

Maximum Likelihood, part Deux

Paul Johnson

19th August 2004

1 Review

Please review the earlier handout on maximum likelihood analysis of the OLS model. Take special note of the result that the maximum likelihood estimator for is identical to the OLS estimator for the linear model.

2 Some new terms! Some old terms!

I have just noticed that the terms employed by econometricians, who have supplied the biggest part of my training in statistics, are different than the ones employed by statisticians and researchers in the applied sciences. The Generalized Linear Model is a creature of statisticians and applied researchers, so perhaps I should have not been surprised to find it has such a different vocabulary. If you look in the leading text by William Greene, *Econometric Analysis, 5ed*, you don't find a chapter on the GLM. You find many elements that are mathematically equivalent, but the research strategy is different.

I find that the stats books are sometimes "opaque" to my eye, possibly because I don't have the training they presuppose. They make many casual claims that appear, to me, to be completely unfounded. For example, recently while studying the GLM and maximum likelihood, I saw many assertions about the so-called score function that seemed ungrounded.

I figure if I don't get it, then it is almost certainly true of the grad students in political science won't either. So I'm trying to summarize the "big new" ideas as they might find application in books on the GLM. (But, in all honesty, when I really need to understand a mathematical result, I almost always find myself back with Greene's book, or another by Greenberg and Webster called *Advanced Econometrics: A Bridge to the Literature*.)

All of these things are mathematically identical between fields, the only differences are terminology.

2.1 Likelihood and log Likelihood functions

The sample is $y = (y_1, y_2, \dots, y_N)$. Each y_i is drawn from some distribution. The probability each observation is given by a probability model that you provide. Lets suppose that the parameters of the distribution are $\theta = (\theta_1, \theta_2)$ and the probability is given by a formula $f(y_i|\theta)$.

The Likelihood of observing the sample of size N is

$$L(\theta) = \prod_{i=1}^N f(y_i|\theta) \quad (1)$$

Apply the log to convert the big product (\prod) to a sum (\sum)

$$\ln L(\theta) = \sum_{i=1}^N \ln(f(y_i|\theta)) \quad (2)$$

2.2 The Score

The vector of first partial derivatives of the log likelihood is called the **score function**, sometimes Fischer's score function, in honor of a famous statistician who pioneered maximum likelihood. People often refer to the score function as $U(\theta)$. Don't forget it is really a vector, with one term for each parameter being estimated:

$$U(\theta) = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_1} \\ \frac{\partial \ln L}{\partial \theta_2} \end{bmatrix}$$

Recalling that $\ln L$ is a sum of N terms, and that the derivative is a linear operator, then it is true that

$$\frac{\partial \ln L}{\partial \theta_1} = \frac{\partial f(y_1|\theta)}{\partial \theta_1} + \frac{\partial f(y_2|\theta)}{\partial \theta_2} + \dots + \frac{\partial f(y_N|\theta)}{\partial \theta_N} \quad (3)$$

So you could think of the score function as the sum of scores of individual observations. You might call the score for an individual observation $u_i(\theta)$, or some other letter if the u bothers you.

2.3 First Order Conditions

In maximum likelihood analysis, we maximize log Likelihood by choosing the best combination of (θ_1, θ_2) . Take partial derivatives with respect to θ_1 and θ_2 and set them equal to 0.

$$\begin{aligned}\frac{\partial \ln L}{\partial \theta_1} &= 0 \\ \frac{\partial \ln L}{\partial \theta_2} &= 0\end{aligned}\tag{4}$$

These are the **first order conditions** for a maximum point. There are 2 equations with 2 unknowns. Statisticians use the term “maximum likelihood score equations” to refer to the system in 4. Green simply calls them the “likelihood equations.” If one sets the score function equal to 0, as in a matrix equation, one has

$$U(\theta) = \frac{\partial \ln L}{\partial \theta} = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_1} \\ \frac{\partial \ln L}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}\tag{5}$$

2.4 The solution of the score equations is the MLE.

When the score function is set equal to 0, one has the maximum likelihood score equations.

Assuming

1. the probability model is “regular,” in the sense that it is mathematically continuous and differentiable and has finite expected values (Greene, p. 474).
2. the point $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ can be found at which both equations are equal to 0, and
3. at that point, $\ln L(\hat{\theta})$ is a maximum point (rather than a minimum or saddle point)

then $\hat{\theta}$ is a maximum likelihood estimate.

2.5 Interesting Mathematical Fact: $E[U(\theta)] = 0$.

2.5.1 $U(\theta), \frac{\partial \ln L}{\partial \theta}$ are random variables.

