Intermediate Social Statistics Hilary 2009 Lecture 2: Maximum Likelihood Estimation

Raymond Duch

Nuffield College

January 28, 2009

Recall Principles of MLE

- N independent and identically distributed (iid) random variables denoted Y
- Y=$[Y_1, ..., Y_N]$
- a column vector of observed data y=$[y_1, ...y_n]$ drawn from Y
- convention for a PDF is $f(y|\theta)$ denotes probability density function (PDF) that specifies the probability of observing data vector y given the parameter θ.

Recall Principles of MLE

- the principle of MLE is to find the value of θ that maximizes the likelihood of observing those data, y
- the general form of the likelihood function is :

$$lnL(\theta|y) = \sum_{i=1}^{N} lnf(y_i|\theta) \tag{1}$$

An Illustration with the Normal Distribution: Some Data

- randomly sample 10 individuals from the labour force who are working and record their wages
- taking the natural log of these wages gives a normal distribution
- y=[1.91, 1.54, 1.71, 1.55, 3.02, 1.76, 2.5, 1.84, 1.61, 1.25]$'$

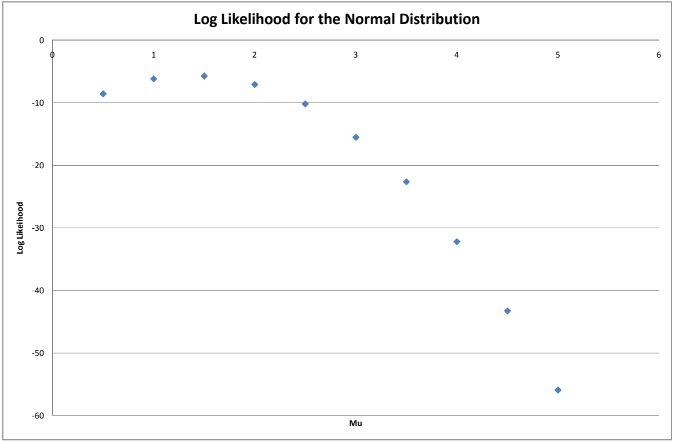
Plotting the Likelihood Function: Log Wages

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(y-\mu)^2/\sigma^2]}$$

So:

$$lnL(y|\mu, \sigma^2) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}[\frac{(y_i - \mu)^2}{\sigma^2}]$$

Plotting the Likelihood Function: Log Wages



ML and the normal linear regression model

For our familiar regression model:

$$y_i = \mathbf{x_i'}\beta + \varepsilon_i$$

The likelihood function for a sample of n independent, identically and normally distributed disturbances is:

$$L = (2\pi\sigma^2)^{-n/2} e^{-\varepsilon'\varepsilon/(2\sigma^2)}$$

Since $\varepsilon_i = y_i - x_i'\beta$, we can substitute to get:

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(y-X\beta)'(y-X\beta)}$$

ML and the normal linear regression model

Take logs (in part, because it'll be easier to see how to maximize), we get the log-likelihood function for the classical normal regression model:

$$lnL(\beta,\sigma^2|y) = -\frac{n}{2}ln2\pi - \frac{n}{2}ln\sigma^2 - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}$$

ML and the normal linear regression model

Take partials with respect to β and σ^2 to get:

$$\frac{\partial lnL}{\partial \beta} = \frac{X'(y - X\beta)}{\sigma^2} = 0$$

$$\frac{\partial lnL}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4} = 0$$

ML and the normal linear regression model

Solving these simultaneously for β and σ^2, we get:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y = b$$

and:

$$\hat{\sigma}^2_{ML} = \frac{e'e}{n}$$

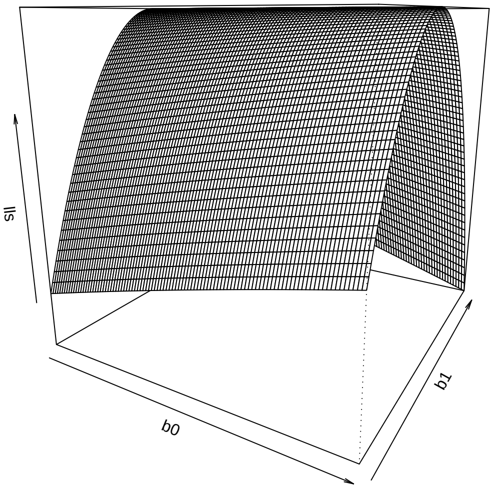
Bi-variate example

Consider a vector of independent variables (years of completed school) for a bivariate regression

x= [15, 12, 12, 11, 18, 16, 16, 14, 12]

and $\mu = 1.87$ and $\sigma = 0.67$

Bivariate example



ML and OLS Compared

Regression coefficient estimator will be the same:

$$\sum_{i=1}^{N} \varepsilon^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{1i})^2 \tag{2}$$

$$\hat{\beta}_1 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})} \tag{3}$$

Properties of ML estimators

- The properties of ML estimators, because they are derived from functions and the partials thereof, depend on the behavior of those functions.
- **unbiasedness**: $E(\hat{\theta} = \theta)$
- **consistency**: $limP(|\hat{\theta} - \theta| > \delta) = 0$
- **efficiency**: an efficient estimator is one that has achieved the lowest possible variance among all estimators, and thus had the most precision among all estimators

Likelihood Ratio Test

- If θ is a vector of parameters, then $\hat{\theta}_U$ is the ML estimator of θ without restrictions, and $\hat{\theta}_R$ is the estimator with the constraints (restrictions).
- We have an "unrestricted" log of likelihood, $\hat{L}_U$. That will be the maximum (by construction); the "restricted" log of likelihood, $\hat{L}_R$, by definition, will be smaller.
- If a restriction is valid, then the likelihood of the restriction won't cause a large reduction in the log of likelihood.

Likelihood Ratio Test

The likelihood ratio is:

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}$$

- Both of these likelihoods are positive
- since $\hat{L}_U > \hat{L}_R$, these two conditions mean that λ must be between zero and one.
- The large-sample distribution of $-2ln\lambda$ is chi-squared with df = number of restrictions imposed.