Maximum Likelihood Estimation (MLE)

1 Specifying a Model

Typically, we are interested in estimating parametric models of the form

$$y_i \sim f(\theta, y_i) \tag{1}$$

where θ is a vector of parameters and f is some specific functional form (probability density or mass function).[1] Note that this setup is quite general since the specific functional form, f, provides an almost unlimited choice of specific models. For example, we might use one of the following distributions:

- **Poisson Distribution**

$$y_i \sim f(\lambda, y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \tag{2}$$

  As you can see, we have only one parameter to estimate: λ. In terms of Eq. 1, $\theta = \lambda$.

- **Binomial Distribution**

$$y_i \sim f(\pi, y_i) = \frac{N!}{y_i!(N-y_i)!}\pi^{y_i}(1-\pi)^{N-y_i} \tag{3}$$

  or

$$y_i \sim f(\pi, y_i) = \binom{N}{y_i}\pi^{y_i}(1-\pi)^{N-y_i} \tag{4}$$

  Again, we have only one parameter to estimate: π. In terms of Eq. 1, $\theta = \pi$.

- **Normal Distribution**

$$y_i \sim f_N(\theta, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \tag{5}$$

  where $\theta = \mu, \sigma^2$ and $\mu = g((\beta, x_i)$. As you can see, we have two parameters to estimate: β and σ^2. In terms of Eq. 1, $\theta = \beta, \sigma^2$.

Obviously the choice of distribution will depend on your theory.

[1] Probability mass functions apply to discrete random variables, whereas probability density functions apply to continuous random variables.

2 Intuition

The following example provides some intuition about maximum likelihood estimation. Suppose our dependent variable follows a normal distribution:

$$y_i \sim N(\mu, \sigma^2) \tag{6}$$

Thus, we have:

$$E[y] = \mu \tag{7}$$

$$\text{var}(y) = \sigma^2 \tag{8}$$

In general, we will have some observations on Y and we want to estimate μ and σ^2 from those data. The idea, as we will see, of maximum likelihood is to find the estimate of the parameter(s) that maximizes the probability of observing the data that we have. Suppose that we have the following five observations on Y:

$$Y = \{54, 53, 49, 61, 58\} \tag{9}$$

Intuitively, we might wonder about the odds of getting these five data points if they were from a normal distribution with $\mu = 100$. The answer here is that it is not very likely – all of the data points are a long way from 100. But what are the odds of getting the five data points from a normal distribution with $\mu = 55$. Now this seems much more reasonable. Maximum likelihood estimation is just a *systematic* way of searching for the parameter values of our chosen distribution that maximize the probability of observing the data that we observe.

Before, we look at the process of maximum likelihood estimation in detail, we need to go over some preliminaries first.

3 Preliminaries

3.1 Fundamental Problem of Inference

Given that $y_i \sim f(\theta, y_i)$, we would like to make inferences about the value of θ.

Note that the problem that we face is the opposite of the typical probability problem. Think back to when you studied probability theory – you wanted to know something about the distribution of y given the parameters of your model (θ). What is the probability of drawing an Ace of Spades from a fair deck? What is the probability of getting 6 heads when tossing a fair coin 8 times? As King (1998, 14) puts it, you want to know $p(y|Model)$ or $p(Data|Model)$. In our terminology, you want to know $f(y|\theta)$

In our case, though, we have the data but want to learn about the model, specifically the model's parameters. In other words, we want to know the distribution of the unknown parameter(s) conditional on the observed data i.e. $p(Model|Data)$ or $f(\theta|y)$. This is known as the 'inverse probability problem'.

3.2 Bayes' Theorem

Recall the following identities from any probability class.

$$p(\theta, y) = p(\theta)p(y|\theta) \tag{10}$$

and

$$p(\theta, y) = p(y)p(\theta|y) \tag{11}$$

We want to know the last term in Eq. 11 i.e. $p(\theta|y)$. It should be obvious from Eq. 10 and 11 that the conditional density of $p(\theta|y)$ is

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta, y)}{p(y)} \\ &= \frac{p(\theta)p(y|\theta)}{p(y)} \end{aligned} \tag{12}$$

Note that the denominator, $p(y)$, is just a function of the data. Since it only makes sense to compare these conditional densities for the same data, we can essentially ignore the denominator. This means that we can rewrite Eq. 12 in its more familiar form

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{13}$$

where $\frac{1}{p(y)}$ is called the constant of proportionality, $p(\theta)$ is the prior density of θ, $p(y|\theta)$ is the likelihood, and $p(\theta|y)$ is the posterior density of θ. Thus, Eq. 13 supplies the Bayesian mantra that 'the posterior is proportional to the prior times the likelihood'.[2] Put differently, the likelihood is the sample information that transforms a 'prior' into a 'posterior' density of θ.

It is important to note that the prior, $p(\theta)$, is fixed before our observations and so can be treated as invariant to our problem. This means that we can rewrite Eq. 13 as

$$p(\theta|y) = k(y)p(y|\theta) \tag{14}$$

where $k(y) = \frac{p(\theta)}{p(y)}$ and is an unknown function of the data. Since $k(y)$ is not a function of θ, it is treated as an unknown positive constant. Put differently, for a given set of observed data, $k(y)$ remains the same over all possible hypothetical values of θ.

4 Likelihood

Without knowing the prior or making assumptions about the prior, we cannot calculate the inverse probability in Eq. 14. However, R. A. Fisher (1912) got around this by introducing the notion of likelihood and the Likelihood Axiom. He defined

$$\begin{aligned} \mathcal{L}(\theta|y) &= k(y)p(y|\theta) \\ &\propto p(y|\theta) \end{aligned} \tag{15}$$

[2] In Bayesian estimation, you make assumptions about the prior (i.e. about $p(\theta)$) and use the likelihood to update according to Bayes' rule. One of the problematic features of Bayesian estimation, though, is that there can be disputes over the appropriate prior.

