Maximum Likelihood Estimation

Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

Outline

General Approaches to Parameter Estimation

There are estimation criteria that produce estimators with good properties

Least Squares (OLS or GLS)

Method of Moments (OLS, GLS, and IV):

$$\theta = g\left(E(Y)\right) \Rightarrow \hat{\theta} = g\left(E_N[y_i]\right)$$

Maximum Likelihood (ML)

It chooses the vector $\hat{\theta}$ which makes the estimation of the probability of the sample most likely

Basic Setup

- Let $\{y_1, y_2, \ldots, y_N\}$ be an iid sample from the population with density $f(Y; \theta_0)$. We aim to estimate θ_0
- Because of the iid assumption, the joint distribution of $\{y_1, y_2, \ldots, y_N\}$ is simply the product of the densities:

$$f(y_1, y_2, \ldots, y_N; \theta_0) = f(y_1; \theta_0) f(y_2; \theta_0) \ldots f(y_N; \theta_0)$$

- The Likelihood Function is the function obtained for a given sample after replacing true θ_0 by any θ

$$L(\theta) = f(y_1; \theta) f(y_2; \theta) \ldots f(y_N; \theta)$$

- $L(\theta)$ is a random variable because it depends on the sample

Definition

The maximum likelihood estimator of θ_0, $\hat{\theta}^{ML}$, is the value of θ that maximizes the likelihood function $L(\theta)$

- It is more convenient to work with the logarithm of the likelihood function

$$l(\theta) = \sum_{i=1}^{N} log(f(y_i;\theta))$$

- Since the logarithmiic transform is monotonic, $\hat{\theta}^{ML}$ also maximizes $l(\theta)$

Example: Bernoulli (1/3)

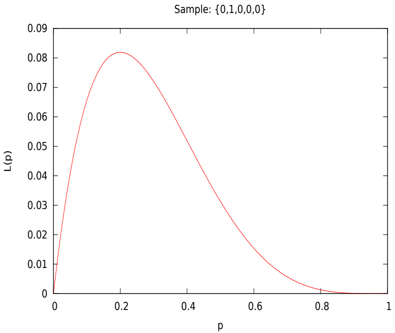
- Assume that Y is Bernoulli: $\begin{cases} 1 & \text{with probability } p_0 \\ 0 & \text{with probability } 1-p_0 \end{cases}$
- Likelihood for observation i : $\begin{cases} p_0 & \text{if } y_i = 1 \\ 1-p_0 & \text{if } y_i = 0 \end{cases}$
- Let n_1 be the number of observations with 1. Then, under iid sampling

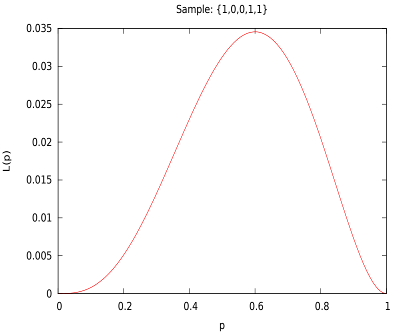
$$L(p) = p^{n_1}(1-p)^{n-n_1}$$

We have a likelihood for each sample

- With $\{0,1,0,0,0\} \Rightarrow L(p) = p(1-p)^4$
- With $\{1,0,0,1,1\} \Rightarrow L(p) = p^3(1-p)^2$

Example: Bernoulli (2/3)





- With $\{0,1,0,0,0\} \Rightarrow \hat{p} = 0.2$
- With $\{1,0,0,1,1\} \Rightarrow \hat{p}^{ML} = 0.6$

Example: Bernoulli (3/3)

- The maximum likelihood estimator is the value that maximizes

$$L(p) = p^{n_1}(1-p)^{n-n_1}$$

- The same $\hat{p}^{ML}$ maximizes the logarithm of the likelihood function

$$l(p) = n_1 log(p) + (n-n_1) log(1-p)$$

$$\frac{\partial l(p)}{\partial p} = 0 \Leftrightarrow \frac{n_1}{\hat{p}^{ML}} = \frac{n-n_1}{1-\hat{p}^{ML}} \Rightarrow \hat{p}^{ML} = \frac{n_1}{n}$$

