Introduction to the Maximum Likelihood Estimation Technique

September 24, 2015

So far our Dependent Variable is Continuous

That is, our outcome variable **Y** is assumed to follow a normal distribution having mean **xb** with variance/covariance $\sigma^2\mathbf{I}$. Many economic phenomena do not necessarily fit this story
Examples:

- Foreign Aid Allocation: Many countries receive aid money and many do not.
- Labor Supply: In your homework, over 1/3 of your sample worked zero hours
- Unemployment claims: The duration of time on the unemployment roles is left skewed and not normal
- Bankruptcy: examining household bankruptcies reveals households are in 1 or 2 categories: bankrupt or not
- School choice: Students pick one of many schools

An important difference here, is that we can't use the model errors as we have so far in the class.

So far our Dependent Variable is Continuous

That is, our outcome variable **Y** is assumed to follow a normal distribution having mean **xb** with variance/covariance $\sigma^2\mathbf{I}$. Many economic phenomena do not necessarily fit this story
Examples:

- Foreign Aid Allocation: Many countries receive aid money and many do not.
- Labor Supply: In your homework, over 1/3 of your sample worked zero hours
- Unemployment claims: The duration of time on the unemployment roles is left skewed and not normal
- Bankruptcy: examining household bankruptcies reveals households are in 1 or 2 categories: bankrupt or not
- School choice: Students pick one of many schools

An important difference here, is that we can't use the model errors as we have so far in the class.

A focus on the Job Choice Example from Mroz

Suppose you estimate the model on the full sample and calculate $\hat{\mathbf{Y}} = \mathbf{xb}$. Compare to $\mathbf{Y}$

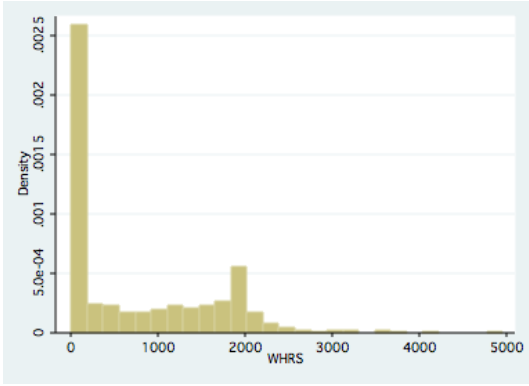


Figure: Actual Working Hours ($\mathbf{Y}$)

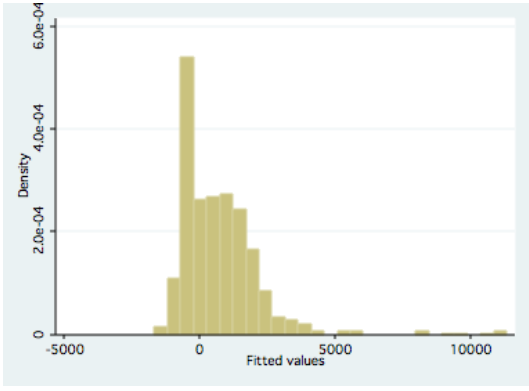


Figure: Predicted Working Hours ($\hat{\mathbf{Y}}$)

Censoring, Truncation, and Sample Selection

The preceding example comes from problems arising from censoring/truncation. In effect part of our dependent variable is continous, but a large portion of our sample is stacked on a particular value (e.g. '0' in our example)

- We don't observe the dependent variable if the individual falls below (or above) a threshold level (truncation)
  Example: We only observe profits if they are positive. Otherwise, they were negative or zero.
- We don't observe a lower (or upper) threshold value for the dependent variable if the "true" dependent variable is below a critical value (censoring)
  Example:The lowest grade level I can assign is an "F". Different students may have different capabilities (albeit not good), but all receive an "F".

For these kinds of problems, use the Tobit or Heckman models.

