# Maximum Likelihood, part Deux 

Paul Johnson

July 15, 2011

These notes were put together while I was studying the Generalized Linear Model (GLM) and building up to study Mixed models and GEE.

## 1 Review

Please review the earlier handout on maximum likelihood analysis of the OLS model.

Note that minimizing the sum of squares (as in OLS) involves minimizing a sum:

$$
\min \sum_{i=1}^{N}\left(y_{i}-\hat{y}_{i}\right)^{2}, \text { where } \hat{y}_{i}=f(\hat{b}, X)
$$

Maximizing a log likelihood for a normally distributed dependent variable ends up being the exact same goal.

$$
\max -\sum\left(y_{i}-\hat{y}_{i}\right)^{2}, \text { where } \hat{y}_{i}=f(\hat{b}, x)
$$

Maximizing the log likelihood leads to the same estimates of the slope parameters as does minimizing the sum of squares. In other words, we are still in the same business of formulating a predictive model $f(\hat{b}, X)$ and finding out how well it fits.

Now, suppose the dependent variable is not normal. Perhaps the observed $y_{i}$ is dichotomous, or a count, or it is truncated or skewed. In those cases, it is harder to stretch the math to make an OLS fit seem sensible, but it is often plain to see that maximizing the likelihood is a not-unreasonable approach. You can stipulate any distribution you like for $y_{i}$ as a function of the data and parameters.

Sometimes we will find that ML solutions are impractical, so we have to use other estimation principles, such as "method of moments" or "Bayesian MCMC." Nevertheless, the ML approach is still preferred when it is practical.

## 2 Some new terms! Some old terms!

My goal here is to put light on some of the terms and results that GLM practitioners commonly refer to, but seldom explain in depth.

Before I worked with psychologists, I though it was frustrating to go between audieneces of economists, statisticians, and political scientists. In some ways, the babel in my head has grown worse as I work with a new group, but it has also started to clear up some points of confusion.

I see now that the underlying mathematical model is generally the same as one travels among audiences, but the names used for the elements are different. The terms and interpretations of the technical fundamentals will differ.

The economists are not so emphatic about the terminology Generalized Linear Model as are statisticians. This caused me a great deal of confusion, because one simply cannot do regression work in R without a solid understanding of the GLM terminology. Consider William Greene's Econometric Analysis, 5ed. There is no chapter on the GLM. Nevertheless, Greene's nearly encyclopedic coverage of the mathematical underpinnings is unparalleled, and if one is trying to find a proof of the supposedly "elementary" or "fundamental" truths of statistics, it may be the best place to look. As I go between Greene's discussion of maximum likelihood and the discussions of statisticians and social scientists, I am often pressed to translate the claims they make into the vocabularies of the other fields. In the following, I name some of these terms and try to justify them.

### 2.1 Likelihood and log Likelihood functions

The sample is $y=\left(y_{1}, y_{2}, \ldots, y_{N}\right)$. 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 $\theta=\left(\theta_{1}, \theta_{2}\right)$ and the probability is given by a formula $f\left(y_{i} \mid \theta\right)$.

The Likelihood of observing the sample of size N is

$$
\begin{equation*}
L(\theta)=\prod_{i=1}^{N} f\left(y_{i} \mid \theta\right) \tag{1}
\end{equation*}
$$

Apply the log to convert the big product $(\Pi)$ to a sum $(\Sigma)$

$$
\begin{equation*}
\ln L(\theta)=\sum_{i=1}^{N} \ln \left(f\left(y_{i} \mid \theta\right)\right) \tag{2}
\end{equation*}
$$

## Dramatic Foreshadowing:

Recall that if $f\left(y_{i}\right)=\exp \left(y_{i}\right)$, then $\ln \left(f\left(y_{i}\right)\right)=y_{i}$. So if you work with distributions that are "exponential" in nature, then you get a RADICAL simplification in the formulae after applying the natural log. To take a look at most of the distributions that you use very often and check to see how much simpler they get after you log $f$. Remember that $\ln \left(a^{b}\right)=b \cdot \ln (a)$.