The likelihood is proportional to the probability of observing the data, treating the parameters of the distribution as variables and the data as fixed. The advantage of likelihood is that it can be calculated from a traditional probability, $p(y|\theta)$, whereas an inverse probability cannot be calculated in any way. **Note that we can only compare likelihoods for the same set of data and the same prior.**

The best estimator, $\hat{\theta}$, is whatever value of $\hat{\theta}$ that maximizes

$$\mathcal{L}(\theta|y) = p(y|\theta) \tag{16}$$

In effect, we are looking for the $\hat{\theta}$ that maximizes the likelihood of observing our sample [recall the intuitive example from earlier]. Because of the proportional relationship, the $\hat{\theta}$ that maximizes $\mathcal{L}(\theta|y)$ will also maximize $p(\theta|y)$ i.e. the probability of observing the data. This is what we wanted from the beginning.

4.1 Likelihood of the Entire Sample

IF the y_i are all independent (or conditionally independent given x_i), then the likelihood of the whole sample is the product of the individual likelihoods over all the observations.

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_1 \times \mathcal{L}_2 \times \mathcal{L}_3 \times \ldots \times \mathcal{L}_N \\
&= \prod_{i=1}^{N} \mathcal{L}_i \\
&= p(y_1|\hat{\theta}) \times p(y_2|\hat{\theta}) \times p(y_3|\hat{\theta}) \times \ldots \times p(y_N|\hat{\theta}) \\
&= \prod_{i=1}^{N} p(y_i|\hat{\theta})
\end{aligned} \tag{17}$$

Instead of writing down the likelihood function, we often write down the log-likelihood function

$$\ln\mathcal{L} = \sum_{i=1}^{N} \ln\, p(y_i|\hat{\theta}) \tag{18}$$

Since all likelihoods are negative, the likelihood and its log have their maxima at the same place. It tends to be much simpler to work with the log-likelihood since we get to sum things up.

4.2 Maximum Likelihood Estimation

Once we have the likelihood (or more normally the log-likelihood) function, we need to find $\hat{\theta}^{ML}$. There are generally three options for this.

1. **Analytic:** Differentiate the likelihood function with respect to the parameter vector and set the resulting gradient vector to zero. Solve the system of equations to find extrema. Take the second derivative to make sure that you have a maximum rather than a minimum. This method only works if there is an analytical solution.

2. **Grid Search:** If you know $\hat{\theta}$ lies in a subspace of $\Re$, do an exhaustive search over that region for the $\hat{\theta}$ that produces the largest likelihood. In other words, try each possible value of $\hat{\theta}$ and see which produces the largest likelihood. The grid search method is a good way of showing that you can find the maximum of the likelihood function by repeated approximation and iteration. However, it is not practical in most cases and becomes much more difficult when the number of parameters rises beyond one or two.

3. **Numerical:** This is the most common. Essentially, you give the computer a set of starting values, $\hat{\theta}^0$, and let a hill climbing algorithm (Newton-Raphson, BHHH, DFP, etc.) find the maximum.[3]

4.3 Poisson Example

Let's work through an example analytically for the Poisson distribution:[4] The Poisson distribution, which is often used as the underlying distribution for a count model, can be written as:

$$y_i \sim f(\lambda, y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \tag{19}$$

For the Poisson distribution, the likelihood function would be

$$\begin{aligned} L &= \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \times \frac{e^{-\lambda}\lambda^{y_2}}{y_2!} \times \frac{e^{-\lambda}\lambda^{y_3}}{y_3!} \times \ldots \times \frac{e^{-\lambda}\lambda^{y_N}}{y_N!} \\ &= \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\ &= \frac{e^{-N\lambda}\lambda^{\sum_{i=1}^{N} y_i}}{\prod_{i=1}^{N} y_i!} \end{aligned} \tag{20}$$

and the log-likelihood function would be

$$\begin{aligned} \ln\mathcal{L} &= \sum_{i=1}^{N} \ln\left(\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) \\ &= \sum_{i=1}^{N} \ln\left(\frac{e^{-N\lambda}\lambda^{\sum_{i=1}^{N} y_i}}{\prod_{i=1}^{N} y_i!}\right) \\ &= -N\lambda + \sum_{i=1}^{N} y_i \ln(\lambda) - \sum_{i=1}^{N} \ln(y_i!) \end{aligned} \tag{21}$$

[3] These numerical methods are described in more detail in a later section of these notes.

[4] Some properties that you might find useful:

1. $e^a \times e^b = e^{a+b}$
2. $\ln(a \times b) = \ln a + \ln b$
3. $\ln r^s = s \ln r$
4. $\ln\left(\frac{r}{s}\right) = \ln r - \ln s$
5. $\ln e = 1$

Now, let's find the analytical solution to our Poisson example. To do this we start by taking the derivative of Eq. 21 with respect to λ.[5]

$$\frac{\partial \ln \mathcal{L}}{\partial \lambda} = -N + \frac{\sum_{i=1}^{N} y_i}{\lambda} \tag{22}$$

We then set this equal to zero and solve for λ.

$$\begin{aligned} -N + \frac{\sum_{i=1}^{N} y_i}{\lambda} &= 0 \\ \hat{\lambda} &= \frac{\sum_{i=1}^{N} y_i}{N} \end{aligned} \tag{23}$$

Finally, we would check that this is a maximum by taking the derivative of Eq. 22 with respect to λ again. I leave this for you to check.