Dichotomous Choice

Consider a model of the unemployed. Some look for work and some may not. In this case the dependent variable is binary (1=Looking for work, 0=not looking for work).
In this case, we model the probability that an individual i is looking for work as

$$Prob(i \in looking) = \int_{-\infty}^{\infty} f(\mathbf{x}_i\beta|\epsilon_i)d\epsilon_i \tag{1}$$

Usual assumptions about the error lead to the Probit (based on the Normal Distribution) or the Logit (based on Generalized Extreme Value Type I).

Multinomial Choice- Choosing among K alternatives

Consider a firm siting decision among K communities. Each community may offer different tax packages, have different amenities, etc. The firm's choice is from among one of the K sites. Now the probability that firm i chooses community k is

$$Prob(k|i) =$$
$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}_{i1}\beta, \dots, \mathbf{x}_{ik}\beta, \dots, \mathbf{x}_{iK}\beta|\epsilon)d\epsilon$$

Usual assumptions about the error lead to the multinomial probit (based on the Normal Distribution) or the multinomial logit (based on Generalized Extreme Value Type I).

Modeling the duration of economic events

Suppose you are interested in the duration of recession i (d_i). The probability that a recession is less than 1 year long is

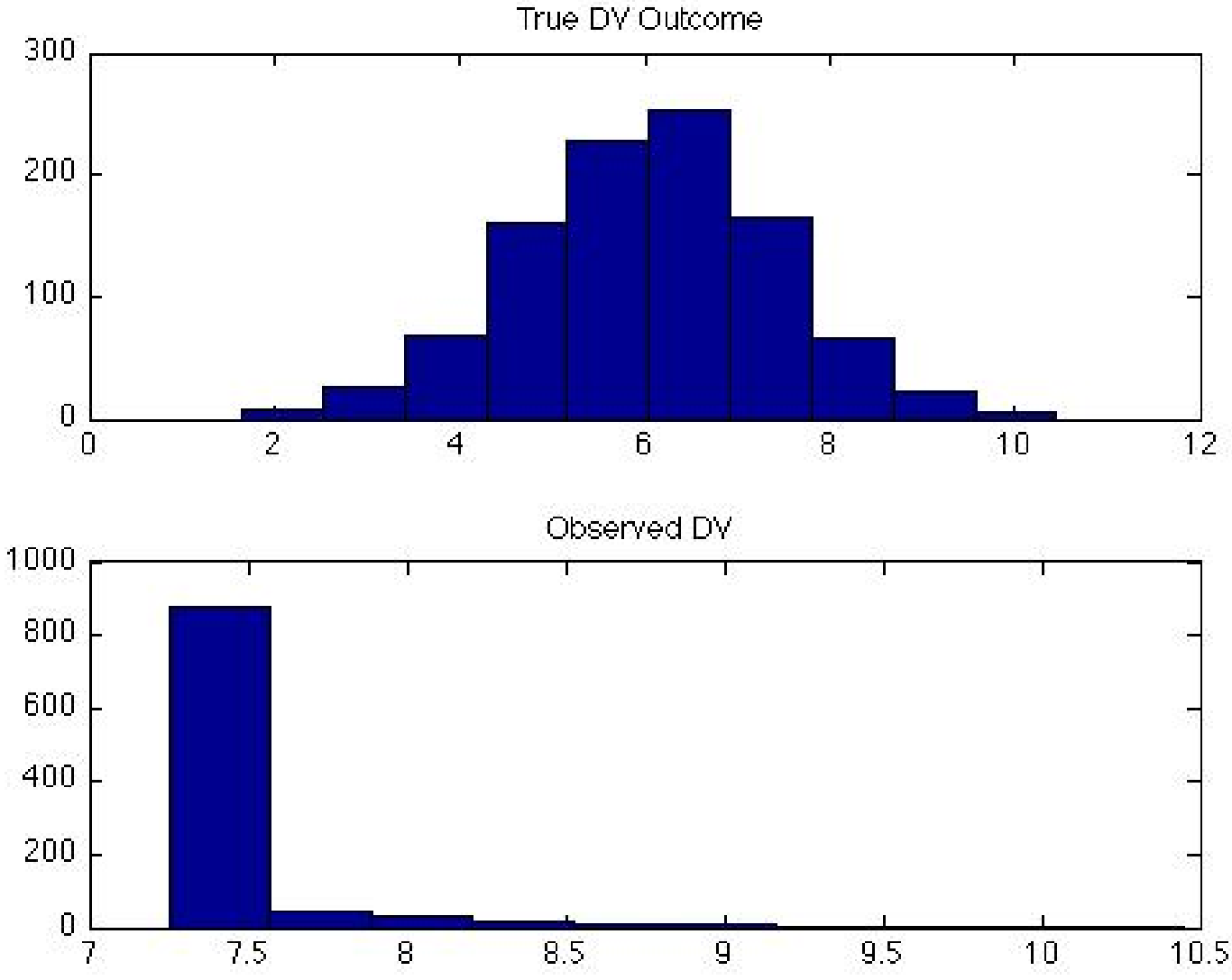
$$Prob(0 < d_i < 12) = \int_0^{12} f(\mathbf{x}_i\mathbf{b}|\epsilon, t)dt \tag{2}$$

