Maximum Likelihood Estimation

Categorical and Limited Dependent Variables

Paul A. Jargowsky

Basic Principles

- Choose estimators that maximize the probability of observing the sample (Fischer 1950).
- The opposite of what you did in statistics class:
  - Given the parameters, calculate the probability of a specific outcome
  - In MLE, you have the outcomes, and you try to estimate what the parameters are
- Note: you are not responsible for the math in these slides, but you need to understand the basic principles.

Binomial Probability Distribution:

$$\Pr(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

Example: If the probability of a ball being orange is 0.3 (p=0.3), what is probability of picking 2 orange ball (x=2) in 3 attempts (n=3)?

$$\Pr(2) = \frac{3!}{2!(3-2)!}(0.3)^2 (1-0.3)^{(3-2)}$$

$$= \frac{6}{(2)(1)}(0.09)(0.7)$$

$$= 0.189$$

Probability of Picking X Orange Balls in 3 Trials as a Function of the True P

Number of Orange Balls:

	0	1	2	3
True P				
0	1	0	0	0
0.1	0.729	0.243	0.027	0.001
0.2	0.512	0.384	0.096	0.008
0.3	0.343	0.441	0.189	0.027
0.4	0.216	0.432	0.288	0.064
0.5	0.125	0.375	0.375	0.125
0.6	0.064	0.288	0.432	0.216
0.7	0.027	0.189	0.441	0.343
0.8	0.008	0.096	0.384	0.512
0.9	0.001	0.027	0.243	0.729
1	0	0	0	1

The Principle of Maximum Likelihood: your best guess about the true population parameter, based on your sample, is the value that maximizes the probability of observing your sample. A sample of 2 orange ball in 3 tries is most likely if the underlying P is 0.7 and not at all likely if the true parameter is 0.1. Your best guess is 0.7 if you observe 2 orange balls.

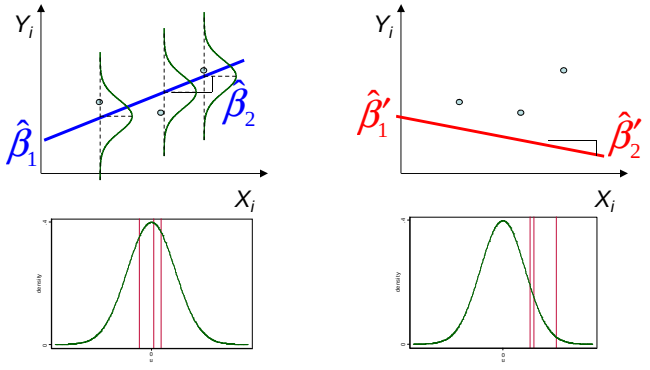
Application of MLE

- You need a function that tells you the probability of observing a particular outcome in terms of the underlying data (fixed) and the parameters (which need to be estimated).
- The probability of the sample is the product of the probabilities of the individual observations, if they are statistically independent
- If possible, solve for an analytic solution.
- If not, maximize the likelihood function numerically (in other words, search for an answer by trying different estimated values of the parameters).

Application to Linear Regression

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad u_i = Y_i - (\beta_1 + \beta_2 X_i)$$

$$u_i \sim N(0, \sigma^2) \qquad = Y_i - \beta_1 - \beta_2 X_i$$



Either outcome is possible, but they are not equally likely.

What is the likelihood of the a single observed point?

$$f(u_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u_i - 0}{\sigma}\right)^2}$$

$$= \frac{1}{\sigma}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u_i}{\sigma}\right)^2}\right)$$

$$= \frac{1}{\sigma}\phi\left(\frac{u_i}{\sigma}\right)$$

$$= \frac{1}{\sigma}\phi\left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma}\right)$$

The likelihood of individual observation is given by the normal density. For convenience, we express it in terms of the standard normal.

