Lecture 6: Maximum Likelihood Estimation

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last time

- Logit Model

Today

- Nonlinear Regression Models
- Maximum Likelihood Estimation
- Properties of the Maximum Likelihood Estimator
- MLE Goodness of Fit and Hypothesis Tests

Upcoming

- Reading for next time
  - Train textbook, Chapter 8
- Problem set
  - Problem Set 1 is posted, due September 24

Recap and Looking Ahead

So far

- Discrete choice framework
- Random utility model
- Logit model

But we still do not know how to estimate the logit model!

Coming up

- Maximum likelihood estimation
- Numerical optimization
- Estimating the logit model

Nonlinear Regression Models

Nonlinear Regression Models

The general formula for a nonlinear regression model

$$y_i = h(x_i, \theta) + \varepsilon_i$$

- OLS (linear model) is a special case
- Some models that appear to be nonlinear can be simplified to a linear model
  - See the Cobb-Douglas production function example from Lecture 1

Examples of Nonlinear Regression Models

Binary discrete choice

$$u_i = \beta' x_i + \varepsilon_i$$

$$y_i = \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

CES production function

$$\ln y_i = \ln \gamma - \frac{\upsilon}{\rho} \ln \left[\delta K_i^{-\rho} + (1 - \delta) L_i^{-\rho} \right] + \varepsilon_i$$

Exponential regression

$$y_i = \beta_0 + \beta_1 e^{\beta_2 x_{i1} + \beta_3 x_{i2}} + \varepsilon_i$$

Nonlinear Regression Assumptions

Assumption 1: Functional form

The conditional mean function of observation y_i given x_i is

$$E(y_i \mid x_i) = h(x_i, \theta)$$

where $h(x_i, \theta)$ is a continuously differetiable function of θ

Nonlinear Regression Assumptions

Assumption 2: Identifiability of the model parameters

The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\theta^0 \neq \theta$ such that $h(x_i, \theta^0) = h(x_i, \theta)$ for all x_i

- There cannot be a second set of parameters, θ^0, that gives the exact same model fit as the true parameters, θ
  - The logit model is not identified without normalizing the scale parameter (setting $\sigma = 1$) because we could multiply the coefficients and scale parameter by the same constant to get a β^0 and σ^0 that fit the model exactly the same
- In the linear model, this was the full rank assumption, but the simple absence of "multicollinearity" among the variables in x_i is not sufficient to produce this condition in the nonlinear regression model

Nonlinear Regression Assumptions

Assumption 3: Zero mean of the disturbance

It follows from Assumption 1 that we may write

$$y_i = h(x_i, \theta) + \varepsilon_i$$

where

$$E[\varepsilon_i \mid h(x_i, \theta)] = 0$$

- The error at observation i is uncorrelated with the conditional mean function, which is not necessarily the same as assuming that the errors and the exogenous variables are uncorrelated
- In the linear model, the conditioning argument of this assumption is the independent variables x_i, but here the entire conditional mean function $h(x_i, \theta)$ is the conditioning argument

Nonlinear Regression Assumptions

Assumption 4: Homoskedasticity and nonautocorrelation

As in the linear model, we assume conditional homoskedasticity

$$E[\varepsilon_i^2 \mid h(x_j, \theta)\ \forall j] = \sigma^2 \text{ (a finite constant)}$$

and nonautocorrelation

$$E[\varepsilon_i \varepsilon_j \mid h(x_k, \theta)\ \forall k] = 0\ \forall j \neq i$$

- In the linear model, the conditioning argument of this assumption is the full set of independent variables, X

Nonlinear Regression Assumptions

Assumption 5: Data generating process

The data generating process for x_i is assumed to be a well-behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts

- The process generating x_i is strictly exogenous to that generating ε_i
- The data on x_i are assumed to be "well behaved"

Nonlinear Regression Assumptions

Assumption 6: Underlying probability model

There is a well-defined probability distribution generating ε_i. At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables ε_i with mean zero and variance σ^2 conditioned on $h(x_i, \theta)$.

- This assumption implies our general model is semiparametric
- We will soon impose additional assumptions for our structural estimation methods

Estimating a Nonlinear Regression Model

Once we specify a nonlinear regression model, we have a few ways to estimate its parameters

- Maximum likelihood estimation (MLE)
- Nonlinear least squares (NLS)
- Generalized method of moments (GMM)

MLE and NLS will impose additional assumptions on ε_i, but GMM is free from additional assumptions

Maximum Likelihood Estimation

Maximum Likelihood Estimation Assumption

MLE Assumption

The probability density function (PDF) for a random variable, y, conditioned on a set of parameters, θ, is

$$f(y \mid \theta)$$

- This function identifies the data generating process that underlies an observed sample of data and provides a mathematical description of the data that the process will produce
- We are making an assumption about the density of y, not just its expectation and variance

