# Maximum Likelihood Estimation

## Econometrics II

Ricardo Mora

Department of Economics
Universidad Carlos III de Madrid
Máster Universitario en Desarrollo y Crecimiento Económico



Notes



## Outline





Notes

## General Approaches to Parameter Estimation

- are there general approaches to estimation that produce estimators with good properties, such as consistency, and efficiency?
- Least Squares: OLS, FGLS, FE
- Method of Moments: Assume $\theta = g(E(Y))$.
  - replaces population by sample moments: $\hat{\theta} = g(E_N[y_i])$.
  - OLS, FGLS, IV, FE
- Maximum Likelihood (ML): loosely speaking, it chooses $\hat{\theta}$ which maximizes the estimate of the empirical density



Notes



## Why Should We Use ML?

- Nice asymptotic results under mild conditions
- Easy to implement for “non-linear” models:
  - labor force participation decision, employment decision
  - marriage/divorce decisions, number of kids a couple want
  - big investment project decision
  - means of transport choice...
- that is, useful for discrete choices...



Notes

## Basic Setup

- Let $\{y_1, y_2, \ldots, y_N\}$ be a sample from a population each with a probability $f(Y; \theta_0)$. We know $f()$ but do not know $\theta_0$
- We assume that observations $\{y_1, y_2, \ldots, y_N\}$ are independent, so that

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

- Likelihood function: the function obtained for a given sample after replacing true $\theta_0$ by any $\theta$

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

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



Notes



## Definition

The maximum likelihood estimator of $\theta_0$, $\hat{\theta}^{ML}$, is the value of $\theta$ that maximizes the likelihood function $L(\theta)$.

- usually, it is more convenient to work with the logarithm of the likelihood function

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



Notes

## Example: Bernoulli (1/4)

- $Y$ is Bernoulli: $\begin{cases} \text{takes value } 1 & \text{with probability } p_0 \\ \text{takes value } 0 & \text{with probability } 1 - p_0 \end{cases}$
- 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

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



Notes



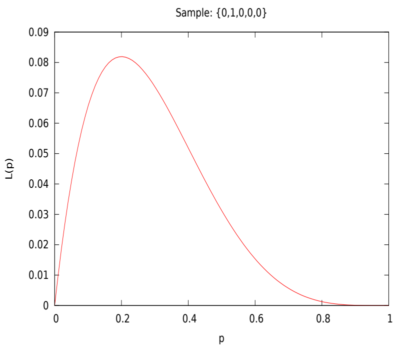
## Example: Bernoulli (2/4)

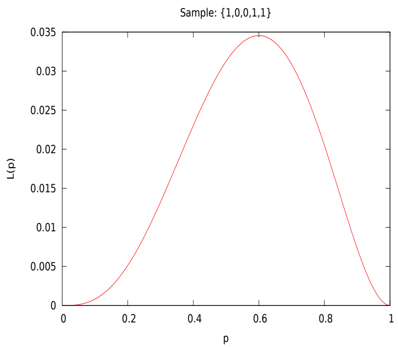
- Each sample gives us one likelihood function
- Suppose we observe this: $\{0, 1, 0, 0, 0\}$
- Then we have this likelihood: $L(p) = p(1-p)^4$
- Suppose we observed instead this sample:$\{1, 0, 0, 1, 1\}$ (ones are more frequent)
- Now we have a different likelihood: $L(p) = p^3(1-p)^2$



Notes

## Example: Bernoulli (3/4)




- with the first sample, $\hat{p} = 0.2$
- with the second sample, $\hat{p} = 0.6$



Notes

Motivation
**Definition & a Basic Example**
Linear Regression Model & ML
Asymptotic Results for ML
Summary

## Example: Bernoulli (4/4)

- the maximum likelihood estimator is the value that maximizes

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

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

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

- $\frac{\partial l(p)}{\partial p} = 0 \Leftrightarrow \frac{n_1}{\hat{p}} = \frac{n-n_1}{1-\hat{p}} \Rightarrow \hat{p} = \frac{n_1}{n}$
- with $\{0,1,0,0,0\} \Rightarrow \hat{p} = \frac{1}{5} = 0.2$
- with$\{1,0,0,1,1\} \Rightarrow \hat{p} = \frac{3}{5} = 0.6$

Ricardo Mora Max. Likelihood Estimation

Notes

## Computing the MLE

- ML estimates are sometimes easy to compute, as in the previous example
- in the linear regression model with normal errors, ML coincides with OLS
- sometimes, however, there is no algebraic solution to the maximization problem
- It is necessary to use some sort of nonlinear maximization procedure: `STATA` will take care of this



