Maximum Likelihood

Maximum Likelihood Estimation

Intuition

- As well as least squares there are other forms of estimation procedure: two most popular alternative estimation methods in econometrics are method of maximum likelihood and method of moments.
- Imagine two possible outcomes, 1 and 0 where the probablility of obtaining 1 is π and the probability of 0 is $(1 - \pi)$.
- Take a random sample of values of size n. Suppose that $n = 5$ and that the sample is $(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1)$. What is the *most likely* value of pi to have generated this sample?

- Take another random sample of values of size 5, $(y_1 = 0, y_2 = 1, y_3 = 1, y_4 = 0, y_5 = 1)$. What is the *most likely* value of *pi* to have generated this sample?
- The intuition behind these simple questions is the intuition behind maximum likelihood, i.e. *what is the most likely value of the parameter to have generated the observed sample.*

Simple Example: take 1

- In the previous example the sample is $(y_1 = 0, y_2 = 1, y_3 = 1, y_4 = 0, y_5 = 1)$. We need to work out the most likely π to have generated this data.
- The probability of this sample being generated given our assumptions is:

$$(1 - \pi) \cdot \pi \cdot \pi \cdot (1 - \pi) \cdot \pi$$

- Suppose that π was 0.1, then the probability of obtaining our sample in a random experiment would be:

$$(1 - 0.1) \cdot 0.1 \cdot 0.1 \cdot (1 - 0.1) \cdot 0.1 = 0.00081$$

- However, if π was 0.2, this probability would be:

$$(1 - 0.2) \cdot 0.2 \cdot 0.2 \cdot (1 - 0.2) \cdot 0.2 = 0.00512$$

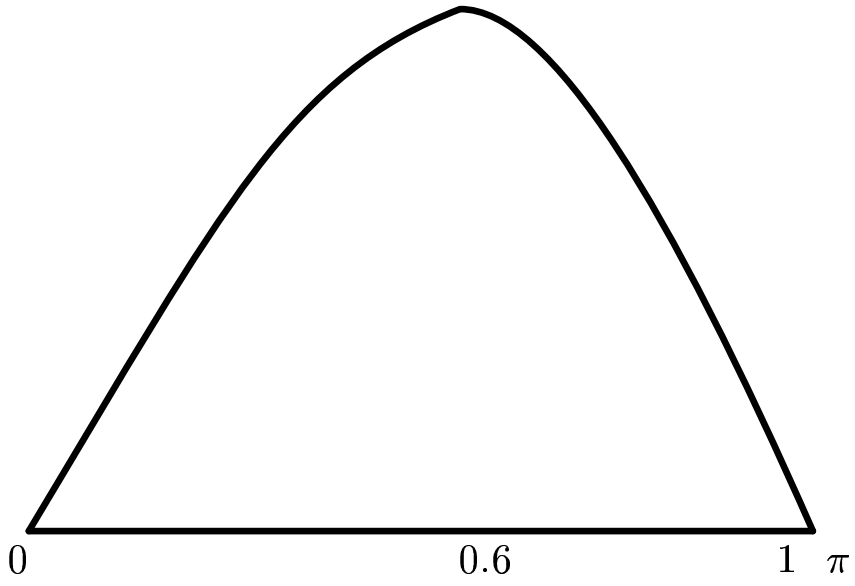
  Thus, it is *more likely*, given our sample, that π is 0.2 than 0.1.

- Continuing this for all values of π gives:

Value of x	Prob of obtaining sample
0	0
0.1	0.0008
0.2	0.005
0.3	0.13
0.4	0.023
0.5	0.03
0.6	0.035
0.7	0.03
0.8	0.02
0.9	0.007
1	0

- Thus the maximum likelihood estimate of π is 0.6.

- We can plot this:



Simple Example: take 2

- In the previous example the sample is $(y_1 = 0, y_2 = 1, y_3 = 1, y_4 = 0, y_5 = 1)$. We need to work out the most likely π to have generated this data. Consider the sample one by one. The probability of the first data point, $y_1 = 0$, occuring is $(1 - \pi)$.
- One way to see this is that the probability of obtaining any value, X, is given by

  $$\pi^X (1 - \pi)^{(1-X)}$$

  Inserting $X = 1$ into this formula gives the result π. You can think of this formula as the *density function* for the distribution represented in this example.

- Now consider data point $y_2 = 1$. Inserting this into the formula gives $(1 - \pi)$. We can continue for each data point in our sample and work out the probability of each value from the density function.
- We can then work out the *joint density* of the sample by multiplying all of these values together. This is the probablity that our sample would arise in any random experiment. In our case, we have:

$$\left(\pi^{y_1}(1-\pi)^{(1-y_1)}\right) \cdot \left(\pi^{y_2}(1-\pi)^{(1-y_2)}\right) \cdot \ldots$$

- Thus we have:

  $$\pi^3(1-\pi)^2$$

  as our joint density function for our sample. We now simply find the π that will maximise this function. We can do this by tabulating/plotting or by differentiating with respect to π and finding the stationary point.