### 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:

$$
U(\theta)=\left[\begin{array}{c}
\frac{\partial \ln L}{\partial \theta_{1}} \\
\frac{\partial \ln L}{\partial \theta_{2}}
\end{array}\right]
$$

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

$$
\begin{equation*}
\frac{\partial \ln L}{\partial \theta_{1}}=\frac{\partial f\left(y_{1} \mid \theta\right)}{\partial \theta_{1}}+\frac{\partial f\left(y_{2} \mid \theta\right)}{\partial \theta_{1}}+\ldots+\frac{\partial f\left(y_{N} \mid \theta\right)}{\partial \theta_{1}} \tag{3}
\end{equation*}
$$

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 maximimum likelihood analysis, we maximize log Likelihood by choosing the best combination of $\left(\theta_{1}, \theta_{2}\right)$. Take partial derivatives with respect to $\theta_{1}$ and $\theta_{2}$ and set them equal to 0 .

$$
\begin{align*}
& \frac{\partial \ln L}{\partial \theta_{1}}=0  \tag{4}\\
& \frac{\partial \ln L}{\partial \theta_{2}}=0
\end{align*}
$$

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

$$
U(\theta)=\frac{\partial \ln L}{\partial \theta}=\left[\begin{array}{c}
\frac{\partial \ln L}{\partial \theta_{1}}  \tag{5}\\
\frac{\partial \ln L}{\partial \theta_{2}}
\end{array}\right]=\left[\begin{array}{l}
0 \\
0
\end{array}\right]
$$

### 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 $\hat{\theta}=\left(\hat{\theta}_{1}, \hat{\theta}_{2}\right)$ 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 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. It turns into a $2 \times 2$ matrix:

$$
H(\theta)=\frac{\partial^{2} \ln L}{\partial \theta \partial \theta^{\prime}}=\left[\begin{array}{cc}
\frac{\partial^{2} \ln L}{\partial \theta_{1}^{2}} & \frac{\partial^{2} \ln L}{\partial \theta_{1} \partial \theta_{2}}  \tag{6}\\
\frac{\partial^{2} \ln L}{\partial \theta_{1} \partial \theta_{2}} & \frac{\partial^{2} \ln L}{\partial \theta_{2}^{2}}
\end{array}\right]
$$

Of course, if you had 10 parameters, you would have a $10 \times 10$ matrix.
The Hessian is also thought of as $\partial U / \partial \theta^{\prime}$.
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 $\frac{\partial^{2} \ln L}{\partial \theta_{1} \partial \theta_{1}}$ and $\frac{\partial^{2} \ln L}{\partial \theta_{2} \partial \theta_{2}}$ must be negative. There is also a condition that restricts the values of the other terms to be within a certain range.
(make a sketch)

### 2.6 Root Finding: The Score and the Hessian work together

It is easy to sit and say that you will find a value of $\hat{\theta}$ that maximizes the likelihood.
It is sometimes very difficult to actually make the calculations. For some models, you can actually solve algebraically to get $\hat{\theta}$ in a clear, algebraic formula. Many times that cannot be done.

Review my handout on "approximations". We need to find the "roots" of the score equations, the values at which $U(\theta) \hat{=} 0$. Applying Newton's method to find the value $\hat{\theta}$ for which $U(\hat{\theta})=0$, one applies an algorithmic process

$$
\hat{\theta}_{\text {new }}=\hat{\theta}_{\text {old }}-H\left(\hat{\theta}_{\text {old }}\right)^{-1} \cdot U\left(\hat{\theta}_{\text {old }}\right)
$$

Do that over and over again, until there is only minimal change in the value of the score. It becomes close to 0 , but because of rounding errors, it is never exactly 0 .

The various approaches to maximization are variations on that theme.
Alternative methods of doing these calculations are usually just slightly different. The method of Fisher Scoring replaces the Hessian matrix with the expected value of the second derivative matrix.

## 3 Estimate the Variance of $\hat{\theta}$.

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 $\theta$ because the neighboring values of $\theta$ are nearly as good. In that case, the diagonal will have negative numbers of small magnitude.