Obviously, since the observations on y_i are random, and those values are used to calculate the probability of a particular outcome, then the derivative is also a random variable.

2.5.2 At the MLE $\hat{\theta}, E[U(\theta)] = 0$.

I have seen this claim assumed in many stats books and I always wondered why. It turns out it is not an “obvious” thing. There’s a proof in Greene (p. 475). Actually, Greene proves a stronger result. Greene shows that the derivative of $\ln L(y_i|\theta)$, the score value for each observation, has an expected value of 0. That is, at the MLE,

$$E \left[\frac{\partial \ln L(y_i|\theta)}{\partial \theta_i} \right] = 0 \text{ for all } i.$$

And, naturally, the sum of those expected values is 0, so the expected value of the score is as well: $E[U(\hat{\theta})] = 0$.

Greene’s explanation of this is not all that complicated. It relies only on results you could find in a first-year calculus book. If you are willing to just believe the result, move on. I did for a long time. Otherwise read Greene. Or consider this “story” about it, and then you will understand fully if you read Greene.

Remember that expected value means a “probability weighted sum of observations.” For a discrete variable y ,

$$E[y] = \sum f(y_i) \cdot y_i$$

or, for a continuous variable,

$$E[y] = \int f(y) \cdot y \, dy.$$

The same works for expectations of functions. Supposing $U(y)$ is a function of y :

$$E[U(y)] = \sum f(y_i) \cdot U(y_i)$$

or

$$E[U(y)] = \int f(y) \cdot U(y) \, dy$$

The probability may depend on some parameter (or collection of parameters), θ , and that is written $f(y_i|\theta)$.

The definition of a probability distribution is

$$\int f(y_i|\theta) \, dy_i = 1 \text{ or } \int f(y_i|\theta) - 1 = 0$$

Take the derivative with respect to θ :

$$\frac{\partial \int f(y_i|\theta) \, dy_i}{\partial \theta} = 0$$

Next, apply Leibnitz Theorem:

$$\frac{\partial}{\partial \theta} \int f(y_i|\theta) dy_i = \int \frac{\partial f(y_i|\theta)}{\partial \theta} dy_i = 0$$

Leibnitz theorem, usually covered in the first year of calculus: The derivative of an integral is the integral of the derivative (or something roughly like that, where I'm assuming away the problem about the limits of the integral that might change as a function of θ .)

Then comes the sneaky part, the part I would not have thought of on my own. Greene observes:

$$\int \frac{\partial f(y_i|\theta)}{\partial \theta} dy_i = \int f(y_i|\theta) \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy_i \quad (6)$$

How do you get from the left to the right? Recall $\frac{\partial \ln(y)}{\partial y} = \frac{1}{y}$ and by the chain rule, $\frac{\partial \ln[f(y)]}{\partial y} = \frac{1}{f(y)} \frac{\partial f(y)}{\partial y}$. Rearrange that to solve for $\frac{\partial f(y)}{\partial y}$

$$\frac{\partial f(y)}{\partial y} = \frac{\partial \ln[f(y)]}{\partial y} f(y) \quad (7)$$

Use that little tidbit in the left hand side of 6, and you get the right hand side. And that means the proof is finished, because the right hand side is equal to the expected value of the partial derivative of $\ln[f(y_i|\theta)]$.

There's another exciting result, quite a bit like this, that awaits concerning the variance of $U(\theta)$. Just wait.

2.6 Second Order Conditions: The Hessian

The Hessian matrix is the matrix of second derivatives. Take each element of $U(\theta)$ as represented in 5. Differentiate each element by each of the parameters, you arrive at a partial derivatives turns into a 2x2 matrix:

$$H(\theta) = \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln L}{\partial \theta_2^2} \end{bmatrix} \quad (8)$$

Of course, if you had 10 parameters, you would have a 10x10 matrix.

The Hessian provides the **second order conditions** that indicate whether the point at which the partial derivatives are equal to 0 is a maximum. If we have found a maximum point, then we know for sure that $\frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_2}$ and $\frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_2}$ must be negative.

(make a sketch)

It could be one has found a minimum or a “saddlepoint.” I have several econometrics texts that would help you to understand the meaning of the second order conditions. But it seems to me that you can understand this on a conceptual level, and that is good enough.

Here’s the “big idea.” Consider the score equation, 5. Suppose that the MLE is found and, furthermore, suppose that the score function is very sharply peaked at the solution point. In that case, one is highly confident with the choice of a particular estimate; at the top of a sharp mountain, it is clear where the maximum is to be found. The estimate is precise. If that is the case, the diagonal elements of $H(\theta)$ will be negative numbers that are large in magnitude. The fit of the model is changing dramatically as one moves away from the solution.