Likelihood Function

The joint density of n independent and identically distributed (i.i.d.) from this process is

$$f(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

We switch the conditioning and define the likelihood function as a function of the unknown parameters, θ, conditioned on the data we observe, y

$$L(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

It will usually be easier to work with the log of the likelihood function, which is called the log-likelihood function

$$\ln L(\theta \mid y) = \sum_{i=1}^{n} \ln f(y_i \mid \theta)$$

Maximum Likelihood Estimator

The maximum likelihood estimator is the estimator that maximizes the likelihood function

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, L(\theta \mid y)$$

Because natural log is a montonic function, the estimator also maximizes the log-likelihood function

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \ln L(\theta \mid y)$$

A necessary condition for maximizing $\ln L(\theta \mid y)$ is

$$\frac{\partial \ln L(\theta \mid y)}{\partial \theta} = 0$$

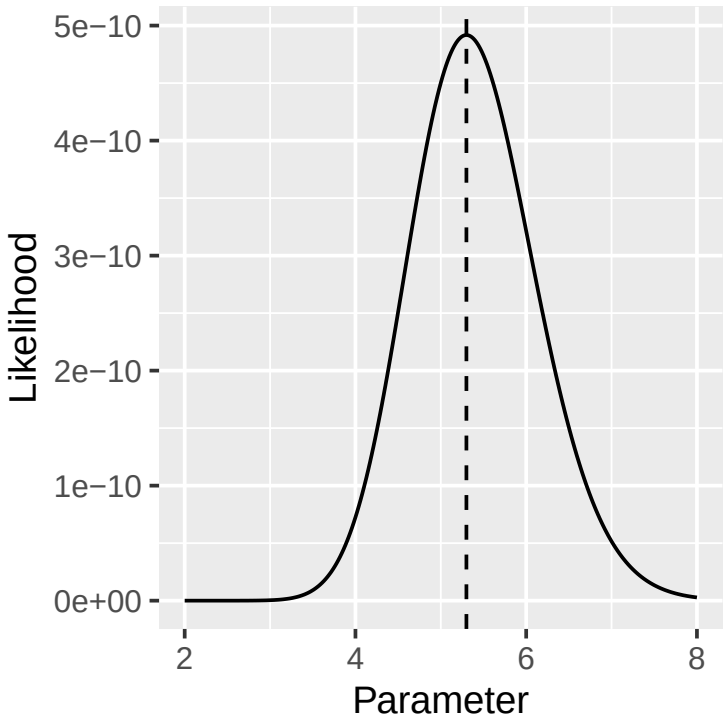
This estimator gives the parameter value(s) that maximizes the likelihood of having observed the data that we did observe

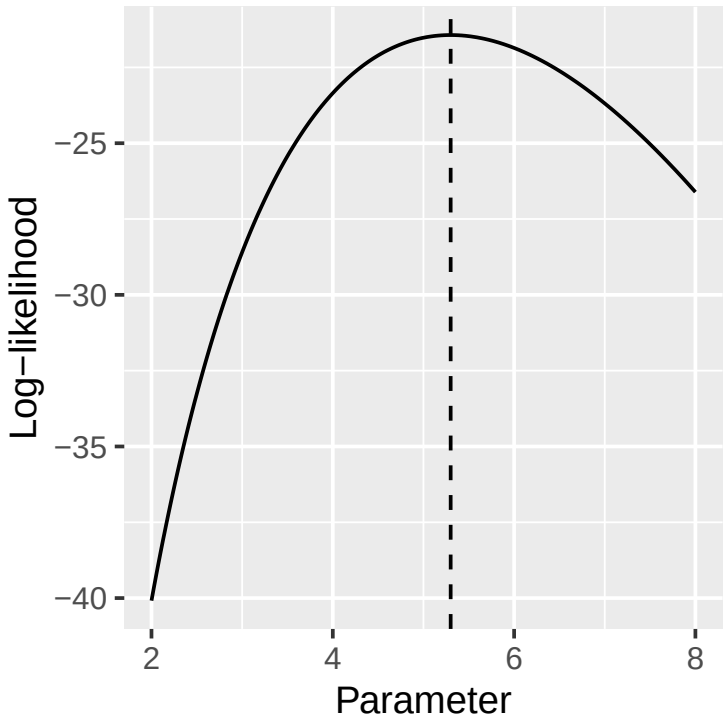
Graphical Example of Maximum Likelihood

We have some data from a Poisson distribution. What is the λ parameter of that distribution?

$$L(\lambda \mid y) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

$$\ln L(\lambda \mid y) = -n\lambda + \ln \lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!)$$