- Tabulation gives:

Value of x	$\pi^3(1-\pi)^2$
0	0
0.1	0.0008
0.2	0.005
0.3	0.13
0.4	0.023
0.5	0.03
0.6	0.035
0.7	0.03
0.8	0.02
0.9	0.007
1	0

ML Estimation of the normal linear model

- We want to estimate α and β in:

  $$y_t = \alpha + \beta x_t + u_t$$

  where the u_t are assumed to be normally distributed. We have a sample of size n of y_t and x_t. It is easier to think of this as:

  $$u_t = y_t - \alpha - \beta x_t$$

  which is now a normally distributed variable.

- We need to find the α and β that are most likely to have generated this sample.

- The density function of each of the normally distributed u_t is

  $$f(u_t) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2\sigma^2}(u_t)^2\right]$$

- This may be written as:

$$f(y_t - \alpha - \beta x_t) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2\sigma^2}(y_t - \alpha - \beta x_t)^2\right]$$

- However, we have a sample of size n so in parallel with our earlier example, we need to find the *joint* density and then maximise it. To do this we need to multiply the individual density functions for each observation together as we did before. This gives:

$$L = \prod_{t=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2\sigma^2}(y_t - \alpha - \beta x_t)^2\right]$$

  This is the *likelihood function* for the sample.

- To maximise this, it is convenient to take logs (convince yourselves that you know why). This gives (on simplification):

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{\sum(y - \alpha - \beta x_t)^2}{2\sigma^2}$$

- We need to choose the α and β that maximise this log-likelihood function. Notice that only the final term contains α and β so the other terms will drop out of the differentiation. Thus, finding the ML estimates is the same as maximising the final term, or minimising:

$$\sum \frac{(y_t - \alpha - \beta x_t)^2}{2\sigma^2}$$

- Note finally that this will give the same estimators as the linear OLS regression model. Thus in the case of the linear model, OLS and ML estimates of α and β are equivalent.

- However, the ML estimate of σ^2 in *not* the same as the OLS estimate. To see this, we maximise the log-likelihood with respect to σ.

- Differentiating gives:

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum (y_t - \alpha - \beta x_t)^2}{\sigma^2}$$

Setting equal to zero, we get the estimator for σ^2,

$$\hat{\sigma}^2 = \frac{\sum (y_t - \hat{\alpha} - \hat{\beta} x_t)^2}{n} = \frac{\sum \hat{u}_t^2}{n}$$

This is not the same as:

$$\hat{\sigma}^2_{\text{OLS}} = \frac{\sum \hat{u}_t^2}{n - k}$$

where k is the number of regressors. In fact, the ML estimator of the variance is *biased* but is *asymptotically unbiased*. As $n \to \infty$, $n - k$ and n become equivalent so the estimators are the same.

Logit/Probit Models

- A class of models which are typically estimated using ML methods are *limited dependent variable models.* Often these are models in which the dependent variable only takes on the values 1 and 0. Examples are trade union membership, voting, smoking.
- OLS methods are not advised because there will be heteroskedasticity and the predictions will often lead to predicted values of y which do not make sense. (See Carter Hill *et al* pp. 369-370.
- Recall from above that for a random variable, y, which takes the value 1 or 0, the density function is:

  $$f(y) = \pi^y (1-\pi)^{1-y} \quad y = 0, 1$$

  Logit and probit models allow us to estimate the probability π in various ways.

- Recall that the expected value of a variable is equal to the (distributed) sum of values multiplied by probabilities. In this case:

$$E(y) = 0 \cdot (1 - \pi) + 1 \cdot \pi = \pi$$

- We break the dependent variable into a bit that can be explained and a part that cannot. The above is the explained part. In the logit model, we assume that this explained part is related to the regressors in a *non-linear* way using a function known as the *logistic* function.

$$E(y) = F(\alpha + \beta x)$$

where F is the logistic function:

$$F(l) = \frac{1}{(1 + e^{-l})}$$

- This function ensures that the predictions in this model are between 1 and 0 which is what we require for a probablity model.

- How can we interpret the results from such an estimation. In general, we want to know the increase in probability resulting from a particular regressor being present (e.g. does being male increase the probability of being a trade union member? If so by how much?)

- Thus we are interested in the derivative:

$$\frac{dE(y)}{dx}$$

where x is some regressor.

