Likelihood - OLS

Nathaniel Beck

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

QII - Week 3

MLE and OLS

- The standard setup of OLS looks different than what we have just seen,
- but we can easily renotate linear models to fit what we have just done.
- The new notation is

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}) \tag{1}$$

- Assume that $\boldsymbol{\Omega}$ meets the Gauss-Markov assumptions
- Note that we are here assuming that the $\mathbf{y}$ are normally distributed, which we didn't do previously
- But as we shall see, either normality is benign or OLS isn't right.

The ugly calculus

- The derivation of the ML estimation of the $\boldsymbol{\beta}$ is easy.
- The observations are independent
- for a univariate normal (y_i), we have that

$$y_i \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2) \tag{2}$$

- or, for each observation

$$L_i = f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}} \tag{3}$$

$$\log(L_i) = \log(f(y_i) =$$

$$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 \tag{4}$$

so for all observations

$$\log(L) = \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 \tag{5}$$

or, putting it together in one big matrix

$$\log(L) =$$

$$-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{6}$$

The scalar setup

- We then take derivatives

$$\frac{\partial \log(L)}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2} \sum_{i=1}^{N} -(y_i - \mathbf{x}_i \boldsymbol{\beta})\mathbf{x}_i' \tag{7}$$

- This derivative is zero when

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\beta \tag{8}$$

  which are the K OLS normal equations.
- This shows the intimate tie between normality and least squares, and either least squares justifies ML or vice versa

- The calculus to find the value of $\hat{\sigma}^2$ which sets the appropriate derivative to zero is slightly more tedious
- Computing at the ml estimate $\hat{\boldsymbol{\beta}}$ with e_i being the usual OLS residual

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} e_i^2 \tag{9}$$

- which yields when the derivative is set to zero

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N} \tag{10}$$

- Note this differs from OLS formula by dividing by N not $N - K$
- MLE is Not necessarily unbiased

The matrix setup

- Working with Equation 6, we get

$$\frac{\partial \log(L)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) \tag{11}$$

  which when set to zero yields the normal equations.
- For estimating σ^2, we use

$$\frac{\partial \log(L)}{\partial \sigma^2} = \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{12}$$

  which when set to zero yields the same estimator as in the scalar case.
- We can get the VCV matrix by taking the $K \times K$ Hessian
- and then taking the negative of the expectation and inverting (invert information matrix).
- The Hessian is

$$\begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & -\frac{1}{\sigma^4}\mathbf{X}'\boldsymbol{\varepsilon} \\ -\frac{1}{\sigma^4}\boldsymbol{\varepsilon}'\mathbf{X} & \frac{N}{2\sigma^4} - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix} \tag{13}$$

- On expectation, the two off diagonal terms are zero
- ($\mathbf{X}$ and ε are uncorrelated)
- and we have already have the relationship of the expectation of $\varepsilon'\varepsilon$ and σ^2
- so we are inverting a diagonal matrix.
- Thus the VCV matrix of the estimates is

$$\begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{N} \end{bmatrix} \tag{14}$$

- Note what this shows:
- the estimate $\hat{\boldsymbol{\beta}}$ is independent of the estimate $\hat{\sigma^2}$
- (which is why in OLS we can first estimate $\boldsymbol{\beta}$ without worrying about σ^2 and then use those estimates to estimate σ^2
- the independence of the mean and variance is a peculiar (but very handy) feature of normals
- Note also we get standard errors on our estimate of σ^2 (though these are seldom reported).
- So maximum likelihood produces all the OLS produces and more
- where they both produce, they more or less produce the same thing
- Note how easy it is to extend the ML derivation to NLLS, that is, where $y_i = f(\mathbf{x}_i) + \epsilon_i$