The probability of the whole sample

- When flipping a coin three times, what is the probability of getting 3 heads?
- What did you have to assume?

$$\mathcal{L} = \prod_{i=1}^{n}[P_i] = (0.5)(0.5)(0.5) = 0.5^3 = 0.125$$

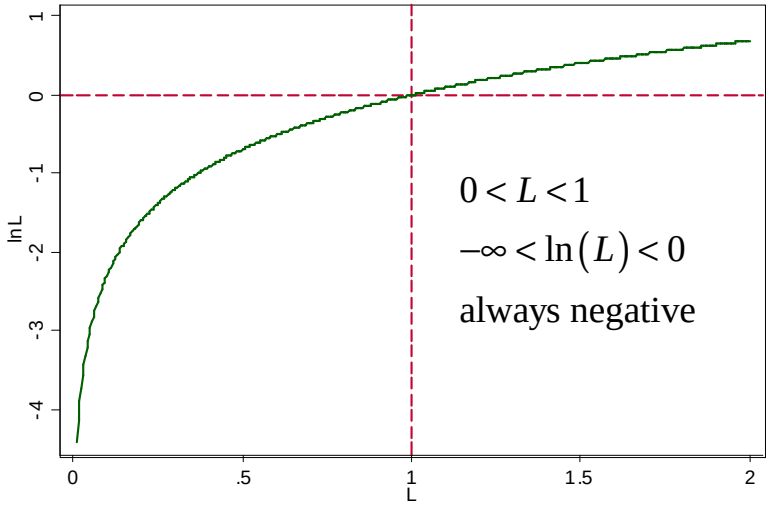
$$\mathcal{L} = \prod_{i=1}^{n}[f(u_i)] = \prod_{i=1}^{n}\left[\frac{1}{\sigma}\phi\left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma}\right)\right]$$

$$\mathrm{L} = \prod_{i=1}^{n}\left[\frac{1}{\hat{\sigma}}\phi\left(\frac{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}{\hat{\sigma}}\right)\right]$$

Log of the likelihood

- Need to maximize likelihood with respect to betas and sigma.
- With many observations, need to use log of likelihood to minimize computational errors, e.g. with N=100 and P=0.5:
  0.5^{100}=.000 000 000 000 000 000 000 000 000 000 789
- Thus, we take the *log* of the likelihood
  - Whatever maximizes *ln(L)* maximizes *L*
  - Products become sums, easier to compute

Range of *ln(L)*



Computing the ln(L) from the data

$$\mathrm{L} = \prod_{i=1}^{n}\left[\frac{1}{\hat{\sigma}}\phi\left(\frac{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}{\hat{\sigma}}\right)\right]$$

Assumes what?

$$\ln(L) = \sum_{i=1}^{n}\ln\left[\frac{1}{\hat{\sigma}}\phi\left(\frac{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}{\hat{\sigma}}\right)\right]$$

$$= \sum_{i=1}^{n}\ln\left[\frac{1}{\hat{\sigma}\sqrt{2\pi}}e^{\left(-\frac{1}{2}\right)\left(\frac{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}{\hat{\sigma}}\right)^2}\right]$$

$$= \sum_{i=1}^{n}\left[\ln\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}{\hat{\sigma}}\right)^2\right]$$

Taking derivatives

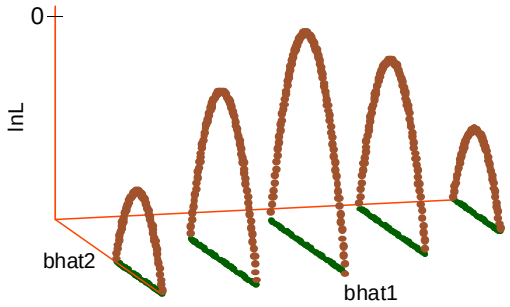
$$\frac{\partial \ln L}{\partial \hat{\beta}_1} = \left(\frac{1}{\hat{\sigma}}\right)\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = 0$$

$$\frac{\partial \ln L}{\partial \hat{\beta}_2} = \left(\frac{1}{\hat{\sigma}}\right)\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)X_i = 0$$

…,etc., for all the parameter estimates. It is clear that these are the same as the normal equations from OLS. Therefore, maximizing the likelihood of a linear equation has an analytic solution and that solution is identical to the one you get when you minimize the sum of squared residuals. Therefore:

$$\hat{\beta}_1^{OLS} = \hat{\beta}_1^{MLE} \quad \hat{\beta}_2^{OLS} = \hat{\beta}_2^{MLE}$$

What if no analytic solution?