4.4 Bernoulli Example

Suppose we have a Bernoulli model in which each observation has a constant and equal chance of success, π. A Bernoulli variable takes on one of two values, conventionally 1 or 0, which indicate a 'success' or 'failure'. The probability distribution for this variable is:

$$\begin{aligned} f(1) &= \pi \\ f(0) &= 1 - \pi \end{aligned} \tag{24}$$

If S stands for the number of successes and F for the number of failures, then the likelihood is:

$$\mathcal{L} = \pi^S (1-\pi)^F \tag{25}$$

and the log-likelihood would be:

$$\ln \mathcal{L} = S \ln(\pi) + F \ln(1-\pi) \tag{26}$$

We then take the derivative of this with respect to π.

$$\frac{\partial \ln \mathcal{L}}{\partial \pi} = \frac{S}{\pi} - \frac{F}{1-\pi} \tag{27}$$

We now set this equal to zero and solve for π:

$$\begin{aligned} \frac{S}{\pi} - \frac{F}{1-\pi} &= 0 \\ S(1-\pi) - F\pi &= 0 \\ \pi &= \frac{S}{S+F} \\ &= \frac{S}{N} \end{aligned} \tag{28}$$

Now that we have our $\hat{\theta}^{ML}$ estimates, we need to calculate standard errors. This will require us to learn about the Information Matrix.

[5] Recall that $(\ln x)' = \frac{1}{X}$ and $\left(e^X\right)' = e^x$.

5 Information Matrix and Standard Errors

The variance of an ML estimator, $\hat{\theta}^{ML}$, is calculated by the inverse of the Information matrix:

$$\text{var}(\theta) = [I(\theta)]^{-1} \tag{29}$$

What is this? This will take a few steps. First, the Information matrix is the negative of the expected value of the Hessian matrix:

$$[I(\theta)] = -E[H(\theta)] \tag{30}$$

So, now what is the Hessian? The Hessian is the matrix of second derivatives of the likelihood with respect to the parameters:

$$H(\theta) = \frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta \partial \theta'} \tag{31}$$

Thus, the variance-covariance matrix of $\hat{\theta}^{ML}$ is:

$$\begin{aligned} \text{var}(\theta) &= [I(\theta)]^{-1} \\ &= (-E[H(\theta)])^{-1} \\ &= \left(-E\left[\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta \partial \theta'} \right] \right)^{-1} \end{aligned} \tag{32}$$

As we'll see in a moment, the standard errors of the estimator, $\hat{\theta}$, are just the square roots of the diagonal terms in the variance-covariance matrix.

> **Cramer-Rao Theorem:** This shows that (given certain regularity conditions concerning the distribution), the variance of any unbiased estimator of a parameter θ must be at least as large as
>
> $$\text{var}(\theta) \geq (-E[H(\theta)])^{-1} \tag{33}$$

This means that any unbiased estimator that achieves this lower bound is efficient and no better unbiased estimator is possible. Now look back at the variance-covariance matrix in Eq. 32. You will see that the inverse of the information matrix is exactly the same as the Cramer-Rao lower bound. This means that MLE is efficient.

5.1 Why? Easy and Hard

Why is this how we calculate the standard errors? The easy way to think about this is to recognize that the curvature of the likelihood function tells us how certain we are about our estimate of our parameters. The more curved the likelihood function, the more certainty we have that we have estimated the right parameter. The second derivative of the likelihood function is a measure of the likelihood function's curvature - this is why it provides our estimate of the uncertainty with which we have estimated our parameters.

The hard, but correct, way to think about this is the following. If we took a linearization of the derivative of the likelihood at the maximum likelihood point, $\hat{\theta}$, around the true value, θ, we would have:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \theta}|\hat{\theta} \\ &= \frac{\partial \mathcal{L}}{\partial \theta}|\theta + \frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}(\hat{\theta} - \theta) \\ \hat{\theta} - \theta &= -\left[\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1} \frac{\partial \mathcal{L}}{\partial \theta} \end{aligned} \tag{34}$$

The variance of $\hat{\theta}$ is just the outer product of the above.

$$\begin{aligned} \text{var}(\hat{\theta}) &= E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \\ &= E\left[\left[\frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1} \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial \mathcal{L}'}{\partial \theta} \left[\frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1}\right] \end{aligned} \tag{35}$$

The score is the gradient of the likelihood $\left(\frac{\partial \mathcal{L}}{\partial \theta}\right)$. If the model is correctly specified, then the expectation of the outer product of the scores (the middle bit) is equal to the information matrix. As a result, we can rewrite Eq. 35 as

$$\begin{aligned} \text{var}(\hat{\theta}) &= E\left[\left[\frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1} \frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'} \left[\frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1}\right] \\ &= \left(-E\left[\frac{-\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]\right)^{-1} \end{aligned} \tag{36}$$

As you can see, this is the negative of the inverse of the information matrix. We can read off the standard errors of $\hat{\theta}$ from the square roots of the diagonal elements of this matrix. Note that these are only correct asymptotically and are hard to calculate in finite samples [we'll return to this a little later in the notes].

5.2 Robust Standard Errors

If the model is not well-specified but the mean function is correctly specified and the variance function is not horribly specified, then maximum likelihood is asymptotically normal with the following variance-covariance matrix

$$\text{var}(\hat{\theta}) = I^{-1} \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial \mathcal{L}'}{\partial \theta} I^{-1} \tag{37}$$

This is the variance-covariance matrix that provides what we call robust variances from Eq. 35. This is the maximum likelihood analogue of White's consistent standard errors.

6 Properties of ML Estimators

If a minimum variance unbiased estimator exists, then the MLE estimator will be it.

6.1 Large Sample Properties

- Consistent: plim $\hat{\theta}^{ML} = \theta$
- Asymptotically normal - $\hat{\theta}^{ML} \overset{a}{\sim} N[\theta, \{I(\theta)\}^{-1}]$
- Variance-Covariance is the Rao-Cramer lower bound (if the model is well-specified) - efficient
- Invariance: If θ^{ML} is the ML estimator of θ, then $\gamma^{ML} = g(\theta^{ML})$ is the maximum likelihood estimator of $\gamma = g(\theta)$. This means that rather than estimating a parameter θ, we can instead estimate some function of it, $g(\theta)$. We can then recover an estimate of θ (that is, $\hat{\theta}$) from $g(\theta)$.

