

**Intermediate Social Statistics: Hilary 2009**  
**Raymond Duch**  
**Nuffield College**

**Lecture 1: Maximum Likelihood Estimation**

## The philosophy of likelihood

Remember the “classical” approach to statistics—taking a (presumably random) sample and deriving statistics that estimate population parameters. The whole emphasis is that there is some unknown parameter that we can use know data to estimate, with some degree of uncertainty.

In contrast, the underlying philosophy of the likelihood approach are the twin notions of the “data generating mechanism” and of “sample information.” That is, the observed data (the sample) contains information about the likely values of the parameters. What values of the parameters would most likely generate the observed data?

Elaborate a bit on how this relates to Bayesian estimation, which is a further and more complete rejection of the classical framework. The notion of presenting a model that is our “best estimate” of underlying parameters is a bit misleading, say the Bayesians. That is, most sets of empirical results that are presented are not the results of deducing the values of fixed parameters, but instead a process of forming prior beliefs, using observations to update those beliefs, and repeating the process. The mathematical construction called Bayes Theorem states that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(This is simply true mathematically.) Extended to situations with data, this becomes:

$$P(\text{parameters}|\text{data}) = \frac{P(\text{data}|\text{parameters})P(\text{parameters})}{P(\text{data})}$$

Philosophically, in this sense we view the data as a fixed set of information that help us to update our prior beliefs about the values of the parameters that are most likely to generate those data.

# The probability density function

$f(y|\theta)$  denote the probability density function (PDF) that specifies the probability of observing data vector  $y$  given the parameter  $\theta$ .  $y$  represents a vector while each element of the vector is represented by  $y_i$ . If individual observations,  $y_i$ s, are statistically independent of one another, i.e., are independent and identically distributed (iid). The pdf of  $y$ , conditioned on a set of parameters  $\theta$ , is  $f(y|\theta)$ . The joint density of  $n$  independent and identically distributed (iid) observations ( $y = (y_1, y_2, y_3, \dots, y_n)$ ) from such a process is simply the product of the individual densities, or:

$$f(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^n f(y_i|\theta) = L(\theta|y)$$

(Why the product? Think of the p(heads) and, on another observation, the p(heads). The joint density of those individual densities is the product .5 times .5 = .25, so long as those observations are independent and identically distributed, which, in the coin flip case, they are.)

To illustrate the idea of a PDF, consider the simplest case with one observation and one parameter, that is,  $m = k = 1$ . Suppose that the data  $y$  represents the number of successes in a sequence of 10 Bernoulli trials (e.g. tossing a coin 10 times) and that the probability of a success on any one trial, represented by the parameter  $\theta$ ; is 0.2. The PDF in this case is given by

$$f(y|n = 10, \theta = 0.2) = \frac{10!}{y!(10 - y)!} (0.2)^y (0.8)^{10-y} (y = 0, 1, \dots, 10) \quad (1)$$

This is known as the binomial distribution with parameters  $n=10$  and  $\theta = .2$  Note that the number of trials is considered a parameter.

We can generate a new PDF by changing the parameters – lets say by setting  $\theta = .7$