$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

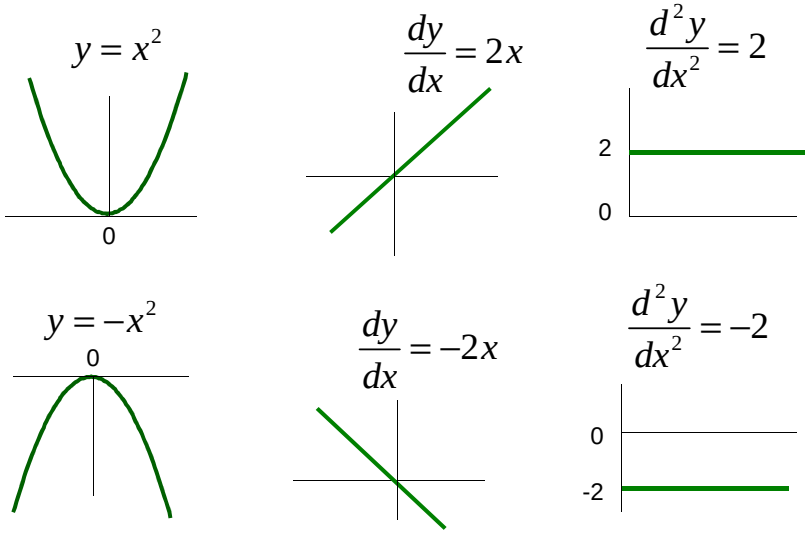
First Order Condition:

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = 0 \quad \rightarrow \quad \frac{\partial \ln \mathcal{L}}{\partial \beta_1} = 0, \quad \frac{\partial \ln \mathcal{L}}{\partial \beta_2} = 0$$

Properties of the MLE estimators

- What did we assume?
  - Constant variance σ (in the linear case), but not necessary in general
  - Observations are independent: cov(u_i,u_j)=0
  - We could express P_i in terms of parameters and the data
- **In general, MLE estimators are**
  - **Consistent**
  - **Asymptotically efficient**
  - **Asymptotically normally distributed**
- Some hairy mathematical details on following slides (optional, feel free to ignore).

Second Order Condition



The Hessian: a matrix of second derivatives of L

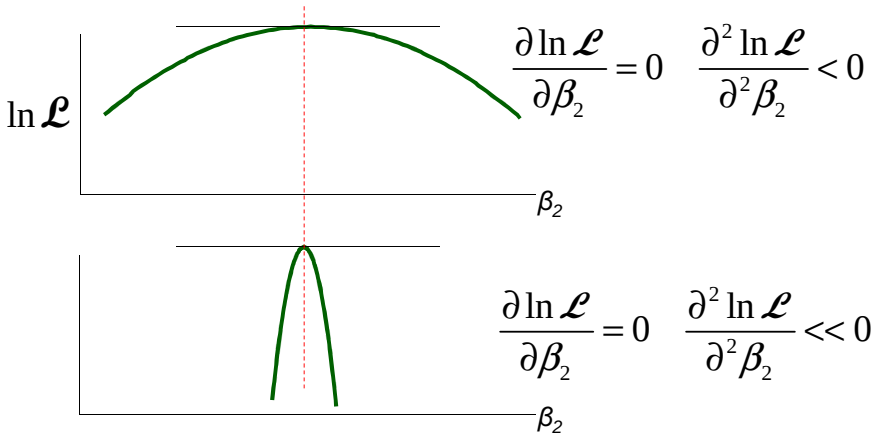
$$\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$$

At the maximum, these will all be negative.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \qquad \mathbf{H}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_1^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_2^2} \end{bmatrix}$$

H gives the rate at which the slopes of the likelihood function are changing.

Comparing two estimators



Which estimator is more precisely measured, i.e. has the least variance?

Variance of ML estimators

$$\text{var}\left(\hat{\boldsymbol{\beta}}\right) = -E\left[\mathbf{H}(\boldsymbol{\beta})\right]^{-1} \qquad \text{var}\left(\hat{\beta}_2\right) = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \beta_2^2}\right]^{-1}$$

- Variance is closely related to second derivatives of the $\mathcal{L}$ function.
- Always positive (at solution)
- The faster the $\ln\mathcal{L}$ changes near the maximum, the smaller the variance of the estimator
- Long 3.6 discusses specifics of numerical estimation techniques. We will trust in Stata.