Suppose, on the other hand, the score is a nearly flat mound. One is not very confident of the estimate of θ because the neighboring values of θ are nearly as good. In that case, the diagonal will have negative numbers of small magnitude.

Later on in the story, you find out that

$$\text{Var}(\hat{\theta}) = [-H(\hat{\theta})]^{-1}$$

So the Hessian is pretty important.

2.7 $Var[U(\theta)] = -E[H]$

This is another result that is frequently asserted and I had never bothered to find out why until recently. It is the companion to the result in section ??.

The argument is described in detail in Greene (p. 475). As in the previous case, he shows the result is true for a single observation i . Begin with the result stated above in ??. That is

$$\int f(y_i|\theta) \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy_i = 0 \quad (9)$$

Differentiating under the integral (Leibnitz rule again),

$$\int \frac{\partial f(y_i|\theta)}{\partial \theta} \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy + \int f(y_i|\theta) \frac{\partial^2 \ln([f(y|\theta)]}{\partial \theta \partial \theta} dy = 0$$

which one can easily see is:

$$- \int f(y_i|\theta) \left[\frac{\partial^2 \ln([f(y|\theta)]}{\partial \theta \partial \theta} \right] dy = \int \frac{\partial f(y_i|\theta)}{\partial \theta} \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy \quad (10)$$

The left hand side is $-E[H]$, so we are almost finished.

Concentrate on the right hand side. Then take a look back at the linchpin “secret trick” in 7. If you use that in the right hand side, it becomes

$$\begin{aligned} \int \frac{\partial f(y_i|\theta)}{\partial \theta} \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy &= \int f(y_i|\theta) \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} \frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} dy \\ &= \int f(y_i|\theta) \left(\frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} \right)^2 dy \end{aligned} \quad (11)$$

The right hand side is $Var[U(\theta)]$. Can you see it? Recall the definition of the variance:

$$Var[U(\theta)] = \int f(y_i|\theta) (U(\theta) - E[U(\theta)])^2 dy$$

And, since $E[U(\theta)] = 0$, that becomes:

$$Var[U(\theta)] = \int f(y_i|\theta) U(\theta)^2 dy$$

which is just

$$Var[U(\theta)] = \int f(y_i|\theta) \left(\frac{\partial \ln[f(y_i|\theta)]}{\partial \theta} \right)^2 dy$$

2.8 The Information Matrix

The result stated in section 2.7 is important because it eventually leads to estimates of the variance of the ML parameter estimates. Because it is so important, the name **information matrix** is given to $-E[H]$. We might as well write it out, for the fun:

$$Info(\theta) = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_2} \end{bmatrix} \quad (12)$$

In many books, one finds the assertion that the Information Matrix is just $-H$, but I believe that is a mistake, or a simplification.

The simplification $Info(\theta) = -H$ is used because the “correct” information matrix may be impossible to calculate.

2.9 Asymptotically, $Var(\hat{\theta}) = Info(\theta)^{-1}$

As the sample size tends to infinity, the variance of the MLE is the inverse of the information matrix. This is another of the ML claims that is frequently asserted, seldom explained.

Greene (p. 478) gives the argument. This requires a Taylor series approximation and an invocation of the Lindberg-Levy central limit theorem, and I have not found a way to explain it all in a simple way. But I can give some hints.

Recall the score equation, evaluated at the MLE $\hat{\theta}$

$$U(\hat{\theta}) = 0$$

Suppose the true parameter value is θ_0 . We want to approximate the score in the vicinity of that value.

The first two terms of the Taylor series approximation of $U(\theta)$ are

$$U(\hat{\theta}) = U(\theta_0) + \frac{\partial U(\tilde{\theta})}{\partial \theta}(\hat{\theta} - \theta_0) = U(\theta_0) + H(\tilde{\theta})(\hat{\theta} - \theta_0) = 0$$

The mean value theorem implies that there is some value $\tilde{\theta}$ which can make the equality hold. Rearrange:

$$U(\theta_0) = -H(\tilde{\theta})(\hat{\theta} - \theta_0)$$

$$(\hat{\theta} - \theta_0) = [-H(\tilde{\theta})]^{-1} U(\theta_0)$$

As I examined Greene, p. 478-9, it seemed to me that was the really critical part. We've got the inverse of the negative Hessian matrix. After a sequence of rearrangements, invoking the fact that MLE are consistent (meaning $\hat{\theta} \rightarrow \theta_0$), and the Lindberg-Levy limit theorem, we arrive at the result that the MLE is Normally distributed, thus:

$$\hat{\theta} N[\theta_0, \text{Info}(\theta_0)^{-1}]$$

The estimate $\hat{\theta}$ converges "in distribution" to the Normal as the sample size approaches infinity, a Normal distribution with mean equal to the true parameter vector θ_0 and variance equal to the inverse of the information matrix evaluated at θ_0 .