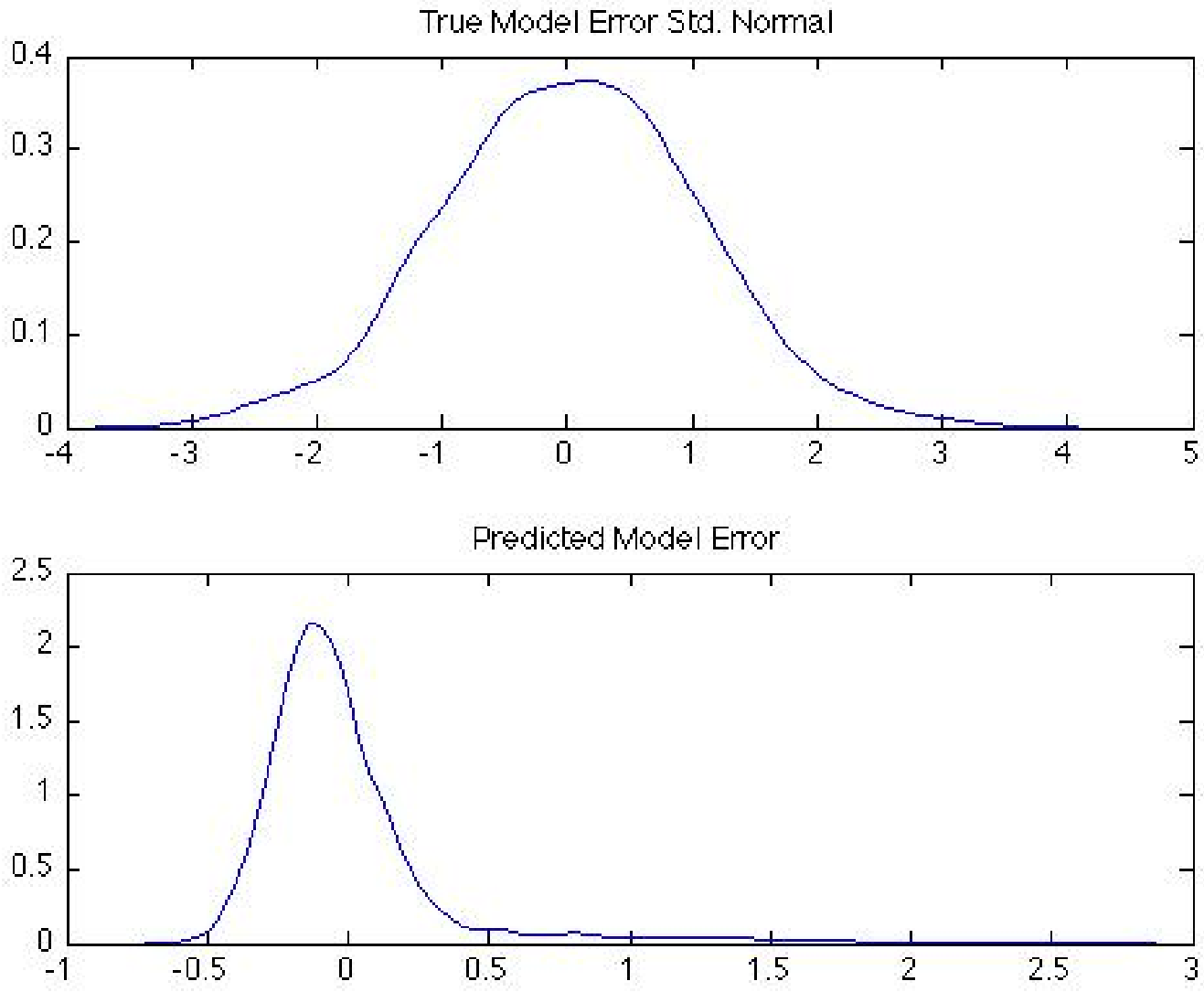
The function $f(.)$ is called the hazard function, and this methodology was adapted from survival analysis from the biological literature.

