_2 x$$
$$= E(y) - \left(\frac{cov(x,y)}{var(x)}\right)E(x) + \frac{cov(x,y)}{var(x)}x$$
$$= E(y) - \left(\frac{cov(x,y)}{var(x)}\right)E(x) + \frac{cov(x,y)}{var(x)}x$$
$$= E(y) - \frac{cov(x,y)}{var(x)}(E(x) - x)$$

The Law of Iterated Expectations

$$\begin{aligned} E_x\big(E(y|x)\big) &= \int E(y|x)f(x)dx \\ &= \int \left[\int y\frac{f(y,x)}{f(x)}dy\right]f(x)dx \\ &= \int \left[\int yf(y,x)dy\right]dx \\ &= \int y\left[\int f(y,x)dx\right]dy \\ &= \int yf(y)dy \\ &= E(y) \end{aligned}$$

Simultaneity Bias

For OLS estimation to work it is necessary for all explanatory variables to be uncorrelated with the disturbance terms u_t. This is automatically true if all explanatory variables are deterministic, as:

$$E(x_t u_t) = x_t E(u_t) = x_t * 0 = 0$$

This does not necessarily hold, however, if the explanatory variables are stochastic. If x and u are at all correlated, then the estimate of β_2 will experience simultaneity bias to the extent of $\frac{\text{cov}(x,u)}{\text{var}(x)}$. This requires the addition of another equation to the model.

Types of Variables

Endogenous variable: Variable determined within the system
Exogenous variable: Variable determined outside of the system
Predetermined variable: Lagged endogenous variable

Reduced Form

The reduced form of a simultaneous equation model is rewritten so that all endogenous variables are on the left of the equals sign as subjects of their equation, and all constants, exogenous variables and predetermined variables are on the right. It is often useful to use matrices to do this.

Two-Stage Least Squares

Estimation of the parameters of a system of equations by OLS results in biased and inconsistent parameter estimates, as at least one of the explanatory variables in each equation is correlated with its disturbance term. To avoid this, we can instead estimate a system of equations using two-stage least squares. To do this, we first must obtain the reduced form of the system of equations.

The two stages of TSLS are as follows:

- Estimate each equation of the reduced form by OLS to produce estimated values of each endogenous variables
- In the original equations, replace the RHS endogenous variables by the reduced form predicted values and estimate each equation by OLS

Thus, the following equations:

$$\begin{aligned} q_t &= \beta_1 + \beta_2 p_t + \beta_3 x_t + u_t \\ p_t &= \alpha_1 + \alpha_2 x_t + \alpha_3 y_t + v_t \end{aligned}$$

Are altered to become:

$$q_t = \beta_1 + \beta_2(\alpha_1 + \alpha_2 x_t + v_t) + \beta_3 x_t + \alpha_3 y_t + u_t$$

The method of two stage least squares is also known as Instrumental Variables (IV), as all of the endogenous variables in the original equations are proxied by 'instruments', which are exogenous and pre-determined variables. Note that when doing this in Eviews, each original equation must be estimated separately in a different TSLS window.

Model Identification

A statistical model is "identified" if the known information available implies that there is one best value for each parameter in the model whose value is not known. We will have 'enough information' if the number of parameters estimated in the final-form equation (e.g. after all endogenous variables are replaced by their reduced form equations) is at least as large as the number of parameters in the original equation. If we have more than we need, this is called over-determination. Fewer than we need leads to under-determination, and prevents the model from being estimated using TSLS. Note that we need to make this comparison with each original equation.

The following equation is over-identified:

$$\begin{aligned}
q_t &= \beta_1 + \beta_2 p_t + \beta_3 x_t + u_t \\
p_t &= \alpha_1 + \alpha_2 z_t + \alpha_3 y_t + v_t \\
q_t &= \beta_1 + \beta_2(\alpha_1 + \alpha_2 z_t + v_t) + \beta_3 x_t + \alpha_3 y_t + u_t \\
q &= f(\text{const}, z_t, x_t, y_t) \\
4 &> 3 \therefore \text{overidentified}
\end{aligned}$$

The following equation is under-identified:

$$\begin{aligned}
q_t &= \beta_1 + \beta_2 p_t + \beta_3 x_t + u_t \\
p_t &= \alpha_1 + \alpha_2 x_t + v_t \\
q_t &= \beta_1 + \beta_2(\alpha_1 + \alpha_2 x_t + v_t) + \beta_3 x_t + u_t \\
q &= f(\text{const}, x_t) \\
2 &< 3 \therefore \text{underidentified}
\end{aligned}$$