Econometrics, Lecture 3. Distributions, maximum likelihood ML and non-linearity

Ron Smith
EMS, Birkbeck, University of London

Autumn 2020

Last time

- Model

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\,\underset{k\times 1}{\beta} + \underset{T\times 1}{u}$$

- With assumptions

$$\begin{aligned} E(u) &= 0, \\ E(u\ u') &= \sigma^2 I_T, \end{aligned}$$

  X is of rank k and exogenous, independent of u.
- $\widehat{\beta} = (X'X)^{-1}X'y$, $V(\widehat{\beta}) = \sigma^2(X'X)^{-1}$, and $\widehat{\beta}$ is BLUE.
- Notice we did not assume anything about the distributions beyond the second moment exists and is constant. Now we introduce distributions.

Two regressions

- The original use of regression was by Francis Galton (1822-1911) as getting a conditional expectation from a joint distribution of two random variables, RV, e.g. of parents and childrens heights.
- The name came from observed regression towards the mean. Tall parents had children who were taller than average, but closer to the average.
- For Galton regression was not necessarily a causal relationship. The later use, which is dominant in econometrics, came from RA Fisher who treated the X as fixed by the experimenter, e.g. amount of fertisliser put on a plot.
- We now look at the conditional expectation interpretation.

Conditional distributions

- The joint distribution of the RV, $Y_t, \mathbf{X}_t$, can be written as the product of the distribution of Y_t conditional on $\mathbf{X}_t$, and the marginal distribution of $\mathbf{X}_t$:

$$D_j(Y_t, \mathbf{X}_t; \theta_j) = D_c(Y_t \mid \mathbf{X}_t; \theta_c) D_m(\mathbf{X}_t; \theta_m) \qquad (1)$$

  θ_j is a vector of parameters of the joint distribution, θ_c of the conditional distribution, θ_m of the marginal.
- If we regard causality going from $\mathbf{X}_t$ to Y_t, then the parameters of interest are θ_c, which we will usually denote by θ.
- $\mathbf{X}_t$ is weakly exogenous if there is no information in the marginal distribution of $\mathbf{X}$ about θ_c. Here exogeneity, is about the distributions of the observables $Y_t, \mathbf{X}_t$ not the unobservable u as in $\mathbf{X}$ is independent of u.
- We are interested in the first two moments of $D_c(Y_t \mid \mathbf{X}_t; \theta_c)$, the conditional expectation, $E(Y_t \mid \mathbf{X}_t; \theta_c)$, regression function, and variance. In the LRM $\theta_c = \theta = (\beta, \sigma^2)$.

Joint Normal (Gaussian) 1

- If scalar Y_t and k vector $\mathbf{X}_t$ have a joint normal distribution:

$$\begin{bmatrix} Y_t \\ \mathbf{X}_t \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right) \tag{2}$$

  $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}'_{yx}$ are $k \times 1$, $\boldsymbol{\Sigma}_{xx}$ is $k \times k$.

- The conditional expectation of Y_t is a linear function of $\mathbf{X}_t$:

$$E(Y_t \mid \mathbf{X}_t) = \mu_y + \left[\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right](\mathbf{X}_t - \boldsymbol{\mu}_x)$$

  Note $\left[\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right] = \boldsymbol{\beta}'$ corresponds to $(X'X)^{-1}X'y$.

- The error $u_t = Y_t - E(Y_t \mid \mathbf{X}_t) = Y_t - \boldsymbol{\beta}'\mathbf{X}_t$ is by the properties of conditional expectation uncorrelated with $\mathbf{X}_t$. So we get the LRM:

$$Y_t = \boldsymbol{\beta}'\mathbf{X}_t + u_t; \;\; t = 1, 2, ..., T. \tag{3}$$

Joint Normal 2

- If they are joint normal with independent sample, the conditional variance $V(Y_t \mid \mathbf{X}_t)$ is constant:

$$E(Y_t - E(Y_t \mid \mathbf{X}_t))^2 = E(u_t^2) = \sigma^2 = \sigma_y^2 - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}. \quad (4)$$

  and the conditional distribution is also normal.