Thus, ML is best asymptotically normal.

6.2 Small Sample Properties

We don't know much about MLE's small sample properties. In general, though, if there is a good small sample estimator, it will look like maximum likelihood. Note that maximum likelihood may not be unbiased in small samples.

6.3 An Aside on Bootstrapping

As noted above, all standard errors in MLE are asymptotic. You probably should not use MLE when your sample is small (maybe less than 100). However, if your sample is small and you still use MLE, you should at least try to deal with the fact that your standard errors are not estimated correctly. One way to do this is 'bootstrapping'. A bootstrap provides a way to perform a statistical inference by resampling from the sample.[6]

The idea is relatively simple. Suppose we wanted to obtain a bootstrap estimate of the standard error of an estimator $\hat{\theta}$. Suppose we had 400 random samples from the population. From these, we could get 400 different estimates of $\hat{\theta}$ and let the standard error of $\hat{\theta}$ be the standard deviation of these 400 estimates.

The problem is that we normally have only one sample from the population available. The bootstrap procedure essentially generates multiple samples by resampling from the current sample. In effect, our current sample is viewed as the population and we obtain multiple samples from this population by resampling (obviously with replacement since we would only have one sample if we didn't do this). Given the 400 bootstrap resamples, we can obtain 400 estimates and then estimate the standard error of $\hat{\theta}$ by the standard deviation of these 400 estimates.

[6]The most common use of the bootstrap is to provide standard error estimates when analytical expressions are complicated. You can then use these standard errors to construct confidence intervals and test statistics.

To be more specific, let $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ denote the estimates, where $B = 400$ in this case. Then the bootstrap estimate of the variance of $\hat{\theta}$ is:

$$\widehat{\text{var}}_{\text{boot}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \overline{\hat{\theta}^*})^2 \tag{38}$$

where $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b$ is the average of the B bootstrap estimates. The square root of $\widehat{\text{var}}_{\text{boot}}(\hat{\theta})$ is called the bootstrap standard error of $\hat{\theta}$.[7] The bootstrap standard errors may well give a more accurate assessment of the variance of the parameter estimates.

The STATA command to obtain bootstrap standard errors is:

```
probit Y X, vce(boot, reps(400) seed(10101))
```

where reps(400) says to use 400 bootstrap resamples and seed(10101) sets the seed to enable replication. You will get the same output as before except now the standard errors and test statistics are based on the bootstrap standard errors.

If you want to obtain clustered bootstraps, the command is:

```
probit Y X, vce(boot, cluster(Z) reps(400) seed(10101))
```

It is possible to obtain different bootstrap confidence intervals. The STATA commands shown above will give you what is called a "normal-based" (N) 95% confidence intervals for θ that equal

$$[\hat{\theta} - 1.96 \times \text{se}_{\text{boot}}(\hat{\theta}), \hat{\theta} + 1.96 \times \text{se}_{\text{boot}}(\hat{\theta})] \tag{39}$$

and is a standard Wald asymptotic confidence interval. However, you can obtain alternative confidence intervals.

The percentile (P) method uses the relevant percentiles of the empirical distribution of the B bootstrap estimates $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. Thus, a percentile 95% confidence interval is $(\hat{\theta}^*_{0.025}, \hat{\theta}^*_{0.975})$. This confidence interval is asymmetric around $\hat{\theta}$ and is generally considered a better approximation than the normal-based confidence interval.

Two alternative confidence intervals are the bias-corrected (BC) confidence interval, which is a modification of the percentile method that incorporates a bootstrap estimate of the finite-sample bias in $\hat{\theta}$, and the BC accelerated (BCa) confidence interval, which is an adjustment of the BC method that allows the asymptotic variance of $\hat{\theta}$ to vary with θ.

In general, you are probably fine using the N or P method for calculating confidence intervals. To see all of the confidence intervals in STATA, type:

```
quietly probit Y X, vce(boot, reps(400) seed(10101))

estat bootstrap, all
```

[7] It is important to note that the bootstrap procedure that has just been outlined assumes independence of observations or of clusters of observations. This allows for dependence via clustering, so long as observations are combined into clusters that are independent and the bootstrap is over the clusters.

7 OLS Example

Let's work our way through a more complicated example for the normal distribution.

7.1 Finding the Log-Likelihood

As we saw earlier, the normal distribution is written as:

$$y_i \sim N(\mu_i, \sigma^2 I) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \tag{40}$$

where $\mu_i = x_i\beta$. This can be rewritten as:

$$y_i \sim N(\mu_i, \sigma^2 I) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i\beta)^2}{2\sigma^2}} \tag{41}$$

Thus, the likelihood for a single observation is:

$$\mathcal{L}(y_i|x, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i\beta)^2}{2\sigma^2}} \tag{42}$$

It follows that the likelihood for the whole sample is:

$$\mathcal{L}(y|x, \beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i\beta)^2}{2\sigma^2}} \tag{43}$$

In matrix form the likelihood is:

$$\frac{1}{(\sqrt{2\pi\sigma^2})^N} \; e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)} \tag{44}$$

Now taking logs, the log-likelihood is:

$$\begin{aligned} \ln\mathcal{L}(y|x, \beta, \sigma^2) &= \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i\beta)^2}{2\sigma^2}}\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i - x_i\beta)^2}{\sigma^2}\right] \end{aligned} \tag{45}$$

In matrix form, the log likelihood is:

$$\ln\mathcal{L} = \frac{-N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \tag{46}$$