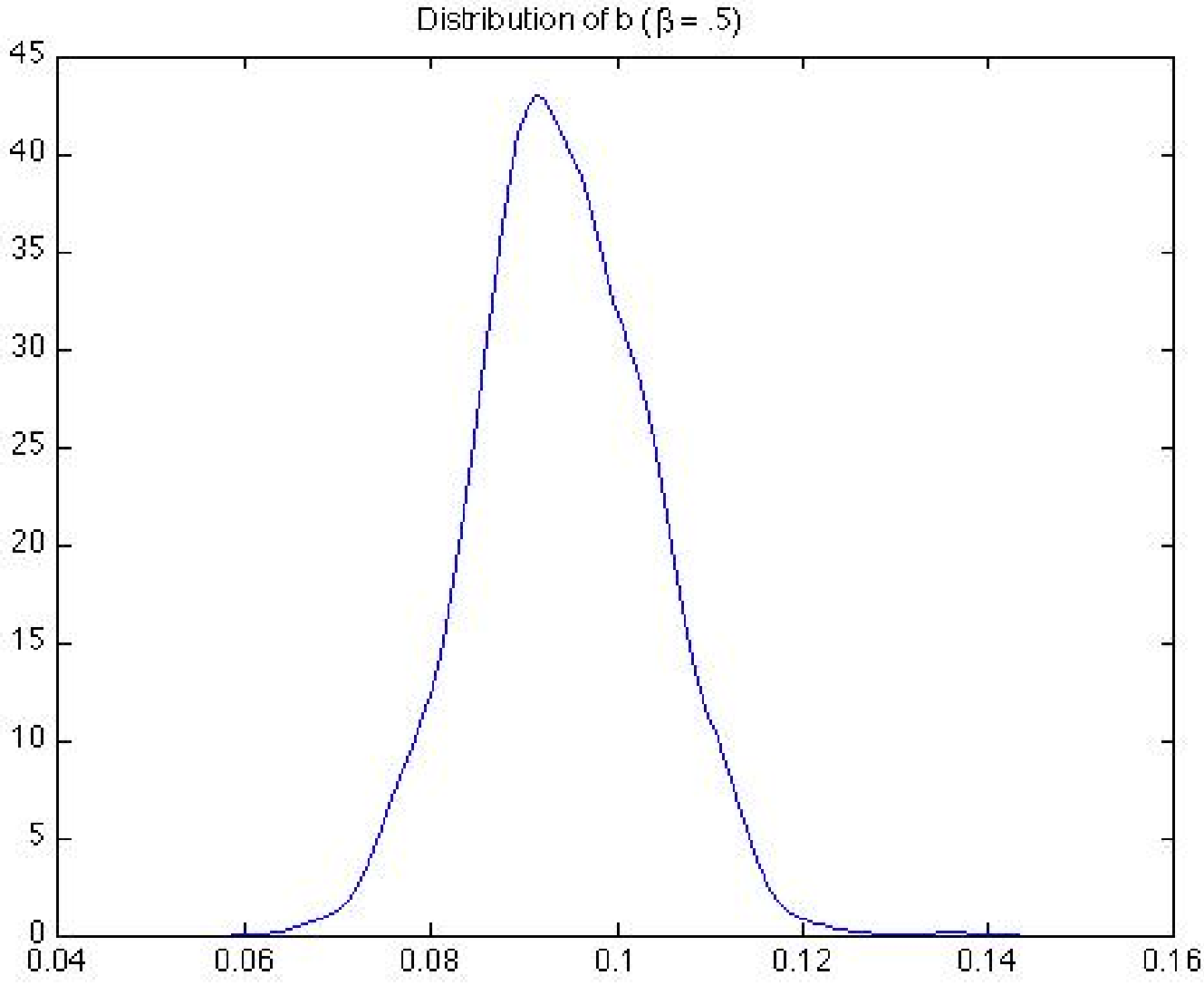
A Monte Carlo Experiment

I have performed a Monte Carlo experiment following this setup.
Data Generation Process for $N = 1000$:

1. Generate vector $\mathbf{x}$ of independent variables
2. Generate the vector ϵ where ϵ is distributed $N(0, \sigma^2 I)$.
3. Calculate "True" Dependent Variable as
   $y_{N\times 1} = 5 + .5x_{N\times 1} + \epsilon_{N\times 1}$
4. Calculate Observed Independent Variable (Y^*) as
   - $Y^* = Y$ if $Y > 7.25$
   - $Y^* = 7.25$ if $Y \leq 7.25$


True DV Outcome
300
200
100
0
0
2
4
6
8
10
12
Observed DV
1000
800
600
400
200
0
7
7.5
8
8.5
9
9.5
10
10.5


True Model Error Std. Normal
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
5
Predicted Model Error
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1
1.5
2
2.5
3


Distribution of b (β = .5)
45
40
35
30
25
20
15
10
5
0
0.04
0.06
0.08
0.1
0.12
0.14
0.16

BIG FAIL for OLS, IV Estimation, and Traditional Panel Estimators

The Maximum Likelihood Approach

The idea:

- Assume a functional form and distribution for the model errors
- For each observation, construct the probability of observing the dependent variable y_i conditional on model parameters $\mathbf{b}$
- Construct the Log-Likelihood Value