3 Re-arranging the Normal distribution

This is just some fun stuff to remind you of some math.

3.1 Suppose $y \sim N(\mu, \sigma^2)$.

The probability of observing y_i depends on two parameters, μ and σ^2 .

$$\text{prob}(y_i|\mu, \sigma^2) = N(\mu, \sigma^2)$$

The usual formula is:

$$\text{prob}(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

Sometimes you will see it rearranged, putting σ symbols under squares or taking them out of square roots:

$$\text{prob}(y_i|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2} \tag{13}$$

We are accustomed to thinking of μ as the "mean" of y_i and σ^2 is the variance of y_i . But you might try to think of them as "just numbers", although it is shown below that the mean is, in fact, a maximum likelihood estimate of μ .

3.2 Put everthing under the exponential.

My favorite! Its useful in thinking about GLMs and ML. Regroup everything inside the exponential:

$$prob(y_i|\mu, \sigma^2) = exp \left[-\frac{1}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mu)^2 \right] \quad (14)$$

Convince yourself! Recall from the elementary algebra that

1. $exp(A + B) = exp(A) \cdot exp(B)$

2. $A = exp(ln(A))$,

and as a consequence of those 2 rules,

$$A * exp(B) = exp(ln(A)) * exp(B) = exp(ln(A) + B). \quad (15)$$

3. $ln(A * B) = ln(A) + ln(B)$

4. $ln(1) = 0$

5. $sqrt(x) = x^{\frac{1}{2}}$

6. $\frac{1}{x^{1/2}} = x^{-\frac{1}{2}}$

7. $ln(A^B) = B \cdot ln(A)$

As a consequence of 1 & 2, 13 can be written as

$$prob(y_i|\mu, \sigma^2) = exp \left[ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2}(y_i - \mu)^2 \right] \quad (16)$$

Focus on the first part inside the brackets:

$$\frac{1}{\sqrt{2\pi\sigma^2}} = (2\pi\sigma^2)^{-\frac{1}{2}}$$

so

$$ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) = ln[(2\pi\sigma^2)^{-\frac{1}{2}}] = -\frac{1}{2}ln(2\pi\sigma^2) = -\frac{1}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2)$$

Insert that into 16

$$prob(y_i|\mu, \sigma^2) = exp \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_i - \mu)^2 \right]$$

Note you could simplify this one step further by “canceling out” the coefficient and exponent in the middle term:

$$\frac{1}{2} \ln(\sigma^2) = \frac{1}{2} \cdot 2 \cdot \ln(\sigma) = \ln(\sigma)$$

but we don’t take that step because we want to keep the σ terms squared.

3.3 Two good things about equation 14.

3.3.1 It is pitifully easy to find the log likelihood!

$$\ln [exp [anything]] = anything$$

3.3.2 You see the Normal as an example of a so-called “exponential distribution).

If you massage 14 around a little bit, you will find that the Normal can be written in a form like this:

$$prob(y_i|\mu, \sigma^2) = exp [a(y_i\mu) + b(y_i) + c(\mu)]$$

Inside the bracket, there is a sum of 3 pieces: a function of the product of y_i and μ , a function of y_i , and a function of μ . It turns out there are many distributions that fit into this framework, and the study of regression models that do so is the topic of Generalized Linear Modeling (GLM).

Don’t fixate on the labels a , b , and c in this expression, because they are not the vital thing. The vital thing is that we can get everything under the $exp()$ and that the parts, y_i , and μ , are arranged in a pattern.

4 The best estimate of the parameter μ is the mean of y !

Because of result 14, it is horribly easy to get maximum likelihood estimates. That’s so because the sum the log likelihoods is so simple.

The log likelihood of the entire sample is the sum of the log likelihoods, and look how easily that breaks down with the normal in equation 14:

$$\begin{aligned}
 \ln L(\mu, \sigma^2) &= \sum_{i=1}^N \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_i - \mu)^2 \right] & (17) \\
 &= -\frac{1}{2} \sum_{i=1}^N \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \mu)^2 \\
 &= -\frac{1}{2} \cdot N \cdot \ln(2\pi) - \frac{1}{2} \cdot N \cdot \ln(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - \mu)^2 \\
 &= -\frac{N}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - \mu)^2
 \end{aligned}$$

Could it get any easier than that? (Don't answer out loud, please.)