7.2 Finding the ML Estimator

In this particular case, we can find a closed form solution for the parameters (β, σ^2). Let's start by using the sample log-likelihood in matrix form:

$$\ln\mathcal{L}(y|X,\beta,\sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2}\left[\frac{(y-X\beta)'(y-X\beta)}{\sigma^2}\right] \tag{47}$$

where y is an $N \times 1$ vector and X is an $N \times k$ matrix. By expanding the numerator of the last term and moving the scalar σ^2 outside, we have:

$$\ln\mathcal{L}(y|X,\beta,\sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}[y'y - 2y'X\beta + \beta'X'X\beta] \tag{48}$$

To find the gradient vector, we need to take the derivative of $\ln\mathcal{L}$ with respect to β and σ^2. Let's start by taking the derivative of Eq. 48 with respect to β.

$$\begin{aligned}
\frac{\partial\ln\mathcal{L}}{\partial\beta} &= -\frac{1}{2\sigma^2}\left[\frac{\partial[y'y - 2y'X\beta + \beta'X'X\beta]}{\partial\beta}\right] \\
&= -\frac{1}{2\sigma^2}[-2X'y + 2X'X\beta] \\
&= \frac{1}{2\sigma^2}[2X'y - 2X'X\beta] \\
&= \frac{1}{\sigma^2}[X'y - X'X\beta]
\end{aligned} \tag{49}$$

We now set this equal to zero:

$$\begin{aligned}
\frac{1}{\sigma^2}[X'y - X'X\beta] &= 0 \\
X'X\beta &= X'y \\
\hat{\beta} &= (X'X)^{-1}X'y
\end{aligned} \tag{50}$$

This is the familiar formula for an OLS coefficient vector. As you can see, OLS and ML give the same estimator for the coefficients.

Second, we take the derivative of Eq. 48 with respect to σ^2.

$$\frac{\partial\ln\mathcal{L}}{\partial\sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(y-X\beta)'(y-X\beta)] \tag{51}$$

We now set this equal to zero:

$$\begin{aligned}
-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(y-X\beta)'(y-X\beta)] &= 0 \\
\frac{1}{2\sigma^4}[(y-X\beta)'(y-X\beta)] &= \frac{N}{2\sigma^2} \\
\frac{1}{\sigma^2}[(y-X\beta)'(y-X\beta)] &= N
\end{aligned} \tag{52}$$

Since we have already solved for $\hat{\beta}$, we can solve for σ^2 by replacing β with its estimate:

$$\begin{aligned} \frac{1}{\sigma^2}[(y - X\hat{\beta})'(y - X\hat{\beta})] &= N \\ \frac{1}{\sigma^2}[(y - \hat{y})'(y - \hat{y})] &= N \\ \frac{1}{\sigma^2}[e'e] &= N \\ \hat{\sigma^2} = \frac{e'e}{N} & \end{aligned} \tag{53}$$

Recall that the OLS estimate, which is unbiased, of $\hat{\sigma}^2 = \frac{e'e}{N-K}$. Thus, the OLS and ML estimator of σ^2 are different. Specifically, the MLE estimate is biased downwards in small samples. However, it is relatively easy to see that the OLS and ML estimators are asymptotically equivalent i.e. they converge as N goes to infinity.

Gradient Vector: This is the vector that contains the first derivative of the log-likelihood function with respect to our parameters. Thus, the gradient vector from the OLS example is the following

$$G = \frac{\partial \ln \mathcal{L}}{\partial \theta} = \begin{bmatrix} \frac{\partial \ln \mathcal{L}}{\partial \beta} \\ \frac{\partial \ln \mathcal{L}}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{X'(y - X\beta)}{\sigma^2} \\ -\frac{N}{2\sigma^2} + \frac{(y-X\beta)'(y-X\beta)}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ \sigma^2 \end{bmatrix} \tag{54}$$

7.3 Finding the Variance-Covariance Matrix

Recall that variance-covariance matrix is:

$$[I(\theta)]^{-1} = (-E[H(\theta)])^{-1} \tag{55}$$

Thus, the first thing we do is find the Hessian, the matrix of second derivatives with respect to our parameters.

We start with the gradient vector:

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \begin{bmatrix} \frac{X'(y - X\beta)}{\sigma^2} \\ -\frac{N}{2\sigma^2} + \frac{(y-X\beta)'(y-X\beta)}{2\sigma^4} \end{bmatrix} \tag{56}$$

We now need to take the derivative of each element of the gradient vector with respect to β and with respect to σ^2. Let's start with β.

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} = \frac{\partial \left[\frac{X'(y - X\beta)}{\sigma^2}\right]}{\partial \beta} = \frac{\partial \left[\frac{X'y - X'X\beta)}{\sigma^2}\right]}{\partial \beta} = -\frac{X'X}{\sigma^2} \tag{57}$$

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \sigma^2} = \frac{\partial \left[\frac{X'(y - X\beta)}{\sigma^2} \right]}{\partial \sigma^2} = -\frac{X'(y - X\beta)}{\sigma^4} = -\frac{X'\epsilon}{\sigma^4} \tag{58}$$

since $\epsilon = y - X\beta$.

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}}{\partial \sigma^2 \partial \beta'} &= \frac{\partial \left[-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y'y - 2y'X\beta + \beta' X'X\beta) \right]}{\partial \beta} \\ &= \frac{-2y'X + 2\beta' X'X}{2\sigma^4} \\ &= \frac{-y'X + \beta' X'X}{\sigma^4} = -\frac{1}{\sigma^4}(y' - \beta' X')X = -\frac{\epsilon' X}{\sigma^4} \end{aligned} \tag{59}$$

since $\epsilon = y - X\beta$ and $\epsilon' = y' - \beta' X'$.