- Search over values for model parameters $\mathbf{b}$ that maximizes the sum of the Log-Likelihood Values

MLE: Formal Setup

Consider a sample $\mathbf{y} = \begin{bmatrix} y_1 & \dots & y_i & \dots & y_N \end{bmatrix}$ from the population. The probability density function (or pdf) of the random variables y_i conditioned on parameters θ is given by $f(y_i, \theta)$. The joint density of n individually and identically distributed observation is $\begin{bmatrix} y_1 & \dots & y_i & \dots & y_N \end{bmatrix}$

$$f(\mathbf{y}, \theta) = \prod_{i=1}^{N} f(y_i, \theta) = L(\theta|\mathbf{y}) \tag{3}$$

is often termed the Likelihood Function and the approach is termed Maximum Likelihood Estimation (MLE).

MLE: Our Example

In our excel spreadsheet example,

$$f(y_i, \theta) = f(y_i, \mu|\sigma^2 = 1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}} \tag{4}$$

It is common practice to work with the Log-Likelihood Function (better numerical properties for computing):

$$ln(L(\theta|\mathbf{y})) = \sum_{i=1}^{N} ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}\right) \tag{5}$$

We showed how changing the values of μ, allowed us to find the maximum log-likelihood value for the mean of our random variables $\mathbf{y}$. Hence the term maximum likelihood.