- The distribution for an observation is:

$$\begin{aligned} D_c(Y_t \mid \mathbf{X}_t; \theta) &\sim IN(\beta'\mathbf{X}_t, \sigma^2) \quad (5) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{Y_t - \beta'\mathbf{X}_t}{\sigma}\right)^2\right\} \end{aligned}$$

  and for the whole sample:

$$\begin{aligned} D_c(y \mid X; \theta) &\sim N(X\beta, \sigma^2 I) \quad (6) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}. \end{aligned}$$

Remarks

- Lots of ways of writing normal distribution, e.g.

$$D_c(Y_t \mid \mathbf{X}_t; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left(\frac{u_t^2}{2\sigma^2} \right)$$

- Didn't need to specify independence for the whole sample. The fact that $V(y \mid \mathbf{X}) = \sigma^2 I$ implies that the conditional covariances between Y_t and Y_{t-i} are zero.
- For normally distributed variables zero covariance implies independence, this is not generally the case, independence is a stronger assumption than uncorrelated.

Dependence

- $Var(y \mid X) = E(uu') = \sigma^2\Omega \neq \sigma^2 I$, if the variances are not constant (heteroskedasticity) and/or there is dependence the covariances, are not equal to zero. Under these circumstances, $\widehat{\beta}$ remains unbiased but is not minimum variance (efficient). Its variance-covariance matrix is not $\sigma^2(X'X)^{-1}$, but $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$.
- If $y \sim N(X\beta, \sigma^2\Omega)$ its distribution is given by:

$$2\pi^{-T/2} \mid \sigma^2\Omega \mid^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)'(\sigma^2\Omega)^{-1}(y - X\beta)\right\}.$$

Notice that when $\Omega = I$, then the term in the determinant, $\mid \sigma^2\Omega \mid^{-1/2}$ is just $(\sigma^2)^{-T/2}$ and we get (6).

Bivariate Case

- In the bivariate case:

$$E(Y_t \mid X_t) = \beta_1 + \beta_2 X_t$$

where $\beta_1 = \mu_y - \beta_2 \mu_X$, and $\beta_2 = \sigma_{xy}/\sigma_{xx}$, where σ_{xy} is the covariance of Y_t and X_t and σ_{xx} the variance of X_t.

- We also have

$$E(X_t \mid Y_t) = \gamma_1 + \gamma_2 Y_t$$

where $\gamma_1 = \mu_X - \gamma_2 \mu_y$, $\gamma_2 = \sigma_{xy}/\sigma_{yy}$.

- Also $\beta_2 \gamma_2 = r^2$, the squared correlation coefficient.
- These two conditional expectations need not have a causal interpretation. If Y_t was height and X_t weight; we can both ask (a) what is the expected weight of someone of height Y and (b) what is the expected height of someone of weight X.
- For prediction we do not need exogeneity assumptions but for causal statements we do.

ML estimation introduction

- Given a RV y with pdf, $f(y, \theta)$, where θ is a parameter, we can use $f(y, \theta)$ to give the probability of particular values of y, given θ.
- For instance, given a coin with probability heads: $\theta = 0.5$, what is the probability of observing 10 heads in a row? Answer $(0.5)^{10}$.
- Alternatively, we can use $f(y, \theta)$ to tell us the likelihood of particular values of θ given a particular sample say $y_1, y_2, .., y_T$. If we observe ten heads, how likely is it that $\theta = 0.5$? Again answer $(0.5)^{10}$.
- The first case interprets $f(y, \theta)$ as a function of y given θ and we call it a pdf. The second case interprets $f(y, \theta)$ as a function of θ given y and we call it a likelihood.
- The maximum likelihood (ML) procedure estimates $\widehat{\theta}$ as the value most likely to have given the observed sample. $\theta = 0.5$ is very unlikely to have generated the observed sample of 10 heads. $\theta = 1$ is more likely.