Notes



## Classical Assumptions

**Gauss-Markov Assumptions**

- A1: Linearity: $y = \beta_0 x + \varepsilon$
- A2: Random Sampling
- A3: Conditional Mean Independence: $E[y\,|x] = \beta_0 x$
- A4: Invertibility of Variance-covariance Matrix
- A5: Homoskedasticity: $Var[\varepsilon\,|x] = \sigma_0^2$

**Normality**

- A6: Normality: $y\,|x \sim N(\beta_0 x, \sigma_0^2)$



Notes

## Basic Setup

- Let $\{y_1, y_2, ..., y_N\}$ be an iid sample from the population with density $y\,|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right)$.
- 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 $y\,|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right) \Rightarrow \varepsilon \sim N\left(0, \sigma_0^2\right)$. This implies that

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



Notes



## Density of the Error Term

- We have that $\varepsilon \sim N\left(0, \sigma_0^2\right)$, so what is its density?
- We can use the following trick:

1. $\varepsilon \sim N\left(0, \sigma_0^2\right)$ implies that $\frac{\varepsilon}{\sigma_0} \sim N(0,1)$
2. $\frac{\varepsilon}{\sigma_0} \sim N(0,1)$ implies that $CDF_\varepsilon(z) = Pr\left(\frac{\varepsilon}{\sigma_0} \leq \frac{z}{\sigma_0}\right) = \Phi\left(\frac{z}{\sigma_0}\right)$
3. since the density of any continuous random variable is the first derivative of its CDF:

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



Notes

## 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

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



Notes



## The Log-likelihood (1/2)

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

$$L(\beta,\sigma) = \prod_i \left\{\left(\frac{1}{\sigma}\right)\phi\left(\frac{y_i - \beta x_i}{\sigma}\right)\right\}$$

- taking the log makes the problem easier

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



Notes

## The Log-likelihood (2/2)

- the first term inside the sum is a constant for all observations

$$log\left(L\left(\beta,\sigma\right)\right)=-Nlog\left(\sigma\right)+\sum_i\left\{log\left[\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)\right]\right\}$$

- 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

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



Notes

Motivation
Definition & a Basic Example
Linear Regression Model & ML
Asymptotic Results for ML
Summary

## The ML Estimator: FOC

- The ML estimator is the value for $(\beta,\sigma)$ such that the log-likelihood is maximized
- We obtain the maximum of the likelihood by setting the partial derivatives with respect to $(\beta,\sigma)$ to zero
- With respect to $\beta$, this implies

$$\frac{2}{\hat{\sigma}^2}\sum x_i\left(\frac{y_i-\hat{\beta}x_i}{\hat{\sigma}}\right)=0$$

- which implies

$$\sum x_i\left(y_i-\hat{\beta}x_i\right)=0$$

- With respect to $\sigma$, this implies

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

Ricardo Mora Max. Likelihood Estimation

Notes

## Some Final Comments

- MLE for $\hat{\beta}$ is exactly the same estimator as OLS
- $\hat{\sigma}^2$ is not the same as the unbiased estimator

$$s^2 = \frac{1}{N-1}\sum_i \left(y_i - \hat{\beta}x_i\right)^2$$

  - $\hat{\sigma}^2 = \frac{N-1}{N}s^2$ is biased, but the bias disappears as $N$ increases



Notes



## Consistency

**Assumptions**

- finite-sample identification: $l(\theta)$ takes different values for different $\theta$
- sampling: a law of large numbers is satisfied by $\frac{1}{n}\Sigma_i l_i(\hat{\theta})$
- 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

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



Notes

## Asymptotic Normality

**Assumptions**

- consistency
- $l(\theta)$ is differentiable and attains an interior maximum
- a Central Limit Theorem can be applied to the gradient

- Under these conditions the ML estimator is asymptotically normal
$$n^{1/2}\left(\hat{\theta}-\theta\right) \to N(0,\Sigma)) \quad \text{as } n \to \infty$$
where $\Sigma = -\left(plim\frac{1}{n}\sum H_i\right)^{-1}$



Notes



## 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 Variance-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 (`robust` option in `STATA`)



Notes

## Summary

- ML estimates are the values which maximize the likelihood function
- under the Gauss-Markov assumptions plus normality of the error term, $\hat{\beta}^{ML}$ is exactly the same estimator as $\hat{\beta}^{OLS}$
- under general assumptions, ML is consistent, asymptotically normal, and asymptotically efficient



Notes

Notes