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}}{\partial \sigma^2 \partial \sigma^{2\prime}} &= \frac{\partial \left[-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(y - X\beta)'(y - X\beta)] \right]}{\partial \sigma^2} \\ &= \frac{N}{2\sigma^4} - \frac{(y - X\beta)'(y - X\beta)}{\sigma^6} \\ &= \frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6} \end{aligned} \tag{60}$$

Thus, the Hessian matrix from the OLS example is the following:

$$H = \frac{\partial \ln \mathcal{L}}{\partial \theta \partial \theta'} = \begin{bmatrix} -\frac{X'X}{\sigma^2} & -\frac{X'\epsilon}{\sigma^4} \\ -\frac{\epsilon' X}{\sigma^4} & \frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6} \end{bmatrix} \tag{61}$$

Now we need to take the expectation of the Hessian.

$$E[H] = E \begin{bmatrix} -\frac{X'X}{\sigma^2} & -\frac{X'\epsilon}{\sigma^4} \\ -\frac{\epsilon' X}{\sigma^4} & \frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6} \end{bmatrix} = \begin{bmatrix} E\left[-\frac{X'X}{\sigma^2} \right] & E\left[-\frac{X'\epsilon}{\sigma^4} \right] \\ E\left[-\frac{\epsilon' X}{\sigma^4} \right] & E\left[\frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6} \right] \end{bmatrix} \tag{62}$$

Since one of the Gauss-Markov assumptions state that $X'\epsilon = 0$ (disturbances are not correlated with the explanatory variables X), we know that the expectation of the elements on the off diagonal of the Hessian are zero i.e.

$$E\left[-\frac{X'\epsilon}{\sigma^4} \right] = 0 \tag{63}$$

and

$$E\left[-\frac{\epsilon'X}{\sigma^4}\right] = 0 \tag{64}$$

We know that:

$$E\left[-\frac{X'X}{\sigma^2}\right] = -\frac{X'X}{\sigma^2} \tag{65}$$

because the expectation of a constant is just the constant. Finally, we know that:

$$\begin{aligned} E\left[\frac{N}{2\sigma^4} - \frac{\epsilon'\epsilon}{\sigma^6}\right] &= E\left[\frac{N}{2\sigma^4}\right] - E\left[\frac{\epsilon'\epsilon}{\sigma^6}\right] \\ &= \frac{N}{2\sigma^4} - \frac{N\sigma^2}{\sigma^6} \\ &= -\frac{N}{2\sigma^4} \end{aligned} \tag{66}$$

because $\frac{N}{2\sigma^4}$ is a constant and because $E[\epsilon'\epsilon] = N\sigma^2$ by assumption. Thus, the expectation of the Hessian is:

$$E[H] = \begin{bmatrix} -\frac{X'X}{\sigma^2} & 0 \\ 0 & -\frac{N}{2\sigma^4} \end{bmatrix} \tag{67}$$

The Information matrix is the negative of the expectation of the Hessian. Thus,

$$I[\theta] = -E[H(\theta)] = -\begin{bmatrix} -\frac{X'X}{\sigma^2} & 0 \\ 0 & -\frac{N}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} \tag{68}$$

Finally, the variance-covariance matrix is the inverse of the Information matrix.[8] Thus, we have:

$$I[\theta]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} \tag{69}$$

Some of this should look very familiar to you. The stuff in the top left $\sigma^2(X'X)^{-1}$ is just the variance-covariance matrix of the OLS estimator, $\hat{\beta}$. Thus, the square root of the elements on the diagonal will give you the standard errors associated with your coefficients. The stuff in the bottom right is the variance of σ^2. This was not actually calculated during OLS - thus, MLE provides more information than OLS. Overall, the variance-covariance matrix of the ML estimator

[8] When the off diagonals of a matrix are 0, then we can find the inverse of a matrix by taking the inverse of the elements on the on-diagonal as we do here.

looks something like the following

$$I[\theta]^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) & \text{cov}(\hat{\beta}_1, \hat{\sigma}^2) \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & cov(\hat{\beta}_2, \hat{\beta}_k) & \text{cov}(\hat{\beta}_2, \hat{\sigma}^2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) & \text{cov}(\hat{\beta}_k, \hat{\sigma}^2) \\ \text{cov}(\hat{\sigma^2}, \hat{\beta}_1) & \text{cov}(\hat{\sigma^2}, \hat{\beta}_2) & \dots & \text{cov}(\hat{\sigma^2}, \hat{\beta}_k) & \text{var}(\hat{\sigma}^2) \end{bmatrix} \tag{70}$$

$$= \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) & 0 \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) & 0 \\ 0 & 0 & 0 & 0 & \text{var}(\hat{\sigma}^2) \end{bmatrix} \tag{71}$$

7.4 Recap

- We can write down the normal regression model, which we usually estimate via OLS, as an ML model. First, we write down the log-likelihood function. Second, we take derivatives with respect to the parameters. Third, we set the derivatives equal to zero. Fourth, we solve for the parameters.
- We get an estimator of the coefficient vector which is identical with that from OLS.
- The ML estimator of the variance is, however, different from the OLS estimator. The reason for the difference is that the OLS estimator of the variance is unbiased, while the ML estimator is biased but consistent. In large samples, as assumed by ML, the difference is insignificant.
- We can apply the formula for the Information matrix to get the variance-covariance matrix of the ML parameters. This gives us the familiar formula for the variance-covariance matrix of the parameters, $\sigma^2(X'X)^{-1}$, and a simple if unfamiliar expression for the variance of $\hat{\sigma}^2$.
- Since the parameter estimates are all MLEs, they are all asymptotically normally distributed.
- The square root of the diagonal elements of the inverse of the information matrix gives us estimates of the standard errors of the parameter estimates.
- We can construct simple z-scores to test the null hypothesis concerning any individual parameter, just as in OLS, but using the normal instead of the t-distribution.
- You have now seen a fully-worked, non-trivial application of ML to a model you are familiar with.

