Econometrics, Lecture 4.
ML estimation of the LRM & transformations

Ron Smith
EMS, Birkbeck, University of London

Autumn 2020

Last time

- Looked at the interpretation of regression as a conditional expectation
- Looked at the joint normal case where the conditional expectation of y was a linear function of X.
- Looked at the general principles of ML
- We now apply those principles to the special case of the LRM with a normal distribution. Remember ML can be used with any distribution not just normal.
- Look at model selection and transformations.

Likelihood function

- For an independent sample the Likelihood is the product of the pdfs:

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

- For an observation on $u_t \sim IN(0, \sigma^2)$

$$f(u_t, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{u_t^2}{2\sigma^2}\right\} = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{u_t^2}{2\sigma^2}\right\}$$

- For the sample

$$\begin{aligned} \prod f(u_t, \sigma^2) &= (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{\sum u_t^2}{2\sigma^2}\right\} \\ L(\beta, \sigma^2) &= (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{u'u}{2\sigma^2}\right\} \end{aligned}$$

Likelihood function for LRM with normal errors

- For the normal LRM, the likelihood of the sample is given by the product of normal pdf, but now interpreted as a function of the $k+1$ vector $\theta = (\beta, \sigma^2)$, the unknown parameters:

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

- Noting $u'u = \sum u_t^2 == (y - X\beta)'(y - X\beta)$, the Log-likelihood function is :

$$LL(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}u'u. \qquad (1)$$

- To find the estimates that maximise this we differentiate it with respect to (w.r.t) β and σ^2 and set the derivatives equal zero.

Derivatives beta

- Maximising (1) w.r.t β is the same as minimising $u'u$, as before:

$$u'u = (y - X\beta)'(y - X\beta) = y'y + \beta'X'X\beta - 2\beta'X'y.$$

$$LL(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}u'u.$$

- The score (first derivative) vector for vector β is:

$$\frac{\partial LL(\beta, \sigma^2)}{\partial \beta} = -\frac{1}{2\sigma^2}(2X'X\beta - 2X'y) \tag{2}$$

First order conditions, FOC, beta

- Setting (2) equal zero gives one FOC

$$\begin{aligned} -\frac{1}{2\widehat{\sigma}^2}(2X'X\widehat{\beta} - 2X'y) &= 0 \\ \frac{1}{\widehat{\sigma}^2}(X'y - X'X\widehat{\beta}) &= 0 \end{aligned} \tag{3}$$

  $\widehat{\beta}$ is the value that make the FOCs=0.

- Solving (3) for $\widehat{\beta}$ gives

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

  if X is of full rank so $(X'X)^{-1}$ exists.

- (3) can be written

$$\frac{1}{\widehat{\sigma}^2}X'(y - X\widehat{\beta}) = \frac{1}{\widehat{\sigma}^2}X'\widehat{u} = 0 \tag{4}$$

So as before $X'\widehat{u} = 0$.

Derivatives sigma squared

Note σ^2 is a scalar and we are taking the derivative of (1) w.r.t σ^2, not σ.

$$LL(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}u'u.$$

- We use
$$\frac{\partial \log(\sigma^2)}{\partial \sigma^2} = \frac{1}{\sigma^2}$$
and $-1/2\sigma^2 = -(2\sigma^2)^{-1}$ with derivative $(-1)(-(2\sigma^2)^{-2})$.

$$\frac{\partial LL(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}u'u. \tag{5}$$

First order conditions sigma squared

Setting (5) equal to zero gives

$$-\frac{T}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\widehat{u}'\widehat{u} = 0$$

multiply through by $2\widehat{\sigma}^4$

$$-T\widehat{\sigma}^2 + \widehat{u}'\widehat{u} = 0$$

so our maximum likelihood estimator of the variance is:

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{T}.$$

The ML estimator is biased and we usually use the unbiased estimator $s^2 = \widehat{u}'\widehat{u}/(T-k)$.

Second derivatives

To check second order conditions and construct the information matrix, we take the derivative wrt β of the first derivative wrt β

$$\frac{\partial LL(\beta,\sigma^2)}{\partial\beta} = \frac{1}{\widehat{\sigma}^2}(X'y - X'X\widehat{\beta}) = \frac{1}{\widehat{\sigma}^2}X'\widehat{u}$$

$$\frac{\partial^2 LL(\beta,\sigma^2)}{\partial\beta\ \partial\beta'} = -\frac{1}{\sigma^2}X'X \tag{6}$$

and cross partial wrt σ^2

$$\frac{\partial^2 LL(\beta,\sigma^2)}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}X'u.$$

Notice the derivative of $(\sigma^2)^{-1}X'u$ w.r.t σ^2 is $-(\sigma^2)^{-2}X'u$.