- With $\{0,1,0,0,0\} \Rightarrow \hat{p}^{ML} = \frac{1}{5} = 0.2$
- With $\{1,0,0,1,1\} \Rightarrow\Rightarrow \hat{p}^{ML} = \frac{3}{5} = 0.6$

Basic Setup

- Let $\{y_1, y_2, ..., y_N\}$ be an iid sample from $y|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right)$.
- We aim to estimate $\theta_0 = \left(\beta_0, \sigma_0^2\right)$
- Because of the iid assumption, the joint distribution of $\{y_1, y_2, ..., y_N\}$ is simply the product of the densities:

$$f\left(y_1, y_2, ..., y_N | x_1, ..., x_N; \theta_0\right) =$$
$$f(y_1|x_1;\theta_0) f(y_2|x_2;\theta_0) ... f(y_N|x_N;\theta_0)$$

- Note that $y|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right) \Rightarrow y - \beta_0 x \equiv \varepsilon \sim N\left(0, \sigma_0^2\right)$. This implies that

$$f_{y|x}\left(y_i|x_i;\theta_0\right) = f_\varepsilon\left(y_i - \beta x_i; \theta_0\right)$$

Density of the Error Term

- We have that $\varepsilon \sim N\left(0, \sigma_0^2\right)$, so what is its density $f_\varepsilon\left(z; \theta_0\right)$?

1. $\varepsilon \sim N\left(0, \sigma_0^2\right) \rightarrow \frac{\varepsilon}{\sigma_0} \sim N\left(0, 1\right)$
2. $CDF_\varepsilon\left(z\right) \equiv Pr\left(\varepsilon \leq z\right) = Pr\left(\frac{\varepsilon}{\sigma_0} \leq \frac{z}{\sigma_0}\right)$
3. Hence, $CDF_\varepsilon\left(z\right) = \Phi\left(\frac{z}{\sigma_0}\right)$
4. The density of a continuous random variable is the first derivative of its CDF:
$$f_\varepsilon\left(z; \theta_0\right) = \left(\frac{1}{\sigma_0}\right) \phi\left(\frac{z}{\sigma_0}\right)$$

Density of the Sample

- Since

$$f_{\varepsilon}(z;\theta_0) = \left(\frac{1}{\sigma_0}\right)\phi\left(\frac{z}{\sigma_0}\right)$$

- and

$$f_{y|x}(y_i|x_i;\theta_0) = f_{\varepsilon}(y_i - \beta x_i;\theta_0)$$

- and

$$f(y_1, y_2, \ldots, y_N|x_1, ..., x_N;\theta_0) = f(y_1|x_1;\theta_0)f(y_2|x_2;\theta_0)...f(y_N|x_N;\theta_0)$$

- then we have that

$$f(y_1, y_2, \ldots, y_N|x_1, ..., x_N;\theta_0) = \prod_{i=1}^{N}\left\{\left(\frac{1}{\sigma_0}\right)\phi\left(\frac{y_i - \beta_0 x_i}{\sigma_0}\right)\right\}$$

The Log-likelihood function

- The likelihood replaces the actual values of the parameters for real variables:

$$L(\beta,\sigma)=\prod_{i=1}^{N}\left\{\left(\frac{1}{\sigma}\right)\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)\right\}$$

- taking the log makes the problem easier

$$log(L(\beta,\sigma))=\sum_{i=1}^{N}\left\{log\left(\frac{1}{\sigma}\right)+log\left[\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)\right]\right\}$$

- and given that $\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)=(2\pi)^{-\frac{1}{2}}\exp\left[-\left(\frac{y_i-\beta x_i}{\sigma}\right)^2\right]$ we have that

$$log(L(\beta,\sigma))=Nlog\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}-\sum_{i=1}^{N}\left(\frac{y_i-\beta x_i}{\sigma}\right)^2$$

