Maximum Likelihood Estimation (Soc 504)

We need to have a standard way to estimate parameters in models and to estimate the standard deviation of their sampling distribution (their standard errors) for purposes of inference. The most common method of estimation in sociology is Maximum Likelihood Estimation. ML estimation involves a series of steps, which are generally the same from problem to problem:

1. Construct a likelihood function.
2. Simplify the likelihood function and take its logarithm.
3. Take the partial derivative of the log-likelihood function (with respect to each parameter) and set it equal to 0.
4. Solve to find the parameters.
5. Take the second partial derivatives of the log-likelihood. In multiparameter models, this produces a matrix of partial derivatives (called the Hessian matrix).
6. Take the negative of the expectation of this matrix to obtain the "information matrix."
7. Invert this matrix to obtain estimates of the variances of parameters (get standard errors by square-rooting the diagonal elements of the matrix).

This process seems complicated, and, indeed, it can be. Step 4 can be quite difficult when there are lots of parameters. Generally, some sort of iterative method is required to find the maximum. Below I detail the process of ML estimation.

1 Constructing a likelihood function

If $x_1, x_2 \ldots x_n$ are independent observations in a data set, then we know from the multiplication rule in probability theory that the joint probability ($\cap_{i=1}^{n} x_i$) of these observations is:

$$Likelihood\,Function \equiv L(\theta \mid X) \equiv p(X \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta).$$

The $\prod$ symbol represents repeated multiplication. This is the likelihood function for the model. Notice how the parameter and the data switch places in the $L(.)$ notation versus the $p(.)$ notation. We denote this as $L(.)$, because from a classical standpoint, the parameter is assumed to be fixed. The true joint probability density function is represented by $p(.)$. However, we are interested in estimating θ, given the data we have observed, so we use this notation.

The primary point of constructing a likelihood function is that, given the data at hand, we would like to 'solve' for the values of the parameter that make the joint probability of the data most probable. For example, in a series of coin flips, if we obtain heads 5 times, we can be certain that $p = 0$ and $p = 1$ are not likely values for the parameter p. The joint probability of the data we observed would be 0 if $p = 1$ or $p = 0$.

2 Maximizing a likelihood function

So, how do we obtain the estimates for the parameters after we set up the likelihood function? Just as many pdfs are unimodal and slope away from the mode of the distribution, we expect the likelihood function to look about the same. So, what we need to find is the peak of this curve. Thinking about calculus, we realize that where the curve peaks, the slope of the curve should be 0. Thus, we should take the derivative of the likelihood function with respect to the parameter, set it equal to 0, and find the x-coordinate (the parameter value) for which the curve reaches a maximum.

We generally take the logarithm of the likelihood function before we differentiate, though, because the log function converts the repeated multiplication to repeated addition, and repeated addition is much easier to work with:

$$Log - Likelihood \equiv LogL(\theta \mid X) \equiv Log[p(X \mid \theta)] = \sum_{i=1}^{n} p(x_i \mid \theta).$$

The log-likelihood reaches a maximum at the same point as the original function.

3 Getting Standard Errors

A nice additional feature of the log-likelihood is that a function of the second derivative of the log-likelihood function can be used to estimate the standard error of the sampling distribution for the parameter. Specifically, we have to take the inverse of the negative of the second derivative of the log-likelihood function:

$$\left(-E\left(\frac{\partial^2 LL}{\partial\theta\partial\theta^T}\right)\right)^{-1}.$$

With a single parameter, $\frac{d^2LL}{\partial\theta\partial\theta^T}$ (called the Hessian matrix) is a scalar. With multiple parameters, the result will be a matrix. Taking the negative expectation of this scalar/matrix yields the *information matrix* . Inverting this matrix yields a matrix containing the variances of the parameters on its diagonal, and the asymptotic covariances of the parameters in the off-diagonal positions (see examples below). The square root of the diagonal elements yields the standard errors.

The fact that this expression produces the standard errors is not immediately apparent. But, if you recall that the first derivative is a measure of 'speed,' and the second derivative is a measure of 'acceleration,' you can think of the second derivative as telling you the rate of curvature of the curve. A very steep curve, then, has a very high rate of curvature, which

makes its second derivative large. Thus, when we invert it, it makes the standard deviation small. On the other hand, a very shallow curve has a very low rate of curvature, which makes its second derivative small. When we invert a small number, it makes the standard deviation large. Note that when we evaluate the second derivative, we substitute the MLE estimate for the parameter into the result to obtain the standard error at the estimate.