Likelihood function

- For a random sample with independent observations, we can just multiply the pdfs for each observation together as we did in the coin example and write the Likelihood as:

$$L(\theta) = f(y_1, \theta) f(y_2, \theta) ... f(y_T, \theta)$$

- We then choose θ that maximises this value for our observed sample $y_1, y_2, ..., y_T$.
- It is more convenient to work with the logarithm of the likelihood function. Since logs are a monotonic function the value of θ that maximises the log-likelihood will also maximise the likelihood. Thus the log-likelihood is:

$$LL(\theta) = \sum_{t=1}^{T} \log f(y_t, \theta).$$

Maximising the Likelihood function

- $LL(\theta)$ is a scalar, suppose θ is a $k \times 1$ vector. To find the maximum we take the k derivatives of $LL(\theta)$, this is called the score vector and set them to zero:

$$S(\widehat{\theta}) = \frac{\partial LL(\widehat{\theta})}{\partial \theta} = \frac{\partial \sum \log f(y_t, \widehat{\theta})}{\partial \theta} = 0$$

  then solve for the value of $\theta, \widehat{\theta}$ that makes the derivatives equal to zero.

- For simple examples, like the LRM we can solve these equations analytically, for more complicated examples we solve them numerically.

Second order conditions

- To check that we have found a maximum, we need to calculate the kxk matrix of second derivatives:

$$\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}.$$

  For a maximum this matrix should be negative definite.

- The information in observation t is the negative of the expected value of the matrix of second derivatives:

$$I_t(\theta) = -E(\frac{\partial^2 LL_t(\theta)}{\partial\theta\ \partial\theta'})$$

  which is a symmetric $k \times k$ matrix.

- The average information matrix in the sample of size T is:

$$I_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} I_t(\theta) = -E(\frac{1}{T}\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}).$$

Properties of the ML estimator 1

- A useful result is that for any unbiased estimator (in small samples) or consistent estimator (asymptotically when $T \rightarrow \infty$) the inverse of the information matrix provides a lower bound (the Cramer-Rao lower bound) on the variance covariance matrix of the estimator

$$V(\widehat{\theta}) \geq I(\widehat{\theta})^{-1}.$$

where $\geq$ indicates that $V(\widehat{\theta}) - I(\widehat{\theta})^{-1}$ is a non-negative definite matrix.

- Under certain regularity conditions the ML estimator $\widehat{\theta}$ is consistent, that is for some small number $\epsilon > 0$

$$\lim_{T \rightarrow \infty} \Pr(|\ \widehat{\theta}_T - \theta\ | > \epsilon) = 0.$$

- When we evaluate asymptotic distributions we look at $\sqrt{T}(\widehat{\theta} - \theta)$ as $T \rightarrow \infty$, because since it is consistent the distribution of $\widehat{\theta}$ collapses to a point and scale the information matrix by T.

Properties of the ML estimator 2

- The ML estimator is asymptotically normally distributed and asymptotically attains the Cramer-Rao lower bound (i.e. it is efficient),

$$\begin{aligned} \sqrt{T}(\widehat{\theta}_T - \theta) &\rightarrow N(0, I(\theta)^{-1}), \\ I(\theta) &= \lim_{T \rightarrow \infty} -E\left(\frac{1}{T}\frac{\partial^2 LL(\theta)}{\partial\theta\,\partial\theta'}\right). \end{aligned}$$

- The scaled score $(\sqrt{T})^{-1}S(\theta)$ is also asymptotically normal $N(0, I(\theta))$. We will use these two asymptotic normality properties in testing.
- In addition, $E(S(\theta)S(\theta)') = T \times I(\theta)$.
- ML estimators are also invariant in that for any function of θ, say $g(\theta)$, the ML estimator of $g(\theta)$ is $g(\widehat{\theta})$. Partly because of this ML estimators are not necessarily unbiased. Some are, many are not.

Non-linear estimation

We distinguish

- (1) equations which are non-linear in variables because of transformations, like logarithms or powers, but which can be estimated by a linear regression on the transformed data and
- (2) equations which are non-linear in parameters, where we need a non-linear estimation routine.