A special case: MLE and OLS

Recalling that in an OLS context, $\mathbf{y} = \mathbf{xb} + \epsilon$. Put another way, $\mathbf{y} \sim N(\mathbf{x}\beta, \sigma^2\mathbf{I})$. We can express this in a log likelihood context as

$$f(y_i|\beta, \sigma^2, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \mathbf{x_i}\beta)^2}{2\sigma^2}} \tag{6}$$

Here we estimate the K β parameters and σ^2. By finding the $K + 1$ parameter values that maximize the log likelihood function. The maximumum likelihood estimator b_{MLE} and s^2_{MLE} are exactly equivalent to their OLS counterparts b_{OLS} and s^2_{OLS}

Characterizing the "Maximum" Likelihood

In order to be assured of an optimal parameter vector b_{mle}, we need the following conditions to hold:

1. $\frac{dln(L(\theta|y,x))}{d\theta} = 0$
2. $\frac{d^2 ln(L(\theta|y,x))}{d\theta^2} < 0$

When taking this approach to the data, the optimization algorithm in stata evaluates the first and second derivates of the log-likelihood function to "climb" the hill to the topmost point representing the maximum likelihood. These conditions ignore local versus global concavity issues.

Properties of MLE

The Maximum Likelihood Estimator has the following properties

- Consistency: $plim(\hat{\theta}) = \theta$
- Asymptotic Normality: $\hat{\theta} \sim N(\theta, I(\theta)^{-1})$
- Asymptotic Efficiency: $\hat{\theta}$ is asymptotically efficient and achieves the Rao-Cramer Lower Bound for consistent estimators (minimum variance estimator).
- Invariance: The MLE of $\delta = c(\theta)$ is $c(\hat{\theta})$ if $c(\theta)$ is a continuous differentiable function.

These properties are roughly analogous to the BLUE properties of OLS. The importance of asymptotics looms large.

Hypothesis Testing in MLE: The Information Matrix

The variance/covariance matrix of the parameters θ in an MLE framework depend on

$$I(\theta) = -1 \times \frac{\partial^2 lnL(\theta)}{\partial\theta\partial\theta'} \tag{7}$$

and can be estimated by using our estimated parameter vector θ:

$$I(\hat{\theta}) = -1 \times \frac{\partial^2 lnL(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'} \tag{8}$$

The inverse of this matrix is our estimated variance covariance matrix for the parameters with standard errors for parameter i equal to $s.e.(i) = \sqrt{I(\hat{\theta}_{ii})^{-1}}$

OLS equivalence of var/covariance matrix of the parameters

Suppose we estimate an OLS model over N observations and 4 parameters. The variance covariance matrix of the parameters can be written

$$s^2(\mathbf{x}'\mathbf{x})^{-1} = s^2 \begin{bmatrix} 0.0665 & -0.0042 & -0.0035 & -0.0014 \\ -0.0042 & 0.5655 & 0.0591 & -0.0197 \\ -0.0035 & 0.0591 & 0.0205 & -0.0046 \\ -0.0014 & -0.0197 & -0.0046 & 0.0015 \end{bmatrix} \tag{9}$$

it can be shown that the first $K \times K$ rows and columns of $I(\hat{\theta})$ has the property:

$$I(\hat{\theta})^{-1}_{K\times K} = s^2(\mathbf{x}'\mathbf{x})^{-1} \tag{10}$$

Note: the last column of I contains information about the covariance (and variance) of the parameter s^2. See Green 16.9.1.