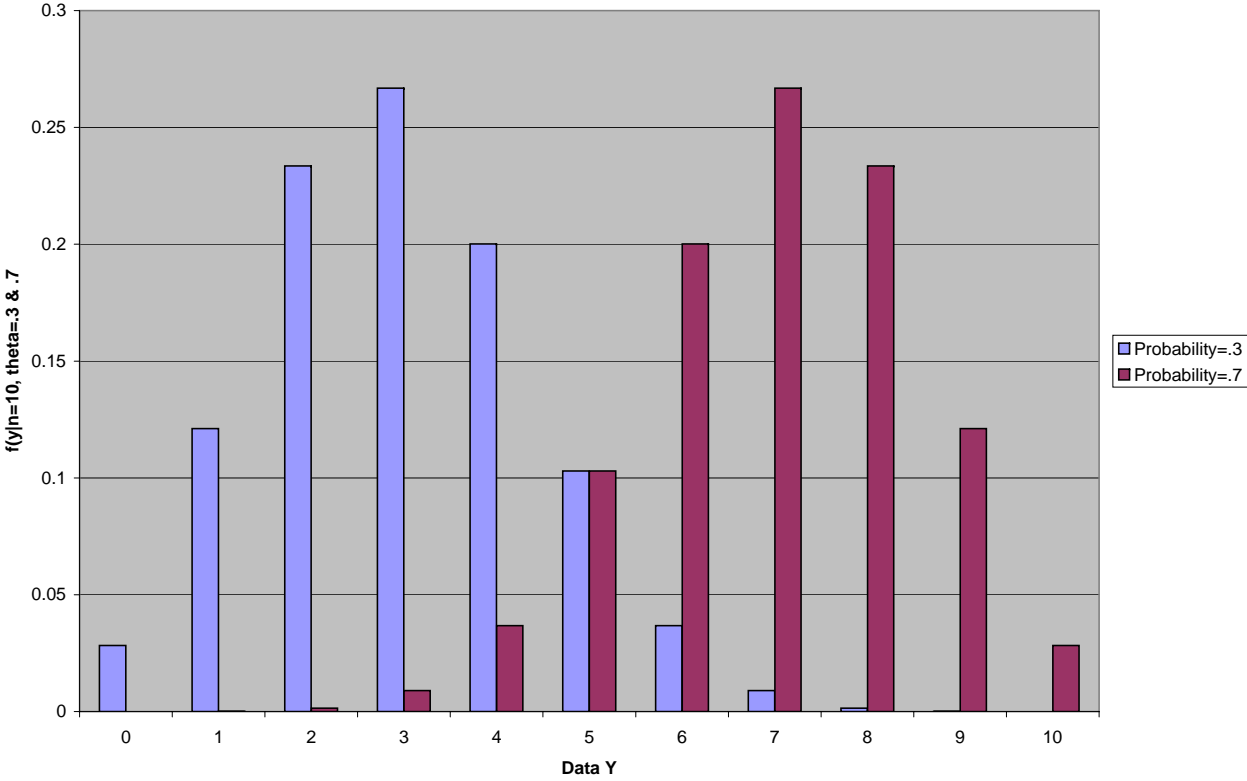
$$f(y|n = 10, \theta = 0.7) = \frac{10!}{y!(10 - y)!} (0.7)^y (0.3)^{10-y} (y = 0, 1, \dots, 10) \quad (2)$$

The following is the general expression of the PDF of the binomial distribution for arbitrary values of  $\theta$  and  $n$ :

$$f(y|n, \theta) = \frac{n!}{y!(n - y)!} (\theta)^y (1 - \theta)^{n-y} (y = 0, 1, \dots, n) \quad (3)$$

which as a function of  $y$  specifies the probability of data  $y$  for a given value of  $n$  and  $\theta$ . The collection of all such PDFs generated by varying the parameter across its range (0 to 1 in this case for  $\theta$ ;  $n \geq 1$ ) defines a model.