In a later section of these notes, I present a description of two vital results in maximum likelihood. Those results lead up to the claim that

$$
\begin{equation*}
\operatorname{Var}(U(\theta))=-E[H] \tag{7}
\end{equation*}
$$

and, from there, it is only a "hop, skip and jump" to the most important claim, which is the variance of the estimated parameter, $\operatorname{Var}[\hat{\theta}]$ can be consistently estimated.

### 3.1 The Information Matrix

The result stated in equation7 (and proven below in section 5.2) 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:

$$
\operatorname{Info}(\theta)=-E\left[\begin{array}{ll}
\frac{\partial^{2} l n L}{\partial \theta_{1} \partial \theta_{1}} & \frac{\partial^{2} l n L}{\partial \theta_{1} \partial \theta_{2}}  \tag{8}\\
\frac{\partial^{2} l n L}{\partial \theta_{2} \partial \theta_{1}} & \frac{\partial^{2} l n L}{\partial \theta_{2} \partial \theta_{2}}
\end{array}\right]
$$

In many books, one finds the assertion that the Information Matrix can be estimated as $-H$, but that is true only for some (I think a broad class) of models.

### 3.2 Asymptotically, $\operatorname{Var}(\hat{\theta})=\operatorname{Info}(\theta)^{-1}$ and $\hat{\theta}$ is Normal!

In words: 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 frequenty 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 $\hat{\theta}$

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

Suppose the true parameter value is $\theta_{0}$. We want to approximate the score in the vicinity of that value (because $\hat{\theta}$ is consistent, then for a large sample it is "tending to" $\theta_{0}$ ).

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

$$
U(\widehat{\theta})=U\left(\theta_{0}\right)+\frac{\partial U(\widetilde{\theta})}{\partial \theta}\left(\widehat{\theta}-\theta_{0}\right)=U\left(\theta_{0}\right)+H(\widetilde{\theta})\left(\widehat{\theta}-\theta_{0}\right)=0
$$

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

$$
\begin{aligned}
U\left(\theta_{0}\right) & =-H(\widetilde{\theta})\left(\widehat{\theta}-\theta_{0}\right) \\
\left(\widehat{\theta}-\theta_{0}\right) & =[-H(\widetilde{\theta})]^{-1} U\left(\theta_{0}\right)
\end{aligned}
$$

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 $\widehat{\theta} \rightarrow \theta_{0}$ ), and the Lindberg-Levy limit theorem, we arrive at the result that the MLE is Normally distributed, thus:

$$
\begin{equation*}
\hat{\theta} \sim N\left[\theta_{0}, \operatorname{Info}\left(\theta_{0}\right)^{-1}\right] \tag{9}
\end{equation*}
$$

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 $\theta_{0}$ and variance equal to the inverse of the information matrix evaluated at $\theta_{0}$.

## 4 What's all that good for?

### 4.1 Significance tests for single parameters.

Your old friend the $t$ test might be be tempting. Suppose the null hypothesis is 0 . You might calculate:

$$
\frac{\hat{\theta}}{\sqrt{\operatorname{Var}(\hat{\theta})}}=\frac{\hat{\theta}}{\operatorname{Std.Error}(\hat{\theta})}
$$

and act as if it were a t statistic. Many people have done so.
That is not exactly a $t$ variable, however, because, as you recall, a $t$ statistic is actually an "exact distribution" that depends on your number of cases (degrees of freedom). This number here is something else because we don't have an exact estimate of the standard error, but rather an asymptotic approximation of it. At best, it is an asymptotically valid $t$ test.

An alternative is Wald's test:

$$
\frac{\hat{\theta}^{2}}{\operatorname{Var}(\theta)} \sim \chi^{2}
$$

If a variable is $\chi^{2}$, its square root is Normal, so Wald's test looks like your old friend

$$
\frac{\hat{\theta}}{\operatorname{Std.Error}(\hat{\theta})}
$$

which looks an awful lot like a $t$ statistic to me, but it isn't really.