- Since, $E(y) = F(\alpha + \beta x)$, we need to use the chain rule of differentiation:

$$\frac{dE(y)}{dx} = F'(\alpha + \beta x)\beta$$

which gives:

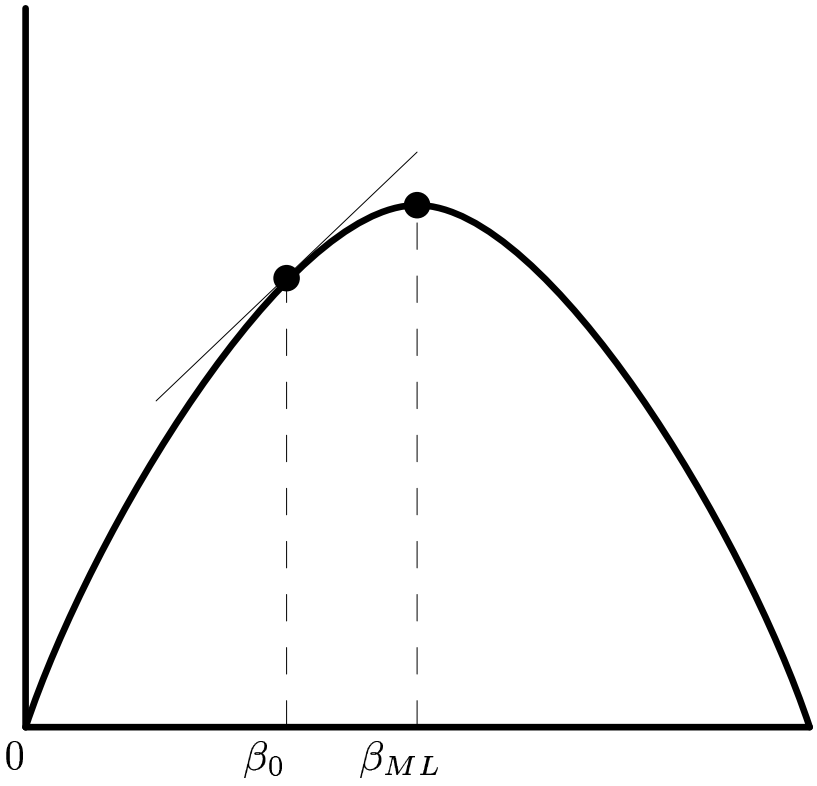
$$\frac{dE(y)}{dx} = \frac{e^{-(\alpha+\beta x)}}{(1 + e^{-(\alpha+\beta x)})^2}\beta$$

The first term here is reported in *Microfit* as the 'factor for calculation of marginal effects'

LM, Wald and LR testing

- These tests, typically thought of as *large sample tests* are motivated by the ML procedure. They are general testing methods applicable in a much wider range of cicumstances than tests such as the F-test (e.g nonlinear restrictions).
- Recall that when testing restrictions, we have an *unrestricted model* and a *restricted model*.
- Take the simple case where we want to test one restriction: $\beta = \beta_0$.
- The *unrestricted* ML estimate of β (call it β_{ML}) is the one formed by maximising the log-likelihood function. The *restricted* value of β is β_0. (Note that the unrestricted value will always be bigger than or equal to the restricted estimate. Why?)
- The LM test is based on the *restricted model*. The idea begind the test is that the slope of the log-likelihood will be zero at the (unrestricted) maximum. Thus if the restrictions are valid, β_{ML} will be 'close' to β_0 and the slope at the restricted estimate will also be close to zero.

- We can see this in a diagram:



- Thus, the LM test takes the slope of the log-likelihood at the restricted estimate, squares it and divides it by (the negative of the expectation of) the second derivative of the log-likelihood (a measure of curvature of the log-likelihood also known as the *information matrix*). Thus the general form for an LM test is:

$$LM = \left(\frac{dL}{d\beta}\right)^2 \frac{1}{d^2L/d\beta^2}$$

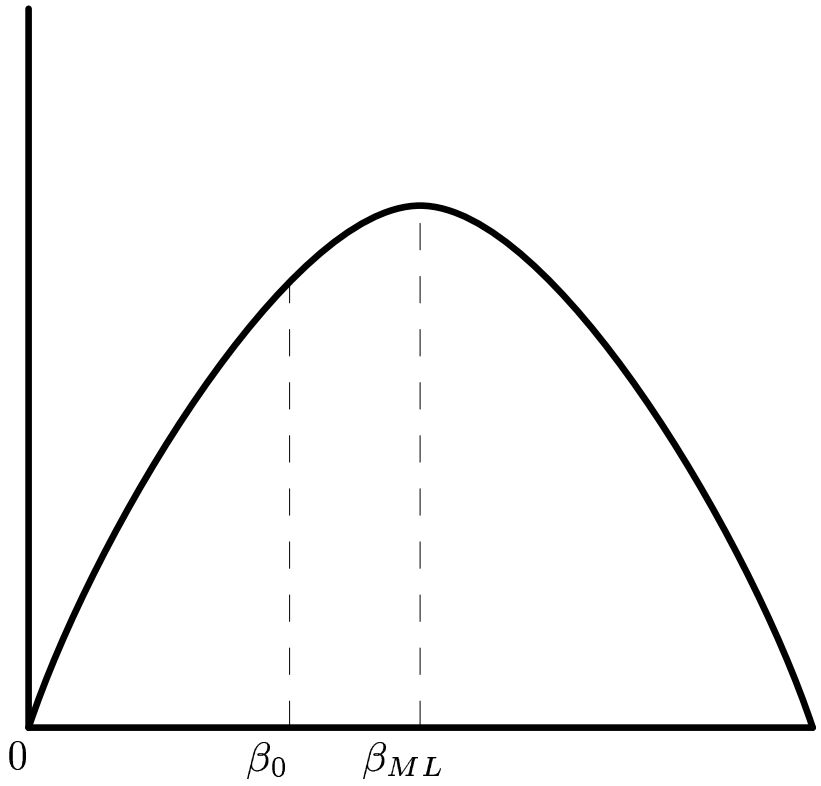
  which can be shown to be (asymptotically) chi-square distributed with r degrees of freedom (in this case $r = 1$).

- Note that since we are comparing the slope to zero, we only need to estimate the *restricted* model to use an LM test. This is one of the reasons why an LM test can often be useful. You only need to estimate the simpler of the two models.

- The Wald test is very similar to the LM. It tests the distance between β_0 and β_{ML}. It also uses the expected curvature as a divisor. The test is:

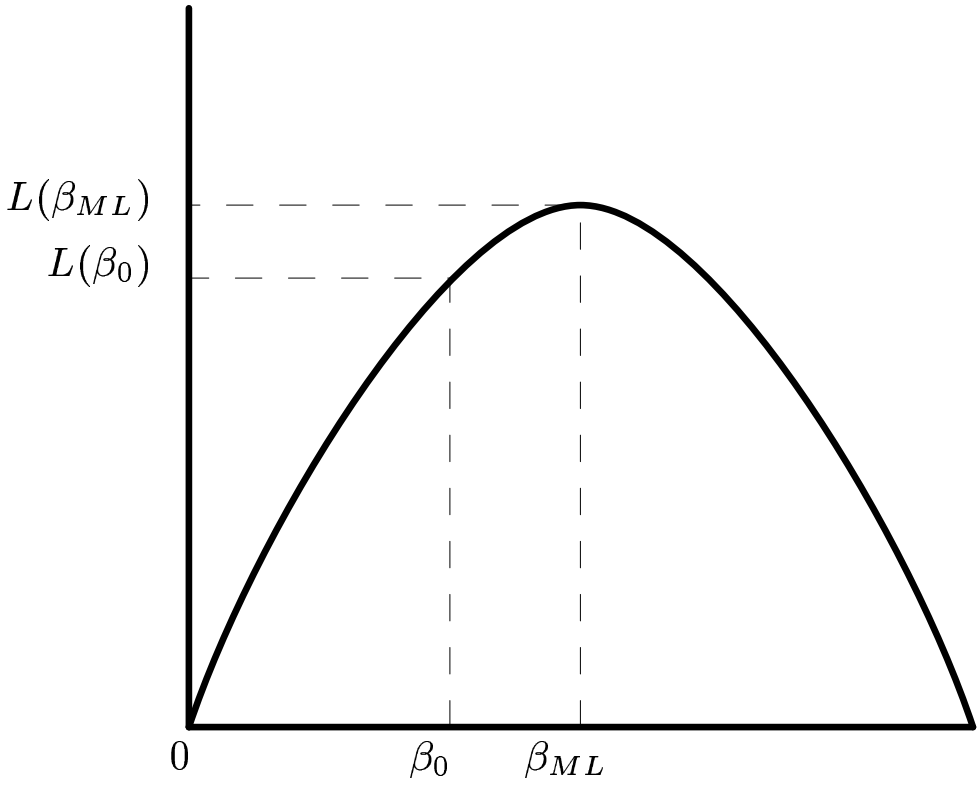
$$W = (\beta_0 - \beta_{ML})^2 \frac{1}{d^2L/d\beta^2}$$

This test is also (asymptotically) distributed as $\chi^2(r)$.



- Note that the W test requires estimation of the *unrestricted* model only. We know the restricted value β_0 so we don't need to estimate the restricted model to find it.

- Finally, the LR test looks at the difference in the values of the log-liklihoods at the restricted and the unrestricted estimates. It thus requires estimation of *both* models.



- The LR test is formed using the following:

  $$2(logL(\beta_{ML}) - logL(\beta_0))$$

  It has the same (asymptotic) distribution as the other two tests above.