8 Inference - Testing Hypotheses

8.1 Individual Parameters

Say we want to test a hypothesis about an individual parameter. We know that ML parameter estimates are asymptotically normally distributed. If we wanted to test the hypothesis $H_0 : \theta_k = \theta^*$, then we would use:

$$z = \frac{\hat{\theta}_k - \theta^*}{\sqrt{V(\hat{\theta})_k}} \overset{a}{\sim} N(0,1) \tag{72}$$

whree θ^* in this case is zero. Under the assumptions justifying MLE, if H_0 is true, the z is distributed asymptotically normal with mean 0 and variance 1. In other words, the test is simply the z-score test you learned in elementary statistics. You'll notice when you do OLS regression that STATA reports t-statistics, but when you switch to MLE, STATA reports z-statistics. It makes little difference, though, whether you use a t-distribution or a normal distribution. As you can see, hypothesis testing about individual coefficients is easy.

8.2 Three Classic Tests

There are three classic tests associated with ML estimation - Likelihood Ratio Test, Wald Test, Lagrange Multiplier Test.

To see the logic behind these tests, consider MLE of parameter θ and a test of the hypothesis $H_0 : c(\theta) = 0$. Look at Figure 17.2 in Greene (2003, 485).

- **Likelihood Ratio Test:** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Thus, we base the test on the 'vertical' difference, $\lambda = \ln\mathcal{L}_U - \ln\mathcal{L}_R$, where $\mathcal{L}_U$ is the value of the likelihood function at the unconstrained value of θ and $\mathcal{L}_R$ is the value of the likelihood function at the restricted estimate. It has been shown that under the null hypothesis, -2λ is distributed χ^2 with degrees of freedom equal to the number of restrictions imposed. To do this test you will have to run two models and get the results.[9]

- **Wald Test:** If the restriction is valid, then $c(\hat{\theta}_{MLE})$ should be close to zero since MLE is consistent. Thus, the test is based on $c(\hat{\theta}_{MLE})$. We reject the hypothesis if this value is significantly different from zero. Essentially, you estimate just the unconstrained model and then test whether the hypothesized constraints are inconsistent with this model. The test basically uses two pieces of information. First, it measures the distance $\hat{\theta}_{MLE} - c(\hat{\theta}_{MLE})$. The larger this distance, the less likely it is that the constraint ($c(\hat{\theta}_{MLE}) = 0$) is true. Second, this distance is weighted by the curvature of the log-likelihood function, which is indicated by the second derivative. The larger the second derivative, the faster the curve is changing.

[9] Note that for the test to work, one model must be nested inside the other model i.e. the simpler model must be the bigger model with some constraints imposed.

If the second derivative is small, this indicates that the log-likelihood curve is relatively flat. Thus, the distance $\hat{\theta}_{MLE} - c(\hat{\theta}_{MLE})$ will be minor relative to the sampling variation. If the curve is steeper, the same distance may be significant.

- **Lagrange Multiplier Test:** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

The three tests are asymptotically equivalent, but may differ in small samples.

9 Numerical Maximization Methods

As I noted earlier, there are three different ways to employ maximum likelihood estimation: analytical, grid search, and numerical. In many cases, there is no analytical solution and grid searches are impractical. As a result, most statistical packages employ some kind of numerical maximization method. But how exactly do these methods work?[10]

9.1 Intuition

Recall that the goal of MLE is to find values of the parameters, say β, that maximize the (log)likelihood function. To do this, we could start with a guess of $\hat{\beta}$ and let's call this $\hat{\beta}_0$. We could then adjust this guess based on the value of the (log)likelihood that it gives us. For simplicity, let's call this adjustment A. Thus, our new guess would be:

$$\hat{\beta}_1 = \hat{\beta}_0 + A_0 \tag{73}$$

Or, more generally,

$$\hat{\beta}_k = \hat{\beta}_{k-1} + A_{k-1} \tag{74}$$

Remember that we want to move the vector $\hat{\beta}$ to the point at which the likelihood is highest. How can we do this? Well, we need to take account of the slope of the likelihood function at each guess. Intuitively, the way that we do this is by incorporating information from the gradient matrix (the matrix of first derivatives with respect to the $\hat{\beta}$s) that we saw earlier. If the gradient matrix is positive, then the likelihood is increasing in $\hat{\beta}$ and so we need to increase our guess of $\hat{\beta}$ some more. And if the gradient matrix is negative, then the likelihood is decreasing in $\hat{\beta}$ and so we need to decrease our guess of $\hat{\beta}$. By repeatedly doing this, we gradually climb to the top of the likelihood function. As you can imagine, as we get to the top, the gradient becomes closer and closer to zero. When the changes get sufficiently small from one iteration to the next, we simply stop, and use the estimates that got us here.

[10] The following notes draw heavily on Zorn's excellent notes and Long (1997).

9.2 More Detail

As we noted before, we start with a guess of $\hat{\beta}$, which we'll call $\hat{\beta}_0$. We then adjust this guess based on the slope of the likelihood function. Before, we left this adjustment vague. But the simplest way to update the parameter estimates is to specify $A_k = \frac{\partial \ln \mathcal{L}}{\partial \beta_k}$. This gives us:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\partial \ln \mathcal{L}}{\partial \beta_k} \tag{75}$$

In other words, we adjust the parameter by a factor equal to the first derivative of the log-likelihood function with respect to the parameters i.e. equal to the gradient matrix. As I noted before, if the gradient matrix is positive, then we increase $\hat{\beta}$ more next time and if the gradient matrix is negative, then we decrease $\hat{\beta}$ next time. This general approach is called the *method of steepest ascent*. Although this method is attractive, it does have one problem, which is that it doesn't consider how fast the slope is changing. As a result, we typically modify this approach.