Figure 1: Probability Density Function



## The likelihood function

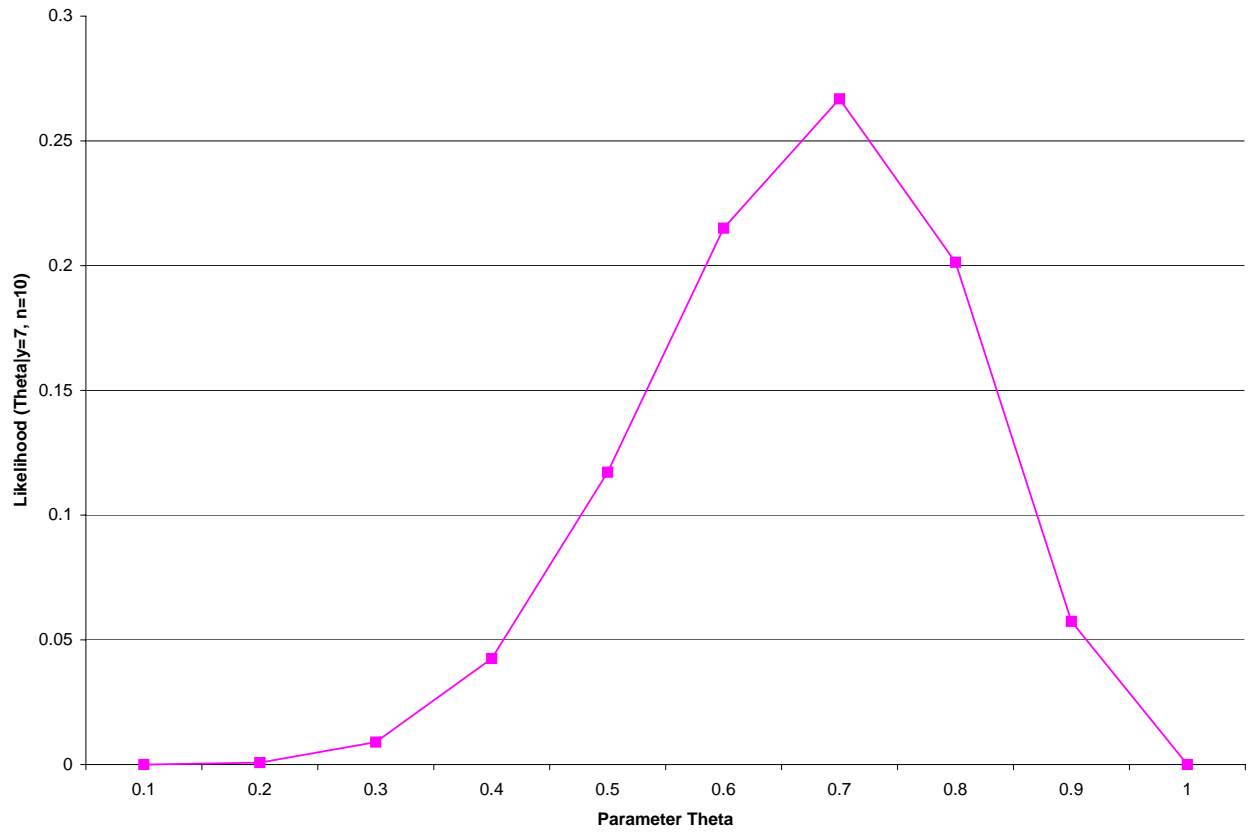
So this joint density is called the likelihood function ( $L$ ), which is defined as a function of the unknown parameter vector  $\theta$ . (Note, how we write the pdf as  $(y|\theta)$ , but the likelihood as  $(\theta|y)$ , which just emphasizes our focus on the parameters and the information that the sample data contains about those parameters. Given the observed data and a model of interest, find the one PDF, among all the probability densities that the model prescribes, that is most likely to have produced the data.

$$L(\theta|y) = f(y|\theta) \tag{4}$$

Thus  $L(\theta|y)$  represents the likelihood of the parameter  $\theta$  given the observed data  $y$ ; and as such is a function of  $\theta$ . For the one-parameter binomial example in Figure 2, the likelihood function for  $y = 7$  and  $n = 10$  is given by

$$L(\theta|n = 10, y = 7) = f(y = 7|n = 10, \theta) = \frac{10!}{7!(3)!} \theta^7 (1 - \theta)^3 (0 \leq \theta \leq 1). \tag{5}$$

Figure 2: Likelihood Function



Note the difference between a PDF and a likelihood function. The PDF in Figure 1 is a function of the data given a particular set of parameter values, defined on the data scale which ranges in this example  $y$  varying from 0 to 10. The likelihood function is a function of the parameter given a particular set of observed data, defined on the parameter scale. Figure 2 tells us the likelihood of a particular parameter value for a fixed data set – in this example  $y=7$ .

## Maximum Likelihood Estimation

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions (Figure 1), we are interested in finding the parameter value that corresponds to the desired probability distribution.

The desired probability distribution is the one that makes the observed data most likely, which means that one must seek the value of the parameter vector that maximizes the likelihood function  $L(\theta|y)$ .

The resulting parameter vector, which is sought by searching the multi-dimensional parameter space, is called the MLE estimate, and is denoted by  $\theta_{mle} = (\theta_{1,mle}, \theta_{2,mle}, \dots, \theta_{k,mle})$ . In our example illustrated in Figure 2  $\theta_{mle} = 0.7$  and the maximized likelihood value is  $L(\theta_{mle} = 0.7 : |n = 10, y = 7) = 0.267$ . The probability distribution corresponding to this MLE is shown in Figure 1 when  $\theta = 0.7$  – this describes the population that is most likely to have generated the observed data of  $y=7$ . Maximum likelihood estimation is a method to seek the probability distribution that makes the observed data most likely.