4 Two Examples

4.1 A Binomial Likelihood: Is a coin fair?

Suppose someone handed you a coin and claimed that the coin was unfairly weighted, because on 10 flips, they had obtained 7 heads. Is this sufficient evidence that the coin is unfair? As we have already discussed previously, the sampling distribution for coin flips is a binomial distribution with parameters n and p (or, each flip follows a Bernoulli distribution with parameter p). Thus, one way to think about this question statistically is to ask whether it is reasonable that the p parameter is .5.

In answering this question, we need to determine our best guess for p using maximum likelihood estimation. The first step in this process is to construct a likelihood function for the data. If each observation (flip) is Bernoulli-distributed, and the observations are independent, then the likelihood function is simply the multiple of 10 Bernoulli densities:

$$L(p \mid X) \propto \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$$

The proportionality symbol is used here, because I have omitted a constant (the combinatorial expression). Notice that the likelihood function here could be obtained by realizing that the result of $x = 7$ heads out of 10 tosses could be represented as a binomial density, in which case:

$$L(p \mid X) = \left(\begin{array}{c} 10 \\ 7 \end{array} \right) p^{x=7}(1-p)^{n-7=3}$$

In this representation, the likelihood function includes the proportionality constant. As it turns out, these representations are equivalent, so long as the observations are considered to be 'exchangeable,' meaning the ordering of the heads versus tails is irrelevant (this is the factor that the combinatorial expression corrects for in the binomial density). Both representations will yield the same results, as we will see, because the constant term drops in the derivative step below.

The likelihood function can be simplified to be:

$$L(p \mid X) \propto p^{\sum_{i=1}^{10} x_i}(1-p)^{n-\sum_{i=1}^{10} x_i}$$

The next step in ML estimation is to take the *log* of the likelihood function:

$$LogL(p \mid X) \equiv LL(p) \propto \left(\sum x_i\right) log(p) + (n - \sum x_i) log(1-p)$$

Next, we take the derivative with respect to p:

$$\frac{dLL}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1 - p}$$

To obtain a maximum, we simply set this expression equal to 0 and solve for p:

$$\frac{n - \sum x_i}{1 - p} = \frac{\sum x_i}{p}$$

Thus,

$$\hat{p} = \frac{\sum x_i}{n}.$$

This shows that the maximum likelihood estimate for p is simply the observed proportion of successes (here, .7). At this point, it may appear that the maximum likelihood estimate supports the individual's contention that the coin is unfair. However, we must consider the uncertainty in our estimate that is introduced by having such a small sample: there may well be cases in which even a fair coin will yield 7 heads on 10 flips. We can capture such uncertainty in our estimate of p by constructing the information matrix and inverting it. The second derivative of the binomial likelihood is:

$$\frac{\partial^2 LL}{\partial p^2} = -\frac{\sum x}{p^2} - \frac{n - \sum x}{(1 - p)^2}$$

Taking expectations yields:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = E\left[-\frac{\sum x}{p^2} - \frac{n - \sum x}{(1 - p)^2}\right]$$

The expectation of these expressions can be computed by realizing that the expectation of $\frac{\sum x}{n}$ is p. Thus:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = -\frac{np}{p^2} - \frac{n - np}{(1 - p)^2}$$

Some simplification yields:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = -\frac{n}{p(1 - p)}$$

At this point, we can negate the expectation, invert it, and evaluate it at the MLE ($\hat{p}$) to obtain:

$$I(p)^{-1} = \frac{\hat{p}(1 - \hat{p})}{n}$$

Taking the square root of this yields the estimated standard error. In this case, the standard error is $\sqrt{\frac{(.7)(.3)}{10}} = .14$. We can construct our usual confidence interval around the maximum likelihood estimate to obtain a 95% interval for our ML estimate, or, alternatively, we can conduct a t test:

$$t = \frac{(.7 - .5)}{.14} = 1.43$$

The observed t statistic is below the 'critical value' of 2.26 (for a two-tailed test; 1.83 for a one-tailed test, $\alpha = .05$), and so we would probably conclude that the sample offers no evidence to suggest the coin is not fair.