Second derivative wrt sigma squared

First derivative is

$$\begin{aligned}\frac{\partial LL(\beta, \sigma^2)}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} u'u \\ &= -\frac{T}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2} u'u \left(\sigma^2\right)^{-2}\end{aligned}$$

Second derivative is

$$\begin{aligned}\frac{\partial^2 LL(\beta, \sigma^2)}{\partial (\sigma^2)^2} &= \frac{T}{2}\left(\sigma^2\right)^{-2} - 2\left(\frac{1}{2} u'u \left(\sigma^2\right)^{-3}\right) \qquad (7) \\ &= \frac{T}{2\sigma^4} - \frac{u'u}{\sigma^6}.\end{aligned}$$

Collecting terms

$\theta = (\beta, \sigma^2)$

$$\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'} = \left[\begin{array}{cc} -\frac{1}{\sigma^2} X'X & -\frac{1}{\sigma^4} X'u \\ \left(-\frac{1}{\sigma^4} X'u\right)' & \frac{T}{2\sigma^4} - \frac{u'u}{\sigma^6} \end{array} \right]$$

This is a $(k+1) \times (k+1)$ matrix, $X'u$ is $(k \times 1)$

Information matrix

- To get $I(\theta)$ we take the negative of the expected value of the second derivative matrix.
- $E(X'u) = 0$, so the off diagonal terms are zero
- $E(u'u) = T\sigma^2$ so the expected value of (7) is

$$\frac{T}{2\sigma^4} - \frac{T\sigma^2}{\sigma^6} = \frac{T}{2\sigma^4} - \frac{T}{\sigma^4} = -\frac{T}{2\sigma^4}$$

- Putting the bits together using (6)

$$I(\theta) = I(\beta, \sigma^2) = -E(\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}) = \left[\begin{array}{cc} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{T}{2\sigma^4} \end{array} \right]$$

Cramer Rao Lower Bound

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

$$I(\beta, \sigma^2)^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}.$$

- This gives the lower bound for the Variance-covariance matrix for estimators of β, σ^2.
- Second derivatives measure curvature, so sharp peaks with high curvature are very precisely estimated, have small variances.
- The estimators of β and σ^2 are independent, their covariances are zero. But there will be non-zero covariances between the elements of $\widehat{\beta}$.
- Same answer for $V(\widehat{\beta})$ as before.

Maximised log likelihood

We can put the ML estimates into the Log-likelihood function, to get the Maximised Log-Likelihood, MLL, reported by most programs

$$\begin{aligned} MLL &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}\widehat{u}'\widehat{u} \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{T\widehat{\sigma}^2}{2\widehat{\sigma}^2} \\ &= -\frac{T}{2}(\log(2\pi)+1) - \frac{T}{2}\log(\widehat{\sigma}^2) \end{aligned}$$

apart from the constant this is just the negative of half the sample size times the log of the ML estimate of the variance. This can be negative or positive.

Model selection

- The MLL has no interpretation in itself but it can be used in testing between models, like the Likelihood Ratio test, and in choosing the ‘best’ model on some ‘model selection’ criterion.
- R^2 and $\overline{R}^2$ are popular, but unreliable, model selection methods. AIC and BIC are better.
- Suppose MLL_i is the maximised log likelihood of model i, k_i is the number of parameters estimated in model i, and T is the sample size.
- The Akaike Information Criterion is

$$AIC_i = MLL_i - k_i.$$

- The Schwarz Bayesian Information Criterion or Posterior Odds Criterion

$$BIC = MLL_i - 0.5k_i \log T$$

- BIC tends to choose a more parsimonious model (fewer parameters) than the AIC.

Model selection

- About half of statistics programs (including Microfit) define AIC and BIC as above, in which case you choose the model with the largest value.
- The other half (including EViews and Stata) use-2 times these values in which case you choose the model with the smallest value.
- Be careful, which way they are defined.
- Later we will often use AIC and BIC to choose models, e.g. different number of lags.

