LECTURE NOTE 6: MAXIMUM LIKELIHOOD ESTIMATION

1 Introduction

Soon we will be discussing methods for analyzing categorical outcome variables, such as logit and probit models. Although one can estimate these models using a non-linear variant of least squares, most statistical software uses another method called maximum likelihood. In this lecture note, we first develop a general framework for maximum likelihood estimation. Then we explore its application to a Bernoulli random variable.

2 General Case

Maximum likelihood estimation (MLE) departs from the way we have been approaching estimation. The least squares approach asked: what parameter value is least wrong in describing the observed data points? In contrast, maximum likelihood asks: what parameter value makes the observed data points most likely?

Let X be a random variable with probability density function $f(x;\theta)$, where θ is a vector of parameters. Suppose we have an i.i.d. sample of observations $\{X_i\}_{i=1}^N$, and each observation i has realization x_i. The maximum likelihood estimator of θ chooses the value $\hat{\theta}$ that maximizes the joint probability of observing $X_i = x_i$ for all i. Because the X_i's are independent, we can write this joint probability as:

$$L = \prod_{i=1}^{N} f(x_i;\theta)$$

L is known as the *likelihood* or the *likelihood function*. Unfortunately, θ is unknown, so we cannot calculate L directly. But we can try to plug in candidate values $\tilde{\theta}$ to see how likely the observed data are, given the value of $\tilde{\theta}$. The maximum likelihood estimator $\hat{\theta}$ is the value of $\tilde{\theta}$ that maximizes the likelihood, i.e. the solution to:

$$\max_{\tilde{\theta}} L = \max_{\tilde{\theta}} \prod_{i=1}^{N} f(x_i;\tilde{\theta})$$

It's difficult to maximize a very long product. We can simplify by taking the logarithm of the likelihood, which turns the product into a sum. Because the logarithm is a strictly increasing transformation, the value of $\tilde{\theta}$ that maximizes L also maximizes $\ln(L)$. So the MLE problem becomes:

$$\max_{\tilde{\theta}} \ln L = \max_{\tilde{\theta}} \sum_{i=1}^{N} \ln\left(f(x_i;\tilde{\theta})\right)$$

The solution $\hat{\theta}$ satisfies $\frac{\partial \ln L}{\partial \hat{\theta}} = 0$.

Under fairly innocuous assumptions, the MLE satisfies:

1. Consistency: $\hat{\theta} \xrightarrow{p} \theta$.
2. Asymptotic Normality: $\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta, \Sigma)$. In large samples, $\hat{\theta}$ is approximately normally distributed with a variance-covariance matrix that we here call Σ. There is an explicit solution for Σ, but you do not need to know it.[1]
3. Asymptotic Efficiency: No other asymptotically unbiased estimator has a smaller variance than $\hat{\theta}$.

Consistency is a basic property we want for nearly every estimator we use. Asymptotic normality implies that we can compute p-values and CIs using the same normal approximations we used with OLS. Asymptotic efficiency tells us that the MLE is in some sense a "best" estimator. It is the reason that most statistical software estimates logits and probits using maximum likelihood rather than non-linear least squares. In general, maximum likelihood is more efficient.

3 Example: Estimating p for a Bernoulli Random Variable

Consider a Bernoulli random variable X_i that takes on value 1 with probability p and 0 with probability $1-p$. Suppose we observe a sample with three i.i.d. observations, $(1,1,0)$. The likelihood function is:

$$\begin{aligned} L &= Pr[X_1 = 1] \cdot Pr[X_2 = 1] \cdot Pr[X_3 = 0] \\ &= p \cdot p \cdot (1-p) \\ &= p^2(1-p) \end{aligned}$$

Thus, the MLE problem becomes:

$$\max_{\tilde{p}} L = \max_{\tilde{p}} \tilde{p}^2(1-\tilde{p})$$

Or in terms of the log-likelihood:

$$\max_{\tilde{p}} \ln L = \max_{\tilde{p}} 2\ln(\tilde{p}) + \ln(1-\tilde{p})$$

[1] In case you are interested: Σ equals minus the second derivative matrix of the log likelihood function. The intuition is that the second derivative measures the curvature of the likelihood function. When the likelihood function is very curved, we can be most confident in the optimum we chose as our MLE.

Both maximization problems lead to the same first order condition, but just for example, consider the first order condition for the second maximization problem: $\frac{d \ln L}{d\hat{p}} = \frac{2}{\hat{p}} - \frac{1}{1-\hat{p}} = 0$. Rearranging terms, we obtain $\hat{p} = \frac{2}{3}$, which is exactly the same as our usual estimator for p, $\bar{X}$.

Let's generalize to an i.i.d. sample of N Bernoulli random variables with probability of success p. Using similar logic to the $N = 3$ example above, if we observe S successes in the sample, the likelihood is:

$$L = p^S (1-p)^{N-S}$$

We then maximize the log-likelihood:

$$\max_{\tilde{p}} \ln L = \max_{\tilde{p}} S \ln(\tilde{p}) + (N-S) \ln(1-\tilde{p})$$

which gives us the first order condition $\hat{p} = \frac{S}{N}$: again exactly the same as the sample average.

4 Maximum Likelihood Estimation in Practice

In the example above, we solved the maximum likelihood problem analytically. But most of the models we estimate by maximum likelihood lack a closed-form solution. We have three options for finding $\hat{\theta}^{MLE}$:

1. Analytic optimization: As already mentioned, we can differentiate the likelihood function, set the first derivative to zero, and solve. We should check the second derivative to make sure we have found a maximum (rather than a minimum).
2. (Undirected) grid search: If we know that θ lies in some range $[\underline{\theta}, \bar{\theta}]$, then we can calculate L for all the values of a grid over $[\underline{\theta}, \bar{\theta}]$ and choose the value that produces the highest L.
3. (Directed) numerical optimization: A number of algorithms exist to use the properties of the likelihood function (e.g., the first and second derivatives) to direct numerical search. So we choose an initial value, $\hat{\theta}^0$, and we then allow the algorithm to find the optimum.

The most attractive method is (1), analytic optimization, but this option is often impossible. Option (2) can be extremely time-intensive, for two reasons. First, we may not have strong prior beliefs about a narrow range $[\underline{\theta}, \bar{\theta}]$, in which case we would have to compute L for many possible parameter values. Second, θ may be a vector of many parameters, in which case we would have to search over a complex, multi-dimensional grid. As a result of these limitations of options (1) and (2), we use option (3) for most maximum likelihood estimation. Numerical optimization works quite well if the likelihood function has only one optimum, as it does in most of the models we study in this course. Of course, if the likelihood function has many local maxima, then we run the risk of choosing the wrong maximum. This concern is not important in most applications.