### 4.2 Significance tests for groups of parameters.

Wald's test can be stated in matrix form so that you can test several parameters at once, as in

$$
\left(\hat{\theta}-\theta_{\text {null }}\right)^{\prime} V^{-1}\left(\hat{\theta}-\theta_{\text {null }}\right)
$$

or, for an example with two coefficients being tested against null values of 0

$$
\left[\hat{\theta}_{1}, \hat{\theta}_{2}\right]\left[\begin{array}{cc}
\operatorname{Var}\left(\hat{\theta}_{1}\right) & \operatorname{Cov}\left(\hat{\theta}_{1}, \hat{\theta}_{2}\right) \\
\operatorname{Cov}\left(\hat{\theta}_{1}, \hat{\theta}_{2}\right) & \operatorname{Var}\left(\hat{\theta}_{2}\right)
\end{array}\right]^{-1}\left[\begin{array}{l}
\hat{\theta}_{1} \\
\hat{\theta}_{2}
\end{array}\right]
$$

## 5 Stupefying Mathematical Facts That are Required if you want to believe the previous sections

### 5.1 $\quad$ Fact: $E[U(\hat{\theta})]=0$.

5.1.1 Remember $U(\theta), \frac{\partial l n L}{\partial \theta}$ are random variables.

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

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 $\ln L\left(y_{i} \mid \theta\right)$, the score value for each observation, has an expected value of 0 . That is, at the MLE,

$$
E\left[\frac{\partial \ln L\left(y_{i} \mid \hat{\theta}\right)}{\partial \theta_{i}}\right]=0 \text { for all } i .
$$

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 presentation 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 $y$,

$$
E[y]=\sum f\left(y_{i}\right) \cdot y_{i}
$$

or, for a continuous variable,

$$
E[y]=\int f(y) \cdot y d y
$$

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

$$
E[U(y)]=\sum f\left(y_{i}\right) \cdot U\left(y_{i}\right)
$$

or

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

The probability may depend on some parameter (or collection of parameters), $\theta$, and that is written $f\left(y_{i} \mid \theta\right)$.

The definition of a probability distribution is

$$
\int f\left(y_{i} \mid \theta\right) d y_{i}=1 \text { or } \int f\left(y_{i} \mid \theta\right)-1=0
$$

Take the derivative with respect to $\theta$ :

$$
\frac{\partial \int f\left(y_{i} \mid \theta\right) d y_{i}}{\partial \theta}=0
$$

Next, apply Leibnitz Theorem:

$$
\frac{\partial}{\partial \theta} \int f\left(y_{i} \mid \theta\right) d y_{i}=\int \frac{\partial f\left(y_{i} \mid \theta\right)}{\partial \theta} d y_{i}=0
$$

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 $\theta$.)

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

$$
\begin{equation*}
\int \frac{\partial f\left(y_{i} \mid \theta\right)}{\partial \theta} d y_{i}=\int f\left(y_{i} \mid \theta\right) \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right)\right]}{\partial \theta} d y_{i} \tag{10}
\end{equation*}
$$

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 $\frac{\partial f(y)}{\partial y}$

$$
\begin{equation*}
\frac{\partial f(y)}{\partial y}=\frac{\partial \ln [f(y)]}{\partial y} f(y) \tag{11}
\end{equation*}
$$

Use that little tidbit in the left hand side of 10, 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 $\ln \left[f\left(y_{i} \mid \theta\right)\right]$.

## 5.2 $\operatorname{Var}[U(\theta)]=-E[H]$

This is another result that is frequently asserted and I had never bothered to find out why until recently. This depends on the result in section 5.1.2.