Functional form

- A crucial feature of the model is the choice of functional form.
- As noted before we distinguish
  - (1) equations which are non-linear in variables because of transformations, but 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 routines.
- We now look at some models that are linear in parameters.
- Interpret coefficients as derivatives or elasticities, which may not be constant, but a function of the variables.
- If you transform the dependent variable, measures of fit are not comparable.
- Most common transformation is logarithmic, usually to base e.
  - We saw across countries life expectancy is a linear function of log per-capita income
  - Apart from the very young, the log of the risk of dying in a year, is a linear function of age.

Logarithmic transformations are common since

1. prices and quantities are non-negative so the logs are defined
2. the coefficients can be interpreted as elasticities, % change in the dependent variable in response to a 1% change in the independent variable, so the units of measurement of the variables do not matter
3. in many cases errors are proportional to the variable, so the variance is more likely to be constant in logs,
4. the logarithms of economic variables are often closer to being normally distributed
5. the change in the logarithm is approximately equal to the growth rate and
6. lots of interesting hypotheses can be tested in logarithmic models.
7. often effects are proportional, which is captured by logarithmic models.

Interpreting regression coefficients with transformed data

Linear β measures ΔY_t resulting from a unit change in X_t : $\Delta X_t = 1$

$$Y_t = \alpha + \beta X_t + u_t,$$

corresponds to the derivative

$$\frac{\partial Y_t}{\partial X_t} = \beta.$$

depends on the units that the variables are measured in.
The elasticity is the percentage change in Y_t resulting from a 1% change in X_t.

$$\eta = \frac{\partial Y_t / Y_t}{\partial X_t / X_t} = \frac{\partial \log Y_t}{\partial \log X_t}$$

It is invariant to units, but in the linear case depends on where measured

$$\eta = \frac{\beta X_t}{Y_t}$$

and is different at every point on the line.

Log-Log

- In a logarithmic regression

$$\log Y_t = \alpha + \beta \log X_t + u_t$$

  then β is the elasticity.

- Multiply the SER by 100 to get percentage error.
- To get a rough comparison of fit. Divide the SER of the linear model by the mean of the dependent variable (assuming it is positive and non-zero) to give a proportionate error.
- Suppose we have a dummy variable in the equation

$$\log Y_t = \alpha + \beta \log X_t + \gamma D_t + u_t$$

  where $D_t = 0$ or $D_t = 1$. The effect on Y_t of the dummy variable going from zero to one is $\exp(\gamma) - 1$.

Percentages

- Suppose Y_t and X_t are both percent, e.g. Y_t is the % inflation rate, and X is the % unemployment rate.
- Then β measures the percentage **point** change in Y_t in response to a one percentage **point** change in X_t. If unemployment rises from 1% to 2%, it increases by one percentage point and 100%.
- Non linear Phillips Curves allow the effect on inflation of a one percentage point change in unemployment to be greater when unemployment is 1% than when it is 9%. We can represent this by using the reciprocal of unemployment

$$\begin{aligned} Y_t &= \alpha + \beta X_t^{-1} + u_t, \\ \frac{\partial Y_t}{\partial X_t} &= -\beta X_t^{-2}. \end{aligned}$$

So at 1% unemployment the effect is just $-\beta$, but at 9% unemployment it is $-\beta/9^2$, very small.

Semi-log

One may not log both dependent and independent variables one might regress life expectancy in years on log per-capita income (which is what Gapminder does), i.e. an equation, sometimes called linear-log, of the form

$$Y_t = \alpha + \beta \log X_t + u_t$$

If X_t changes by 1%, then $\Delta X_t / X_t = 0.01$, so Y_t changes by 0.01β.
One can have it the other way round a log-linear model

$$\log Y_t = \alpha + \beta X_t + u_t.$$

Here a unit change in X_t, $\Delta X_t = 1$, causes a $100\beta\%$ change in Y_t.

Next time

- So far we have focussed on estimation, now we turn to inference, testing.
- We want to test individual hypotheses such as whether a single coefficient is different from zero and joint hypotheses about a number of coefficients.
- The hypotheses may be substantive, like constant returns to scale in a production function, or diagnostic, whether there is serial correlation in the residuals.
- The tests may be exact, small sample, like t or F tests or asymptotic, holding as the sample size gets large, normal or chi-squared tests.
- We will revise distributions related to the normal.