An S shaped function Linear in parameters

If our dependent variable is a proportion, p_t taking values between zero and one, the logistic transformation is often used $\ln(p_t/(1-p_t))$, the logit of p_t. If this is made a function of time,

$$\ln\left(\frac{p_t}{1-p_t}\right) = a + bt + u_t$$

this gives an S shaped curve for p_t over time, which often gives a good description of the spread of a new good (e.g. the proportion of the population that have a mobile phone) and can be estimated by least squares, since it is linear in the parameters. The form of the relationship is

$$p_t = \frac{1}{1 + \exp -(a + bt + u_t)}$$

An S shaped function non-linear in parameters

- We could also set it up as inherently non-linear.
- To estimate a logistic with a saturation level: $p_t = N_t / K$, where N_t is the number of mobile phone owners and K is the saturation level we could estimate

$$N_t = \frac{K}{1 + \exp -(a + bt)} + \varepsilon_t$$

  directly by non-linear least squares.
- Notice the assumption about the errors is different. In the previous case the error was additive in the logit, here it is additive in the number. In practice, unless the market is very close to saturation it is difficult to estimate K precisely.
- Most programs will estimate such non-linear models using an iterative method to find the minimum of the sum of squared residuals or the maximum of the likelihood functio

Non-linear estimation

- For the LRM with normal (Gaussian) errors considered below, the ML estimator has a closed form solution. For many of the models we will consider (MA errors, GARCH, Johansen) this is not the case. In such cases one typically requires some sort of iterative procedure to obtain estimates.
- If $F(\boldsymbol{\theta})$ is the likelihood function starting from some initial guesses, $\boldsymbol{\theta}_0$ the estimates are updated as

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \lambda_i \boldsymbol{\Delta}_i \tag{7}$$

  where λ_i is the step size, in iteration i, and $\boldsymbol{\Delta}_i$ the direction and this continues until it converges to a maximum: you get to the top of the hill.
- The most commonly used algorithms are gradient methods..Define the gradient and Hessian

$$\mathbf{g} = g(\boldsymbol{\theta}) = \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}; \ \mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}.$$

Gradient methods

- The simplest gradient method is Newton's method based on a linear Taylor series expansion around $\boldsymbol{\theta}_0$

$$\begin{aligned} \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &\simeq \mathbf{g}_0 + \mathbf{H}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \\ \boldsymbol{\theta} &\simeq \boldsymbol{\theta}_0 - \mathbf{H}_0^{-1}\mathbf{g}_0. \end{aligned}$$

  In (7) this sets $\lambda_i = 1$ and $\boldsymbol{\Delta}_i = \mathbf{H}_i^{-1}\mathbf{g}_i$. This often works well, but may be improved by adjusting λ_i.

- It may be difficult to calculate $\mathbf{H}_i^{-1}$ and it may not be positive definite. In ML examples the outer product gradient, OPG, method uses $\left[\sum_{t=1}^{T} \mathbf{g}_i \mathbf{g}_i'\right]^{-1}$ instead of $(-H)^{-1}$. This is always positive definite and only requires calculating first derivatives. It is the basis of BHHH, Berndt, Hall, Hall & Hausman.

Issues

- Where to Start? Try to choose sensible initial values, $\boldsymbol{\theta}_0$, e.g. based on linear approximations and try different values to check for local maxima.
- How to climb up hill?Climb depends on choice of λ_i and $\boldsymbol{\Delta}_i$, programs will often switch between procedures. You can choose between 4 in Stata for GARCH.
- When to stop?: determining whether it has converged to a maximum. $\mathbf{g}_i < \varepsilon$, .and $F_i - F_{i-1} < \varepsilon$ are sensitive to scaling, the units the variables are measured in, $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ is less sensitive.
- In the linear case if the parameter is not identified because $X'X$ is singular, it will be obvious. It may not be obvious in the non-linear case and the program may provide estimates even if the likelihood is very flat. This may occur if one has not identified the right sort of non-linearity.
- For less well behaved functions there are algorithms like simulated annealing and genetic algorithms.

Remarks

- ML is a very general procedure which can be used in a wide variety of circumstances as long as we can specify a distribution for the random variables.
- The distribution does not have to be normal, a fat tailed distribution like t is often better for financial data, Poisson for count data.
- Method of moment estimators like GMM, generalised MoM, do not require distributional assumptions.
- Sometimes when the distributional assumption does not hold ML estimators are inconsistent, but sometimes quasi maximum likelihood estimators do well even when the distributional assumption does not hold.

Next time

- Next we will look at a special case, the LRM with normal errors.
- Set up the likelihood function for $\theta = \beta, \sigma^2$, take derivatives, set them to zero and solve for the ML estimators
- Take second derivatives, get the information matrix.
- Look at the maximised log likelihood function.
- Look at transformations