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. That is

$$
\begin{equation*}
\int f\left(y_{i} \mid \theta\right) \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right)\right]}{\partial \theta} d y_{i}=0 \tag{12}
\end{equation*}
$$

Differentiating under the integral (Leibnitz rule),

$$
\int \frac{\partial f\left(y_{i} \mid \theta\right)}{\partial \theta} \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta} d y+\int f\left(y_{i} \mid \theta\right) \frac{\partial^{2} \ln ([f(y \mid \theta)]}{\partial \theta \partial \theta^{\prime}} d y=0
$$

which one can easily see is:

$$
\begin{equation*}
-\int f\left(y_{i} \mid \theta\right)\left[\frac{\partial^{2} \ln ([f(y \mid \theta)]}{\partial \theta \partial \theta^{\prime}}\right] d y=\int \frac{\partial f\left(y_{i} \mid \theta\right)}{\partial \theta} \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta} d y \tag{13}
\end{equation*}
$$

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 11. If you use that same trick:

$$
\int \frac{\partial f\left(y_{i} \mid \theta\right)}{\partial \theta} \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta} d y=\int f\left(y_{i} \mid \theta\right) \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta} \frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta} d y
$$

$$
\begin{equation*}
=\int f\left(y_{i} \mid \theta\right)\left(\frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta}\right)^{2} d y \tag{14}
\end{equation*}
$$

Claim: The last term is $\operatorname{Var}[U(\theta]$.
How do I know that? Recall the definition of the variance:

$$
\operatorname{Var}\left[U(\theta]=\int f\left(y_{i} \mid \theta\right)\left(U(\theta)-E[U(\theta)]^{2}\right) d y\right.
$$

As shown in the previous section, $E[U(\theta)]=0$. So:

$$
\operatorname{Var}\left[U(\theta]=\int f\left(y_{i} \mid \theta\right) U(\theta)^{2} d y\right.
$$

which (remembering the definition of $U(\theta)$ ) is just

$$
\operatorname{Var}[U(\theta)]=\int f\left(y_{i} \mid \theta\right)\left(\frac{\partial \ln \left[f\left(y_{i} \mid \theta\right]\right.}{\partial \theta}\right)^{2} d y
$$

## 6 Maximum Likelihood and the Normal distribution

### 6.1 Definition

Recall the Normal Distribution:

$$
\operatorname{prob}\left(y_{i} \mid \mu, \sigma^{2}\right)=\frac{1}{\sqrt{2 \pi \sigma^{2}}} e^{-\frac{1}{2 \sigma^{2}}\left(y_{i}-\mu_{i}\right)^{2}}
$$

That can be rearranged as

$$
\begin{equation*}
\operatorname{prob}\left(y_{i} \mid \mu, \sigma^{2}\right)=\exp \left[-\frac{1}{2} \ln (2 \pi)-\frac{1}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}}\left(y_{i}-\mu\right)^{2}\right] \tag{15}
\end{equation*}
$$

It is pitifully easy to find the log likelihood! Because:

$$
\ln [\exp [\text { anything }]]=\text { anything }
$$

### 6.2 The best estimate of the parameter $\mu$ is the sample mean!

Because of result 15, 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 simple that is:

$$
\begin{gather*}
\ln L\left(\mu, \sigma^{2}\right)=\sum_{i=1}^{N}\left[-\frac{1}{2} \ln (2 \pi)-\frac{1}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}}\left(y_{i}-\mu\right)^{2}\right]  \tag{16}\\
=-\frac{1}{2} \sum_{i=1}^{N} \ln (2 \pi)-\frac{1}{2} \sum_{i=1}^{N} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}} \sum_{i=1}^{N}\left(y_{i}-\mu\right)^{2} \\
=-\frac{1}{2} \cdot N \cdot \ln (2 \pi)-\frac{1}{2} \cdot N \cdot \ln \left(\sigma^{2}\right)-\frac{1}{\sigma^{2}} \sum_{i=1}^{N}\left(y_{i}-\mu\right)^{2} \\
=-\frac{N}{2} \cdot \ln (2 \pi)-\frac{N}{2} \cdot \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}} \sum_{i=1}^{N}\left(y_{i}-\mu\right)^{2}
\end{gather*}
$$

You want to maximize that by adjusting the values of $\mu$ and $\sigma^{2}$. Ignore the first part (that does not at all depend on either $\mu$ nor $\sigma^{2}$ ). 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"):

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

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

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

Which is -1 times the sum of squared deviations about $\hat{\mu}$. The best estimate, $\hat{\mu}$, is found by solving the first order condition

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

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 $\mu$ is the mean of the observations.