We usually work with the log of the likelihood, because taking logs of both sides means we get to work with sums instead of products. Remember that if  $y = ab$ , then  $\log y = \log a + \log b$ . So:

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

We will allow the density to depend on the  $X$ s, as well, as in:

$$\ln L(\theta|y, \mathbf{X}) = \sum_{i=1}^N \ln f(y_i|x_i, \theta)$$

What does all of this do?

Recall from the curve in Figure 2 we are looking for a maximum. How do we maximize a function? (Remember the calculus.) We take the first derivative and set it equal to zero.

(Now that could get either a minimum or a maximum.) This is called the likelihood equation:

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

$$\theta_i = \theta_{i,mle}$$

for all

$$i = 1, \dots, k.$$

This follows from the definition of maximum or minimum of a continuous differentiable function that implies that its first derivatives vanish at such points.

MLE estimates need not exist nor be unique.

Then, to check that it's a maximum, not a minimum, i.e., a peak not a valley, do the second derivative to see that it's less than zero.

$$\frac{\partial^2 \ln L(\theta|y)}{\partial \theta_i^2} < 0$$

Lets continue our previous one-parameter binomial example from above given a fixed value of n. Take the logarithm of the likelihood function  $L(\theta|n = 10, y = 7)$  we get the following..

$$\ln L(\theta|n = 10, y = 7) = \ln \frac{10!}{7!(3)!} + 7 \ln \theta + 3 \ln(1 - \theta). \quad (6)$$

Then calculate the first derivative...

$$\begin{aligned} \frac{\partial \ln L(\theta|n = 10, y = 7)}{\partial \theta} &= \frac{7}{\theta} - \frac{3}{1 - \theta} = \frac{7 - 10\theta}{\theta(1 - \theta)}. \quad (7) \\ &= -47.62 < 0 \end{aligned}$$

Often it is not possible to find an analytic form solution for the MLE estimate, especially when the model involves many parameters and its PDF is non-linear. In such situations, the MLE estimate must be sought numerically using non-linear optimization algorithms. The basic idea is to quickly find optimal parameters that maximize the log-likelihood. This is done by searching much smaller sub-sets of the multi-dimensional parameter space rather than exhaustively searching the whole parameter space. The intelligent search proceeds by trial and error over the course of a series of iterative steps.

## Log likelihood function and equations for the normal

Remember the normal?