Nested Hypothesis Testing

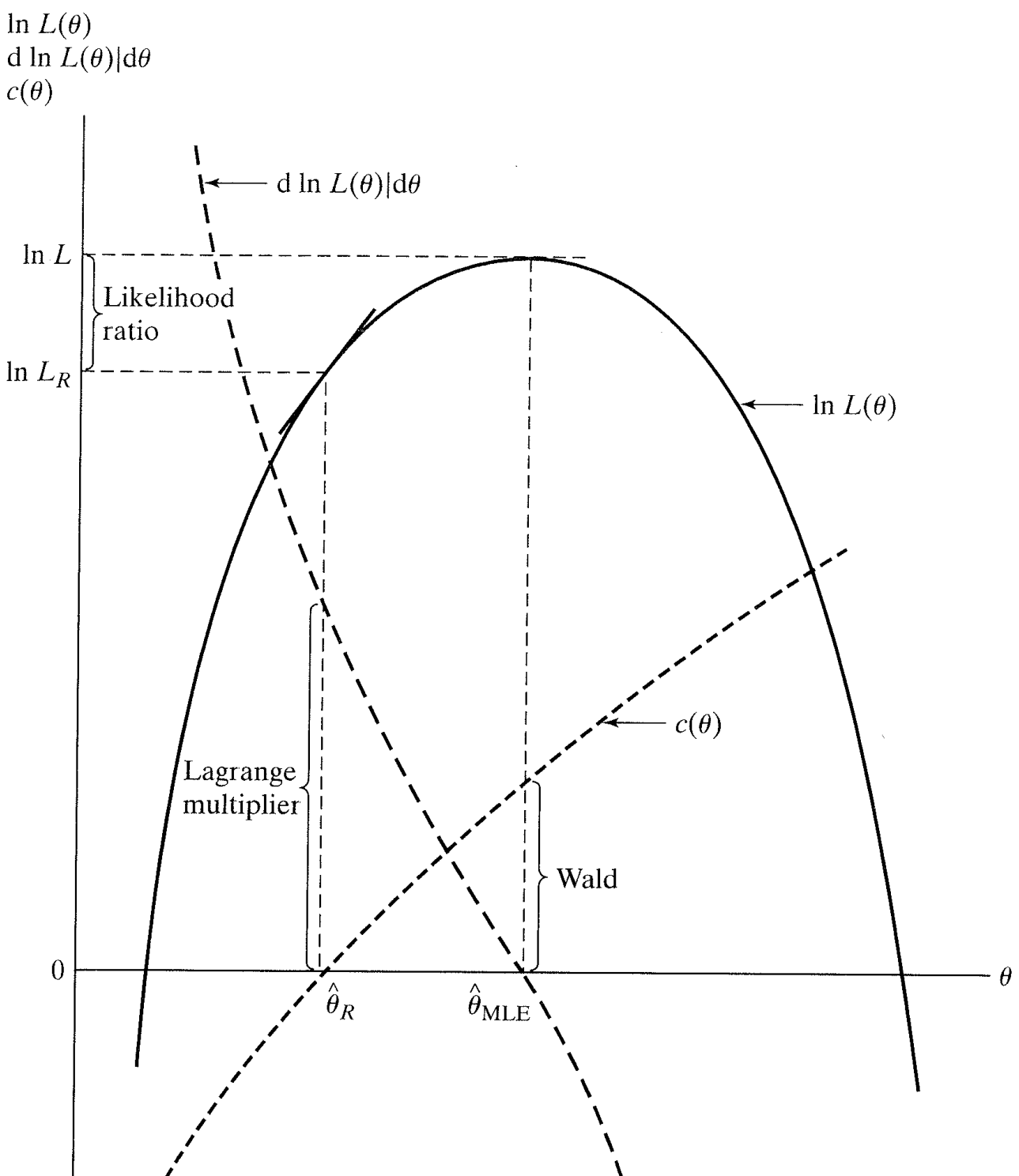
Consider a restriction of the form $c(\theta) = 0$. A common restriction we consider is

$$H_0 : c(\theta) = \theta_1 = \theta_2 = \ldots = \theta_k = 0 \tag{11}$$

In an OLS framework, we can use F tests based off of the Model, Total, and Error sum of squares. We don't have that in the MLE framework because we don't estimate model errors. Instead, we use one of three tests available in an MLE setting:

- Likelihood Ratio Test- Examine changes in the joint likelihood when restrictions imposed.
- Wald Test- Look at differences across $\hat{\theta}$ and θ_r and see if they can be attributed to sampling error.
- Lagrange Multiplier Test- examine first derivative when restrictions imposed.