The ML Estimator: FOC

- With respect to β:

$$\frac{2}{\hat{\sigma}^2}\sum_{i=1}^{N} x_i\left(y_i-\hat{\beta}x_i\right)=0$$

- which implies

$$\sum_{i=1}^{N} x_i\left(y_i-\hat{\beta}x_i\right)=0$$

- With respect to σ, this implies

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{\beta}x_i\right)^2$$

- MLE for $\hat{\beta}$ is exactly the same estimator as OLS;$\hat{\sigma}^2=\frac{N-1}{N}s^2$ is biased, but the bias disappears as N increases

Computing the MLE

- ML estimates are often easy to compute, as in the two previous examples
- Sometimes, however, there is no algebraic solution to the maximization problem
- It is then necessary to use some sort of numerical maximization procedure

Numerical Maximization Procedures

Newton's method

- Start with an initial value $\hat{\theta}^0$
- At any iteration, $\hat{\theta}^{j+1} = \hat{\theta}^j - H^{-1}g$
  - g is the first derivative of the likelihood (i.e. the gradient)
  - H is the second derivative (the Hessian)
- Check if there is convergence

- Which $\Delta\hat{\theta}$ increases the most the quadratic Taylor approximation of $L\left(\hat{\theta}+\Delta\hat{\theta}\right)$,

$$L\left(\hat{\theta}+\Delta\hat{\theta}\right) \simeq L\left(\hat{\theta}\right) + g\left(\hat{\theta}\right)\Delta\hat{\theta} + \frac{1}{2}H\left(\hat{\theta}\right)\Delta\hat{\theta}^2?$$

Quasi-Newton Methods

- Newton's Method will not work well when the Hessian is not negative definite.
- In such cases, one popular way to obtain the MLE is to replace the Hessian by a matrix which is always negative definite
- These approaches are referred to as quasi-Newton algorithms
- `gretl` uses one of them: the BFGS algorithm (Broyden, Fletcher, Goldfarb and Shanno)

Consistency

Assumptions

1. Finite-sample identification: $l(\theta)$ takes different values for different θ
2. Sampling: a law of large numbers is satisfied by $\frac{1}{n}\Sigma_i l_i(\hat{\theta})$
3. Asymptotic identification: max $l(\theta)$ provides a unique way to determine the parameter in the limit as the sample size tends to infinity.

- Under these conditions, the ML estimator is consistent

$$plim\left(\hat{\theta}^{ML}\right) = \theta_0$$

Identificación

- These are the crucial assumptions to exploit the fact that the expected maximum likelihood attains its maximum at the true value θ_0
- If these conditions did not hold, there would be some value θ_1 such that θ_0 and θ_1 generate an identical distribution of the observable data
- Then we wouldn't be able to distinguish between these two parameters even with an infinite amount of data
- We then say that these parameters are observationally equivalent and that the model is not identified

Asymptotic Normality

Assumptions

1. Consistency
2. $l(\theta)$ is differentiable and attains an interior maximum
3. A CLT can be applied to the gradient

- Under these conditions the ML estimator is asymptotically normal

$$n^{1/2}\left(\hat{\theta}-\theta\right) \rightarrow N\left(0,\Sigma\right)) \quad \text{as } n \rightarrow \infty$$

where $\Sigma = -\left(plim \frac{1}{n} \sum H_i\right)^{-1}$

Asymptotic Efficiency and Variance Estimation

If $l(\theta)$ is differentiable and attains an interior maximum

- the MLE must be at least as asymptotically efficient as any other consistent estimator that is asymptotically unbiased

Consistent estimators of the Varianze-Covariance Matrix

- empirical hessian: $var_H(\hat{\theta}) = -\left[\frac{1}{n}\sum H_i^{-1}(\hat{\theta})\right]^{-1}$
- BHHH, $var_{BHHH}(\hat{\theta}) = \left[\left(\frac{1}{n}\sum g_i(\hat{\theta})\right)^T\left(\frac{1}{n}\sum g_i(\hat{\theta})\right)\right]^{-1}$
- the sandwich estimator: valid even if the model is misspecified

Summary

- ML estimates are the values which maximize the likelihood function
- under general assumptions, ML is consistent, asymptotically normal, and asymptotically efficient