The best way to think about this is that the matrix $\frac{\partial \ln \mathcal{L}}{\partial \beta_k}$ is a 'direction' matrix – it tells us which direction to go in to reach the maximum. But it doesn't tell us how far to increase or decrease $\hat{\beta}$ each time. Thus, we need to generalize our approach so that we change our estimates not only in terms of their direction but also by a factor, the 'step size', determining how far we should change them:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \lambda_{k-1} \Delta_{k-1} \tag{76}$$

As Eq. (76) indicates, our adjustment A now has two parts:

- Δ tells us the direction we want to take a step in
- λ tells us the amount by which we want to change our estimates

As you will have recognized, we need a way of determining how fast the slope of the likelihood is changing at that value of $\hat{\beta}$ to do this. We can get this from the second derivative of the log-likelihood function with respect to the $\hat{\beta}$. The second derivative tells us the rate at which the slope of the log-likelihood function is changing. If it is big, then the slope is increasing or decreasing quickly. We need to incorporate the information from the second derivative about how fast the slope is changing into our numerical maximization routine. We do this by saying that if the function is very steep (the second derivative is big), then we don't want to change the parameters very much from one iteration to the next. However, if the function is flat (the second derivative is small), then we can adjust the parameters more from one iteration to the next.

It turns out that we can actually do this in three different ways:

1. **The Hessian**

$$H = \frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} \tag{77}$$

As we saw earlier, this $k \times k$ matrix contains the second derivatives along the main diagonal and the cross-partials of the elements of $\hat{\beta}$ in the off-diagonals. Sometimes, though, figuring out the second derivatives can be hard and so two alternatives are often used.

2. **Information Matrix**

$$I = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'}\right] \tag{78}$$

In some cases, it can be easier to compute the information matrix than the Hessian matrix.

3. **Outer Product Approximation to the Information Matrix**

If the information matrix is also too difficult to calculate, we can use the outer product approximation of the information matrix:

$$\sum_{i=1}^{N} \frac{\partial \ln \mathcal{L}_i}{\partial \beta} \frac{\partial \ln \mathcal{L}_i'}{\partial \beta} \tag{79}$$

In other words, we sum over the 'squares' (outer products) of the first derivatives of the log-likelihoods. This has the advantage that we don't have to deal with the second derivatives at all.

Each of these options for incorporating the 'rate of change' has a maximization algorithm associated with it.

1. **Newton-Raphson uses the (inverse of the) Hessian**

$$\hat{\beta}_k = \hat{\beta}_{k-1} - \left[\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \hat{\beta}_{k-1} \partial \hat{\beta}_{k-1}'}\right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \hat{\beta}_{k-1}}\right] \tag{80}$$

As this illustrates, the new parameter estimates are equal to the old ones. However, they are adjusted in the direction of the first derivative (steepest ascent) and the amount of change is inversely related to the size of the second derivative.

2. **Method of Scoring uses the (inverse of the) information matrix**

$$\hat{\beta}_k = \hat{\beta}_{k-1} - \left[E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \hat{\beta}_{k-1} \partial \hat{\beta}_{k-1}'}\right)\right]^{-1} \frac{\partial \ln \mathcal{L}}{\partial \hat{\beta}_{k-1}} \tag{81}$$

3. **Berndt, Hall, Hall, and Hausman (BHHH) uses the inverse of the outer product approximation to the information matrix**

$$\hat{\beta}_k = \hat{\beta}_{k-1} - \left(\sum_{i=1}^{N} \frac{\partial \ln \mathcal{L}_i}{\partial \beta} \frac{\partial \ln \mathcal{L}_i'}{\partial \beta}\right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \hat{\beta}_{k-1}} \tag{82}$$

As Zorn points out, there are other algorithms such as the Davidson-Fletcher-Powell (DFP) algorithm, in which the step length is chosen in a way to make the updated matrix of parameter estimates positive definite.

Newton-Raphson works pretty well and quickly for simply functions with global maxima. Newton-Raphson is the default in STATA. As a result, STATA actually calculates the gradient and Hessian at each iteration; it can be slow but is quite reliable. You might want to use the scoring method or BHHH if you have a complicated likelihood function or the data are poor (i.e. collinearity). To find out more about the maximization methods in STATA, type HELP MAXIMIZE.

9.3 Potential Problem Solving

If you cannot get the model to converge or the answer you get appears wrong, then there are a number of things that you can look at:

1. Make sure that the model is well-specified and that the variables are constructed correctly.
2. Rescale the variables so that they are roughly on the same scale. It turns out that the larger the ratio between the largest standard deviation and the smallest standard deviation, the more problems you will have in numerical methods.
3. Choose an alternative optimization algorithm.

9.4 Uncertainty Revisited

Recall from earlier that, asymptotically, the variance-covariance matrix of the estimated parameters is equal to the inverse of the negative of the information matrix.

$$\begin{aligned} \text{var}(\hat{\theta}) &= E\left[\left[\frac{-\partial^2 L}{\partial\theta\partial\theta'}\right]^{-1}\frac{-\partial^2 L}{\partial\theta\partial\theta'}\left[\frac{-\partial^2 L}{\partial\theta\partial\theta'}\right]^{-1}\right] \\ &= \left(-E\left[\frac{-\partial^2 \mathcal{L}}{\partial\theta\partial\theta'}\right]\right)^{-1} \end{aligned} \tag{83}$$

The intuition was that the second derivatives tell us something about the rate at which the slope of the likelihood function changes. If the slope is steep, we can be confident that the maximum we have reached is the 'true' maximum; that is, the variance of $\hat{\beta}$ is small. If the slope is flat, we can't be sure of our maximum, and so the variance around our estimate will be larger.

Sometimes it can be hard to calculate Eq. (83). But as we have just seen, we can substitute the outer product approximation for this. Typically, whatever second derivative matrix a maximization method uses is also used for estimating the variance of the estimated parameters:

Method	'Step Size' (∂^2) matrix	Variance Estimate
Newton	Hessian	Inverse of the negative Hessian
Scoring	Information Matrix	Inverse of the negative information matrix
BHHH	Outer product approximation of information matrix	Inverse of the outer product approximation

References

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. London: Sage Publications.