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(y-\mu)^2/\sigma^2]}$$

Figure 3: Density Plot of Age from NYT Survey

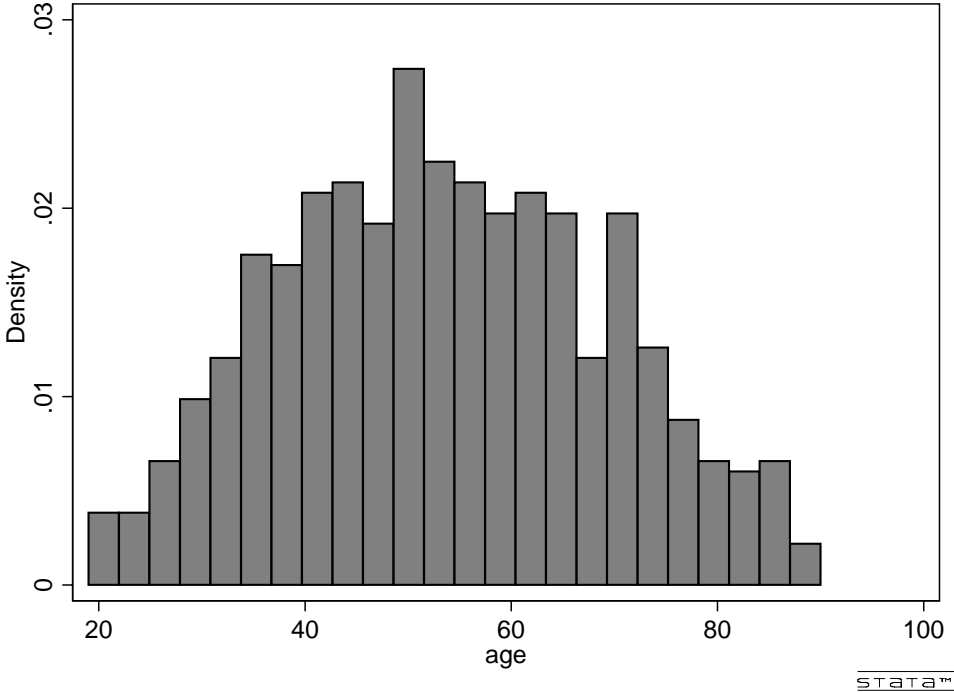


Figure 4: PDF for Age from NYT Survey

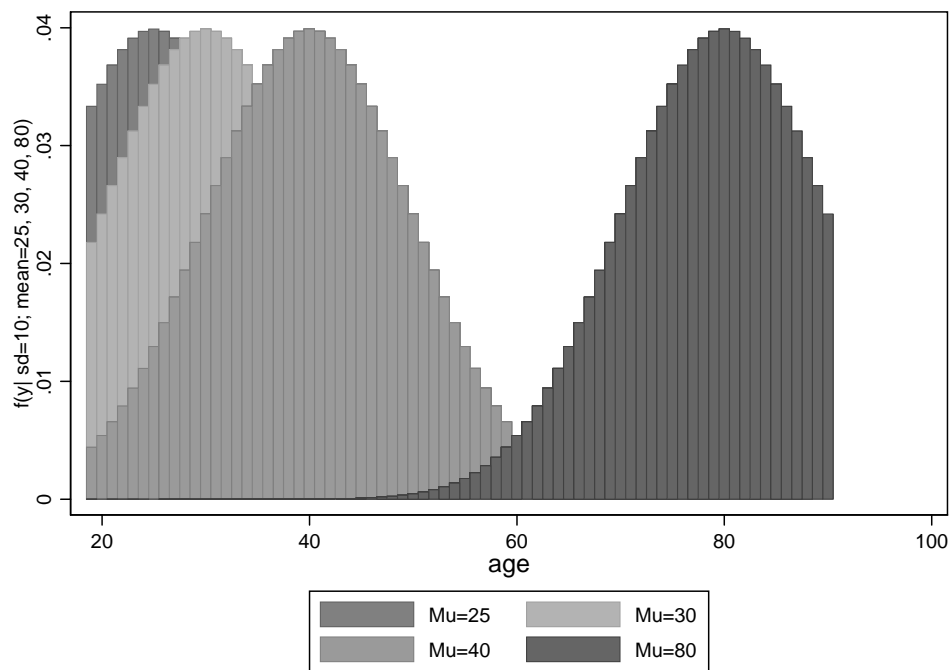
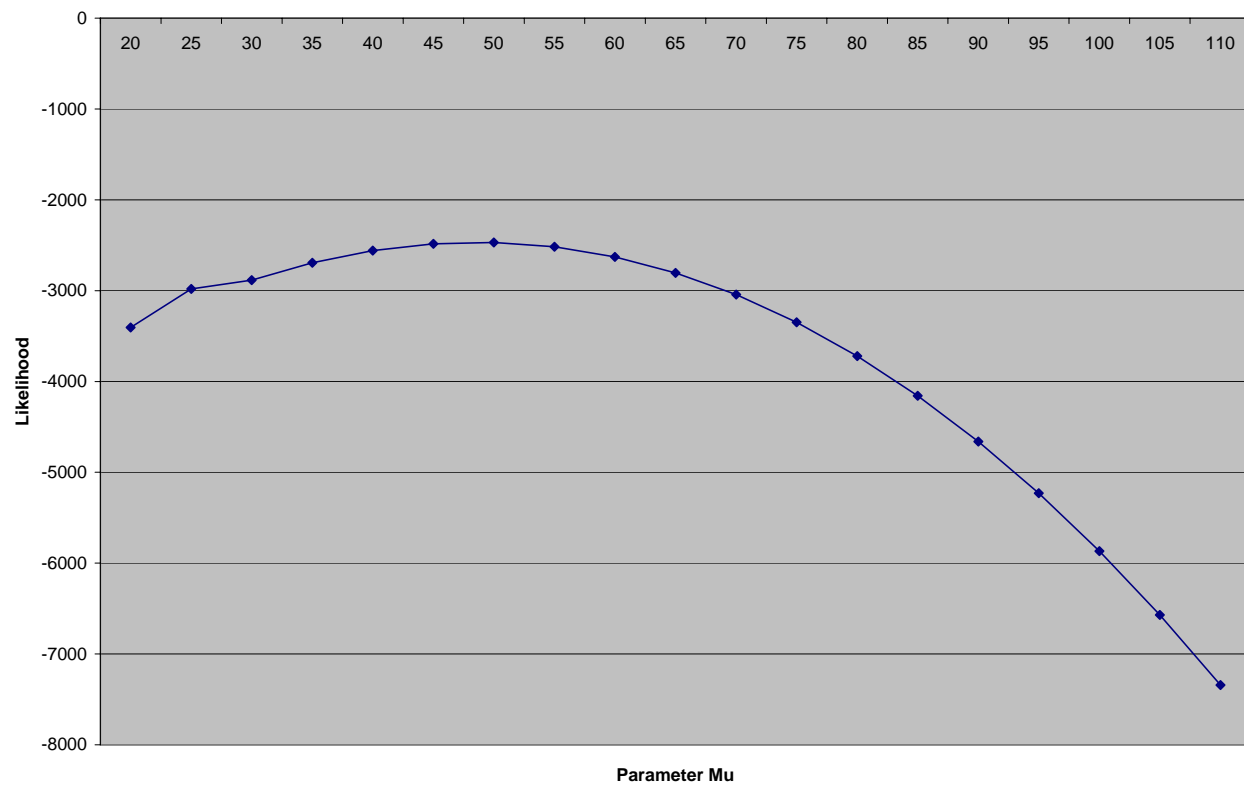