Conditional Likelihood

So far, we have assume y is function of parameters, but we also want to allow it to be a function of (or conditional on) exogenous variables, x

$$f(y \mid x, \alpha)$$

Then we can write the conditional log-likelihood function as

$$\ln L(\alpha \mid x, y) = \sum_{i=1}^{n} \ln f(y_i \mid x_i, \alpha) + \sum_{i=1}^{n} \ln g(x_i \mid \alpha)$$

If these densities depend on mutually exclusive sets of parameters; that is, if we can partition α into $[\theta, \delta]$ and write

$$\ln L(\theta, \delta \mid x, y) = \sum_{i=1}^{n} \ln f(y_i \mid x_i, \theta) + \sum_{i=1}^{n} \ln g(x_i \mid \delta)$$

then the conditional likelihood can almost always be treated as likelihood and the the "conditional" is dropped for convenience

Properties of the Maximum Likelihood Estimator

Asymptotic Properties of MLE

Under certain regularity conditions (we will get to these at the end), the maximum likelihood estimator (MLE) has these properties

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance

Consistency of MLE

The MLE, $\hat{\theta}$, converges in probability to the true parameter value(s), θ_0

$$\hat{\theta} \xrightarrow{p} \theta_0$$

- As our sample size grows (to infinity), the MLE becomes vanishingly close to the true parameter value(s)

Asymptotic Normality of MLE

The asymptotic distribution of the MLE, $\hat{\theta}$, is normal with mean at the true parameter value(s), θ_0, and known variance

$$\hat{\theta} \overset{a}{\sim} N(\theta_0, I(\theta_0)^{-1})$$

where

$$I(\theta_0) = -E_0 \left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta_0'} \right]$$

- The asymptotic variance (or variance-covariance matrix) of the MLE

$$Var(\hat{\theta}) = \left\{ -E_0 \left[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] \right\}^{-1}$$

- We are more certain of the MLE when the likelihood function has more curvature

Asymptotic Efficiency of MLE

The MLE, $\hat{\theta}$, is asymptotically efficient and achieves the Cramér-Rao lower bound

$$Var(\hat{\theta}) = I(\theta_0)^{-1}$$

- No consistent estimator has lower asymptotic mean sqaured error than the MLE

Invariance of MLE

The MLE of $\gamma_0 = c(\theta_0)$ is $c(\hat{\theta})$ if $c(\theta_0)$ is continuous and continuously differentiable, where $\hat{\theta}$ is the MLE of true parameter(s) θ_0

- The MLE of a function of some parameter(s) is the function applied to the MLE of the parameter(s)

Regularity Conditions for MLE

The preceding properties are true only if the following regularity conditions are met

1. The first three derivatives of $\ln f(y_i \mid \theta)$ with respect to θ are continuous and finite for almost all y_i and for all θ. This condition ensures the existence of a certain Taylor series approximation to and the finite variance of the derivatives of $\ln L(\theta)$.
2. The conditions necessary to obtain the expectations of the first and second derivatives of $\ln f(y_i \mid \theta)$ are met.
3. For all values of θ, $\left|\frac{\partial^3 \ln f(y_i|\theta)}{\partial\theta_j\partial\theta_k\partial\theta_l}\right|$ is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.

MLE Goodness of Fit and Hypothesis Tests

Value of Likelihood Function

The likelihood function evaluated at the MLE, $L(\hat{\theta})$, gives us some idea of how well our model fits the data, especially when compared to the likelihood function evaluated at some other parameter values

We can use these comparisons to

- Say something about a model's goodness of fit
- Conduct hypothesis tests

Likelihood Ratio Index

One measure of how well a MLE fits the data is the likelihood ratio index

$$\rho = 1 - \frac{\ln L(\hat{\theta})}{\ln L(0)}$$

where $L(0)$ measures the fit of a model with only a constant term (all other parameters equal to 0)

- This index looks like R^2 and is sometimes called a "pseudo R^2"
- But this name is misleading because this metric is nothing like R^2 other than having the same range of $[0, 1]$

Larger values of ρ imply a better model fit, but this is no different from saying larger values of the likelihood function are better

Likelihood Ratio Test

The likelihood ratio test is used to test hypotheses about parameters

$$H_0\text{: Some restriction on the parameters } \theta$$

We estimate the model with and without the restrictions, yielding

- $\hat{\theta}_U$: Unrestricted MLE
- $\hat{\theta}_R$: Restricted MLE

The likelihood ratio is

$$\lambda = \frac{L(\hat{\theta}_R)}{L(\hat{\theta}_U)}$$

which has the range $(0, 1)$ and is distributed

$$\lambda \sim \chi^2(r)$$

where r is the number of restrictions

Announcements

Reading for next time

- Train textbook, Chapter 8

Upcoming

- Problem Set 1 is posted, due September 24