You want to maximize that by adjusting the values of μ and σ^2 . Ignore the first part (that does not at all depend on either μ nor σ^2). Throw away the $\frac{1}{2}$ in the front of each term, because removing that does not change the location of the maximum. So the log likelihood is proportional to a much simpler thing (the symbol \propto means "is proportional to"):

$$\ln L(\mu, \sigma^2) \propto -N \cdot \ln(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - \mu)^2$$

The variance σ^2 is a "nuisance parameter." In fact, if you look at this, you realize that, **NO MATTER WHAT** value you put in for σ^2 , the optimal value of μ is not affected. The best estimate is the value of $\hat{\mu}$ that maximizes this:

$$- \sum_{i=1}^N (y_i - \hat{\mu})^2$$

Which is -1 times the sum of squared deviations about $\hat{\mu}$. And you find

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

and

$$\sum_{i=1}^N y_i - N \cdot \hat{\mu} = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^N y_i}{N}$$

Result: The maximum likelihood estimate of the parameter μ is the mean of the observations.

5 Back to regression.

5.1 The Likelihood Function

In the regression model, we proceed as though the observations of $y_i \sim N(\mu_i, \sigma^2)$. For each observation, however, there is a different μ parameter in OLS.

Suppose

$$y = X\beta + e$$

As usual, y 's an $N \times 1$ vector, X is a $N \times p$ matrix, β is a $p \times 1$ vector, and e is a vector. Each observation in e , $e_i \sim N(0, \sigma^2)$.

If X_i is the i 'th row,

$$\mu_i = X_i\beta$$

Then one can think of y_i as if it followed the $N(X_i\beta, \sigma^2)$ distribution. That means we need to estimate β , rather than μ . Compare against equation 14

$$\text{prob}(y_i|\beta, \sigma^2) = \exp \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_i - X_i\beta)^2 \right] \quad (18)$$

$$\ln L(\beta, \sigma^2) = \sum_{i=1}^N \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_i - X_i\beta)^2 \right] \quad (19)$$

which implies

$$\ln L(\beta, \sigma^2) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - X_i\beta)^2$$

and if you prefer, you can replace $\sum (y_i - X_i\beta)^2$ with vectors $(y - X\beta)'(y - X\beta)$:

$$\ln L(\beta, \sigma^2) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \quad (20)$$

The score function is the vector of derivatives

$$U(\beta) = \begin{bmatrix} \frac{\partial \ln L}{\partial \beta_1} \\ \frac{\partial \ln L}{\partial \beta_2} \\ \vdots \\ \frac{\partial \ln L}{\partial \beta_p} \end{bmatrix}$$

Why doesn't the score function include σ^2 ? Just convenience, I believe. I think it could be included, but it is not because estimation (in practice) proceeds in two steps. We calculate $\hat{\beta}$ first, then an estimate of σ^2 can be calculated.

The first order condition is

$$U(\beta) = 0$$

5.2 MLE Equals OLS

Give a casual glance at the objective function in 20 and you can tell that the first two terms don't matter because they don't depend on β . As a result, maximizing 20 is the same as maximizing

$$-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

Which is the same as minimizing

$$(y - X\beta)'(y - X\beta)$$

which is just the sum of squared residuals. So MLE is mathematically identical to OLS.

5.3 Solving the system

In Myers, Montgomery, and Vining, p. 32, they show the steps to solve that, although I find had some trouble retracing one step (perhaps there is a typographical error). So maybe it is worthwhile to write it down. This looks like one equation, but really it is a matrix equation with the number of rows equal to p . The solution is the "right" value of $\hat{\beta}$:

$$\ln L(\beta, \sigma) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y'y - \beta'X'y - y'X\beta + \beta'X'X\beta)$$

Since $y'X\beta$ is a scalar (a 1×1 matrix), it is equal to its transpose, $y'X\beta = y'X\beta$, so this reduces to :

$$\ln L(\beta, \sigma) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y'y - 2\beta'X'y + \beta'X'X\beta)$$

The first order condition is

$$U(\hat{\beta}) = \frac{\partial \ln L}{\partial \beta} = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \beta} (-2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}) = 0$$

$$(-2X'y + 2X'X\hat{\beta}) = 0$$

$$2X'X\hat{\beta} = 2X'y$$

$$X'X\hat{\beta} = X'y$$

If the inverse of $(X'X)$ exists, the solution is:

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

And the MLE for σ^2 is found by differentiating 20 with respect to σ^2 .

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} - \frac{1}{2\sigma^4} (y - X\hat{\beta})(y - X\hat{\beta}) = 0$$