### 6.3 Regression with a Normal Variable

Suppose $y_{i} \sim N\left(\mu_{i}, \sigma^{2}\right)$. If $X_{i}$ is a the i'th row of a data set, suppose further that,

$$
\mu_{i}=X_{i} \beta
$$

Then one can think of $y_{i}$ as if it followed the $N\left(X_{i} \beta, \sigma^{2}\right)$ distribution. That means we need to estimate $\beta$, rather than $\mu$. Compare against equation 15

$$
\begin{align*}
\operatorname{prob}\left(y_{i} \mid \beta, \sigma^{2}\right) & =\exp \left[-\frac{1}{2} \ln (2 \pi)-\frac{1}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}}\left(y_{i}-X_{i} \beta\right)^{2}\right]  \tag{17}\\
\ln L\left(\beta \mu, \sigma^{2}\right) & =\sum_{i=1}^{N}\left[-\frac{1}{2} \ln (2 \pi)-\frac{1}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}}\left(y_{i}-X_{i} \beta\right)^{2}\right] \tag{18}
\end{align*}
$$

which implies

$$
\ln L\left(\beta, \sigma^{2}\right)=-\frac{N}{2} \ln (2 \pi)-\frac{N}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}} \sum_{i=1}^{N}\left(y_{i}-X_{i} \beta\right)^{2}
$$

and if you prefer, you can replace $\sum\left(y_{i}-X_{i} \beta\right)^{2}$ with vectors $(y-X \beta)^{\prime}(y-X \beta)$ :

$$
\begin{equation*}
\ln L\left(\beta, \sigma^{2}\right)=-\frac{N}{2} \ln (2 \pi)-\frac{N}{2} \ln \left(\sigma^{2}\right)-\frac{1}{2 \sigma^{2}}(y-X \beta)^{\prime}(y-X \beta) \tag{19}
\end{equation*}
$$

The score function is the vector of derivatives

$$
U(\beta)=\left[\begin{array}{c}
\frac{\partial \ln L}{\partial \beta_{1}} \\
\frac{\partial \ln L}{\partial \beta_{2}} \\
\vdots \\
\frac{\partial \ln L}{\partial \beta_{p}}
\end{array}\right]
$$

Why doesn't the score function include $\sigma^{2}$ ? 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 $\sigma^{2}$ can be calculated.

The first order condition is

$$
U(\beta)=0
$$

### 6.4 MLE Equals OLS

Give a casual glance at the objective function in 19 and you can tell that the first two terms don't matter because they don't depend on $\beta$. As a result, maximizing 19 is the same as maximizing

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

Which is the same as minimizing

$$
(y-X \beta)^{\prime}(y-X \beta)
$$

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

### 6.5 Solving the system

The Normally distributed variable is one for which we have an explicit solution. There is no need for approximation approaches described earlier.

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 $\hat{\beta}$ :

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

Since $y^{\prime} X \beta$ is a scalar (a $1 x 1$ matrix), it is equal to its transpose, $y^{\prime} X \beta=y^{\prime} X \beta$, so this reduces to: :

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

The first order condition is

$$
\begin{gathered}
U(\hat{\beta})=\frac{\partial \ln L}{\partial \beta}=-\frac{1}{2 \sigma^{2}} \frac{\partial}{\partial \beta}\left(-2 \hat{\beta}^{\prime} X^{\prime} y+\hat{\beta}^{\prime} X^{\prime} X \hat{\beta}\right)=0 \\
\left(-2 X^{\prime} y+2 X^{\prime} X \hat{\beta}\right)=0
\end{gathered}
$$

$$
\begin{aligned}
2 X^{\prime} X \hat{\beta} & =2 X^{\prime} y \\
X^{\prime} X \hat{\beta} & =X^{\prime} y
\end{aligned}
$$

If the inverse of $\left(X^{\prime} X\right)$ exists, the solution is:

$$
\hat{\beta}=\left(X^{\prime} X\right)^{-1} X^{\prime} y
$$

And the MLE for $\sigma^{2}$ is found by differentiating 19 with respect to $\sigma^{2}$.

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