Figure 5: Likelihood Function for Normal Distribution with age data from NYT Survey



So:

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^n \left[ \frac{(y_i - \mu)^2}{\sigma^2} \right]$$

Take partials wrt  $\mu$ :

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) = 0$$

And take partials wrt  $\sigma^2$ :

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 = 0$$

Then solve simultaneously by multiplying the partial wrt  $\mu$  by  $\sigma^2$  and solving for  $\mu$ , then substitute into the other and solve for  $\sigma^2$ , and get:

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n$$

and

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2$$

These are the same as those derived by the first moment about zero and the second moment about the mean. That should be comforting.

## ML and the normal linear regression model

For our familiar regression model:

$$y_i = \mathbf{x}_i' \beta + \varepsilon_i$$

The likelihood function for a sample of  $n$  independent, identically and normally distributed disturbances is:

$$L = (2\pi\sigma^2)^{-n/2} e^{-\varepsilon'\varepsilon/(2\sigma^2)}$$

Since  $\varepsilon_i = y_i - x_i'\beta$ , we can substitute to get:

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(y-X\beta)'(y-X\beta)}$$

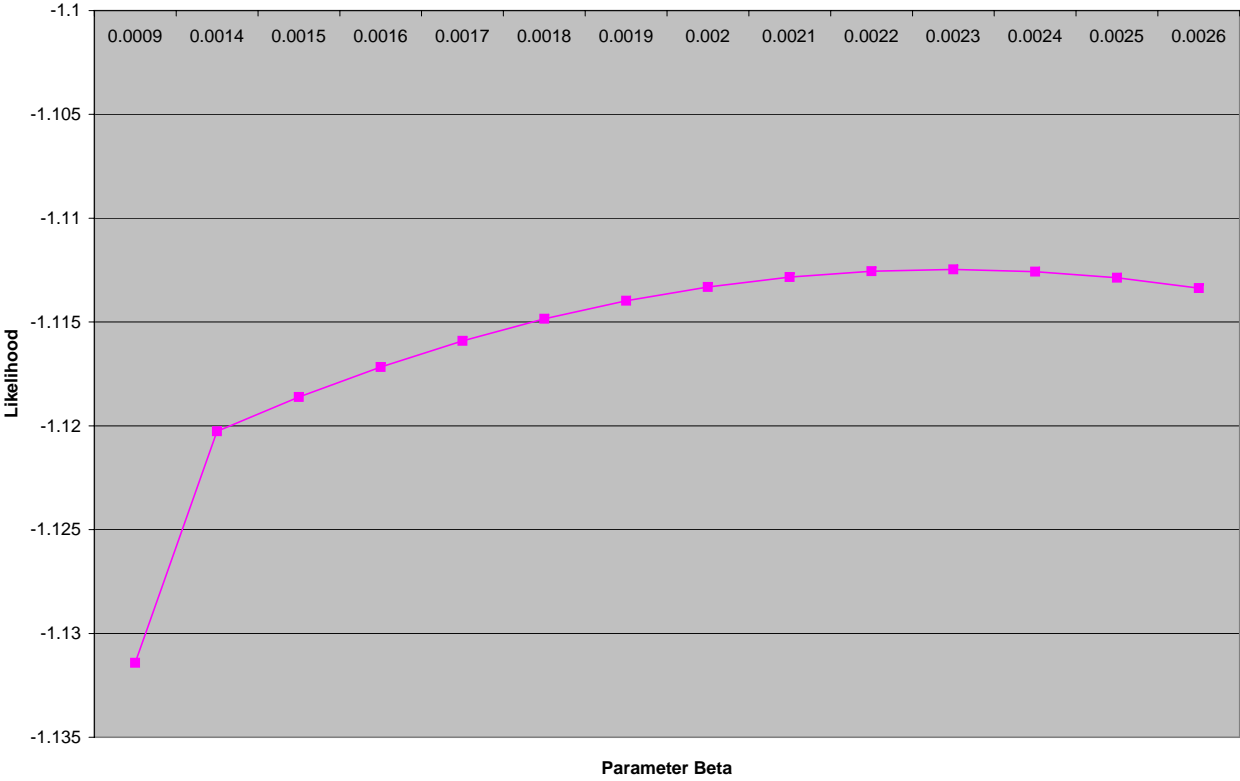
Take logs (in part, because it'll be easier to see how to maximize), we get the log-likelihood function for the classical normal regression model:

$$\ln L(\beta, \sigma^2 | y) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}$$

From the NYT Survey we might want to regress party id (democrat=1; independent=1.5; republican=2) on the age variable we examined earlier.

$$\ln L = \sum_{i=1}^n \left[ -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{[\text{partyid}_i - (b_0 + b_1 * \text{age}_i)]^2}{2\sigma^2} \right]$$

Figure 6: Likelihood Function for Linear Regression Party Id on Age data from NYT Survey



Take partials with respect to  $\beta$  and  $\sigma^2$  to get:

$$\frac{\partial \ln L}{\partial \beta} = \frac{X'(y - X\beta)}{\sigma^2} = 0$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4} = 0$$

Solving these simultaneously for  $\beta$  and  $\sigma^2$ , we get:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y = b$$

and:

$$\hat{\sigma}_{ML}^2 = \frac{e'e}{n}$$

So if our disturbances are normally distributed, the LS and ML estimators of  $\beta$  are identical.

But the ML estimator of  $\sigma^2$  is not equal to the LS estimator of  $\sigma^2$ . The LS estimator is  $s^2 = e'e/(n - K)$ , which we proved is an unbiased estimator of  $\sigma^2$ . So the ML estimator of  $\sigma^2$  is biased downward (i.e., it will be too small). But this is only problematic in small samples, as the difference between  $s^2$  and  $\sigma_{ML}^2$ , which is  $-K/n$ , will disappear in large samples.

## Properties of ML estimators

The properties of ML estimators, because they are derived from functions and the partials thereof, depend on the behavior of those functions. These are called “regularity conditions”

Consistent; asymptotically normal; asymptotically efficient; invariant

## Test procedures

There are three tests, all of which are asymptotically equivalent:

### Likelihood ratio test

If  $\theta$  is a vector of parameters, then  $\hat{\theta}_U$  is the ML estimator of  $\theta$  without restrictions, and  $\hat{\theta}_R$  is the estimator with the constraints (restrictions). We have an “unrestricted” log of likelihood,  $\hat{L}_U$ . That will be the maximum (by construction); the “restricted” log of likelihood,  $\hat{L}_R$ , by

definition, will be smaller. If a restriction is valid, then the likelihood of the restriction won't cause a large reduction in the log of likelihood. The likelihood ratio is:

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}$$

Both of these likelihoods are positive, and since  $\hat{L}_U > \hat{L}_R$ , these two conditions mean that  $\lambda$  must be between zero and one.

The large-sample distribution of  $-2\ln\lambda$  is chi-squared with  $df =$  number of restrictions imposed.

**Wald**

**LM test**