4.2 A Normal Likelihood

Because the normal distribution will be used repeatedly throughout the remainder of the semester, I keep the following example at a general level. Suppose you have n observations $x_1, x_2, \ldots, x_n$ that you assume are normally distributed. Once again, if the observations are assumed to be independent, a likelihood function can be constructed as the multiple of independent normal density functions:

$$L(\mu, \sigma \mid X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

We can simplify the likelihood as:

$$L(\mu, \sigma \mid X) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

The log of the likelihood is:

$$LL(\mu, \sigma \mid X) \propto -nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

In the above, I have eliminated the $-\frac{n}{2}log(2\pi)$, an irrelevant constant. In this example, we have two parameters, μ and σ, and hence the first derivative must be taken with respect to each parameter. This will leave us with two equations (one for each parameter). After taking the derivatives with respect to each parameter, we obtain the following:

$$\frac{\partial LL}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial LL}{\partial \mu} = -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)$$

Setting these partial derivatives each equal to 0 and doing a little algebra yields:

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

These estimators should look familiar: the MLE for the population mean is the sample mean; the MLE for the population standard deviation is the sample standard deviation (note: the MLE is known to be biased, and hence a correction is added, so that the denominator is $n - 1$ rather than n).

Estimates of the variability in the estimates for the mean and standard deviation can be obtained as we did in the binomial example. However, as noted above, given that we have two parameters, our second derivate matrix will, in fact, be a matrix. For the purposes of avoiding taking square roots until the end, let $\tau = \sigma^2$, and we'll construct the Hessian matrix in terms of τ. Also, let θ be a vector containing both μ and τ. Thus, we must compute:

$$\frac{\partial^2 LL}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2 LL}{\partial\mu^2} & \frac{\partial^2 LL}{\partial\mu\partial\tau} \\ \frac{\partial^2 LL}{\partial\tau\partial\mu} & \frac{\partial^2 LL}{\partial\tau^2} \end{bmatrix}$$

Without showing all the derivatives, the elements of the Hessian matrix are then:

$$\frac{\partial^2 LL}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{-n}{\tau} & \frac{n\mu - \sum_{i=1}^{n} x_i}{\tau^2} \\ \frac{n\mu - \sum_{i=1}^{n} x_i}{\tau^2} & \frac{n}{2\tau^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\tau^3} \end{bmatrix}$$

In order to obtain the information matrix, which can be used to compute the standard errors, we must take the expectation of this Hessian matrix and take its negative. Let's take the expectation of the diagonal elements first. Those elements can be rewritten as: $\frac{n(\mu-\bar{x})}{\tau^2}$. The expectation of $\mu - \bar{x}$ is 0, making the off-diagonal elements of the information matrix equal to 0. This should make some sense: there need be no relationship between the mean and variance in a normal distribution.

The first element, $\left(\frac{-n}{\tau}\right)$, is unchanged under expectation. Thus, after substituting σ^2 back in for τ and negating the result, we obtain $\frac{n}{\sigma^2}$ for this element of the information matrix.

The last element, $\frac{n}{2\tau^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\tau^3}$, requires a little thinking. The only part of this expression that changes under expectation is $\sum_{i=1}^{n}(x_i-\mu)^2$. The expectation of this expression is $n\tau$. That is, $E(x_i-\mu)^2$ is τ, and we are taking this value n times (notice the summation). Thus, this element, after a little algebraic manipulation, negation, and substitution of σ^2 for τ, becomes: $\frac{n}{2\sigma^4}$. So, our information matrix appears as:

$$I(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

To obtain standard errors, we need to a) invert this matrix, and b) take the square root of the diagonal elements (variances) to obtain the standard errors. We have already discussed the complexities of matrix inversion. Fortunately, however, the process in this case is quite simple, given that the off-diagonal elements are equal to 0. In this case, the inverse of the matrix is simply the inverse of the diagonal elements.

So, ultimately, if we invert and square root the elements of the information matrix, we find that the estimate for the standard error for our estimate $\hat{\mu}$ is $\frac{\hat{\sigma}}{\sqrt{n}}$, and our estimate for the standard error for $\hat{\sigma^2}$ is $\hat{\sigma^2}\sqrt{\frac{2}{n}}$. The estimate for the standard error for $\hat{\mu}$ should look familiar: as we discussed previously, it is the standard deviation of the sampling distribution for a mean.