Economics 215, 2015 Allin Cottrell

Maximum Likelihood Estimation: some basics

We start with the linear regression model expressed in matrix form,

$$y = X\beta + u \tag{1}$$

and we will assume that the error term is not only i.i.d. but also follows the normal distribution: for each element, u_i, of the vector u we have

$$u_i \sim N(0, \sigma^2)$$

For a random variable x with mean μ and standard deviation σ, the normal density (pdf) is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We now specialize this to observation i in relation to the model given by (1). The mean or expected value of y_i equals $X_i\beta$ (since $E(u_i) = 0$), and the standard deviation of y_i is just the standard deviation of the error term, σ. So the pdf is

$$f(y_i, \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - X_i\beta)^2}{2\sigma^2}\right) \tag{2}$$

This is known as the *likelihood* of observation i (conditional on X_i, and given the parameters β and σ). We almost always work with the log of the likelihood (written as ℓ). Taking the log of (2) we get

$$\ell(y_i, \beta, \sigma) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - X_i\beta\right)^2 \tag{3}$$

To find the *joint likelihood* of an entire data sample we take the *product* of the likelihoods of all the individual observations (like probabilities, likelihoods obey the multiplication rule). In log terms, this means *adding up* the contributions given by (3); for a sample of n observations this gives

$$\ell(y, \beta, \sigma) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 \tag{4}$$

Now, from the Maximum Likelihood point of view, the task of estimation is to find an estimate $\hat{\beta}$ that maximizes the joint likelihood of the sample data (which is equivalent to maximizing the log likelihood). Let's use the notation $\mathrm{SSR}(\hat{\beta})$ to indicate $\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$. So we want to choose $\hat{\beta}$ to maximize

$$\ell(y, \hat{\beta}, \sigma) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{SSR}(\hat{\beta}) \tag{5}$$

But what about the (unknown) σ? A common first step is to *concentrate* the log likelihood with respect to this term. This means finding the partial of $\ell()$ with respect to σ, setting it to zero, and then solving for σ. The relevant equation is

$$\frac{\partial\ell(y, \hat{\beta}, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\,\mathrm{SSR}(\hat{\beta}) = 0$$

which gives

$$\sigma^2 = \frac{\mathrm{SSR}(\hat{\beta})}{n}$$

Substituting the above expression into (5) we get

$$\ell^c(y, \hat{\beta}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\mathrm{SSR}(\hat{\beta}) + \frac{n}{2}\log n - \frac{n}{2}$$

Now note that $\hat{\beta}$ enters the log likelihood only via the SSR function. This means that maximizing the likelihood amounts to the same thing as minimizing the SSR, which of course is what OLS does. This equivalence is specific to the linear model.