These are all asymptotically equivalent and all are NESTED tests.


ln $L(\theta)$
d ln $L(\theta)$|dθ
$c(\theta)$
d ln $L(\theta)$|dθ
ln L
Likelihood ratio
ln L_R
ln $L(\theta)$
$c(\theta)$
Lagrange multiplier
Wald
0
θ
$\hat{\theta}_R$
$\hat{\theta}_{\text{MLE}}$

The Likelihood Ratio Test (LR Test)

Denote $\hat{\theta}_u$ as the unconstrained value of θ estimated via MLE and let $\hat{\theta}_r$ be the constrained maximum likelihood estimator. If $\hat{L}_u$ and $\hat{L}_r$ are the likelihood function values from these parameter vectors (not Log Likelihood Values), the likelihood ratio is then

$$\lambda = \frac{\hat{L}_r}{\hat{L}_u} \tag{12}$$

The test statistic, $LR = -2 \times ln(\lambda)$, is distributed as $\chi^2(r)$ degrees of freedom where r are the number of restrictions. In terms of log-likelihood values, the likelihood ratio test statistic is also

$$LR = -2 * (ln(\hat{L}_r) - ln(\hat{L}_u)) \tag{13}$$

The Wald Test

This test is conceptually like the Hausman test we considered in the IV sections of the course. Consider a set of linear restrictions (e.g. $R\theta = 0$).
The Wald test statistic is

$$W = \left[R\hat{\theta} - 0\right]' \left[R[Var.(\hat{\theta})]R'\right]^{-1} \left[R\hat{\theta} - 0\right] \tag{14}$$

W is distributed as $\chi^2(r)$ degrees of freedom where r are the number of restrictions.
For the case of one parameter (and the restriction that it equals zero), this simplifies to

$$W = \frac{(\hat{\theta} - 0)^2}{var(\hat{\theta})} \tag{15}$$

The Lagrange Multiplier Test (LM Test)

This one considers how close the derivative of the likelihood function is to zero once restrictions are imposed. If imposing the restrictions doesn't come at a big cost in terms of the slope of the likelihood function, then the restrictions are more likely to be consistent with the data.
The test statistic is

$$LM = \left(\frac{\partial L(R\hat{\theta})}{\partial \hat{\theta}} \right)' I(\hat{\theta})^{-1} \left(\frac{\partial L(R\hat{\theta})}{\partial \hat{\theta}} \right) \tag{16}$$

LM is distributed as $\chi^2(r)$ degrees of freedom where r are the number of restrictions. For the case of one parameter (and the restriction that it equals zero), this simplifies to

$$LM = \frac{\left(\frac{\partial L(\hat{\theta}=0)}{\partial \hat{\theta}} \right)^2}{var(\hat{\theta})} \tag{17}$$

Non-Nested Hypothesis Testing

If one wishes to test hypothesis that are not nested, different procedures are needed. A common situation is comparing models (e.g. probit versus the logit). These use Information Criteria Approaches.

Akaike Information Criterion (AIC) :

$$-2ln(L) + 2K$$

Bayes/Schwarz Information Criterion (BIC) : $-2ln(L) + Kln(N)$

where K is the number of parameters in the model and N is the number of observations. Choosing the model based on the lowest AIC/BIC is akin to choosing the model with best adjusted R^2-although it isn't necessarily based on goodness of fit, it depends on the model.

Goodness of fit

Recall that model R^2 uses the predicted model error. Here, while we have errors, we don't model them directly. Instead, there has been some work related to goodness of fit in maximum likelihood settings. McFadden's Pseudo R^2 is calculated as

$$Psuedo\ R^2 = 1 - \frac{ln(L(\hat{\theta}))}{ln(L(\hat{\theta}_{constant}))} \tag{18}$$

Some authors (Woolridge) argue that these are poor goodness of fit measures and one should tailor goodness of fit criteria for the situation one is facing.