PLSC 503: "Multivariate Analysis for Political Research"

Introduction to Maximum Likelihood
March 30, 2017

Introduction

Consider a model that looks like this:

$$Y_i \sim N(\mu, \sigma^2)$$

So:

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ \mathrm{Var}(Y) &= \sigma^2 \end{aligned}$$

Suppose you have some data on Y, and want to estimate μ and σ^2 from those data...

The whole idea of *likelihood* is to *find the estimate of the parameter(s) that maximizes the probability of the data.*

Example: Suppose Y is income of assistant professors (in thousands of dollars), and we have a random sample of five data points:

$$\begin{aligned} Y &= 64 \\ &\quad 63 \\ &\quad 59 \\ &\quad 71 \\ &\quad 68 \end{aligned}$$

Intuitively, what are the odds that these five data points were drawn from a normal distribution with $\mu = 120$? (Answer: Not very likely).

What about $\mu = 65$ (which happens to be the empirical mean from this sample)? (More likely – WHY?).

What maximum likelihood is, is a systematic way of doing exactly this.

Think of the salaries as draws from a normal distribution.

We can write:

$$\Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \quad (1)$$

This is the density, or probability density function (PDF) of the variable Y.

- The probability that, for any one observation i, Y will take on the particular value y.
- This is a function of μ, the expected value of the distribution, and σ^2, the variability of the distribution around that mean.

We can think of the probability of a single realization being what it is, e.g.:

$$\begin{aligned}\Pr(Y_1 = 64) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(64-\mu)^2}{2\sigma^2}\right] \\ \Pr(Y_2 = 63) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(63-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

etc.

Now, we're interested in getting estimates of the parameters μ and σ^2, based on the data...

If we assume that the observations on Y_i are independent (i.e. not related to one another), then we can consider the joint probability of the observations as simply the product of the marginals.

Recall that, for independent events A and B:

$$\Pr(A, B) = \Pr(A) \times \Pr(B)$$

So:

$$\Pr(Y_1 = 64, Y_2 = 63) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(64-\mu)^2}{2\sigma^2}\right] \times \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(63-\mu)^2}{2\sigma^2}\right]$$

More generally, for N independent observations, we can write the joint probability of each realization of Y_i as the product of the N marginal probabilities:

$$\Pr(Y_i = y_i \forall i) \equiv L(Y|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \quad (2)$$

This product is generally known as the *Likelihood* $[L(Y)]$, and is the probability that each observation is what it is, given the parameters.

Estimation

Of course, we don't know the parameters; in fact, they're what we want to figure out. That is, we want to know the likelihood of some values of μ and σ^2, given Y. This turns out to be proportional to $L(Y|\mu, \sigma^2)$:

$$L(\hat{\mu}, \hat{\sigma}^2|Y) \propto \Pr(Y|\hat{\mu}, \hat{\sigma}^2)$$

We can get at this by treating the likelihood as a function (which it is). The basic idea is *to find the values of μ and σ^2 that maximize the function; i.e., those which have the greatest likelihood of having generated the data.*

How do we do this?

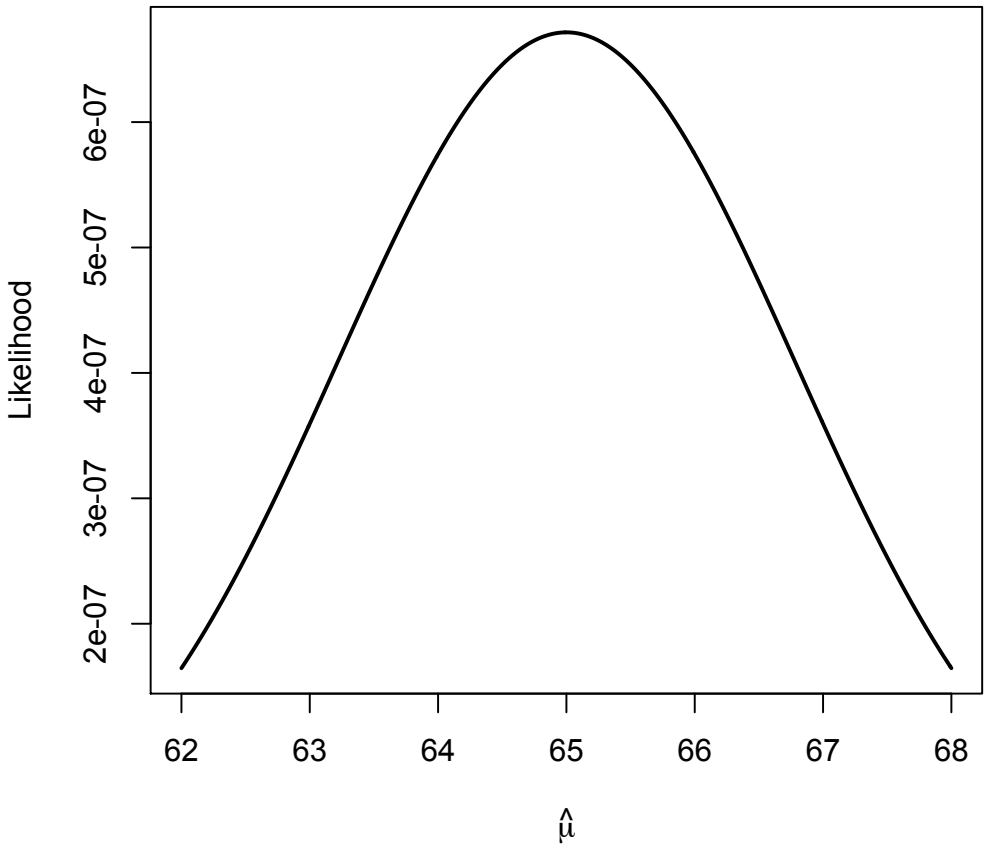
One way would be to start plugging in values for μ and σ^2 and seeing what the corresponding likelihood was...

E.g.: For $\hat{\mu} = 68$ and $\hat{\sigma} = 4$, we would get:

$$\begin{aligned} L &= \frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(64-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(63-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(59-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(71-68)^2}{32}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(68-68)^2}{32}\right] \\ &= 0.0000001645725 \end{aligned}$$

More generally, we can graph the likelihood for these five observations (assuming, for the moment, that $\sigma^2 = 16$) for different possible values of μ:

Figure 1: Likelihood of $\hat{\mu}$ for five sample data points



Note that the likelihood is maximized at $\mu = 65$, the empirical mean.

But...

- ...likelihood functions often look scary,
- ...products are generally hard to deal with, and
- ...we run into issues with numeric precision when we get into teeeeny joint probabilities. Here, the range of values for the (joint) likelihood is from 0.0000002 to 0.00000065 (in other words, pretty small).

Fortunately, it turns out that if we find the values of the parameters that maximize any monotone increasing transformation of the likelihood function, those are also the parameter

values that maximize the function itself.

Most often we take natural logs,[1] giving something called the *log-likelihood*:

$$\begin{aligned}
\ln L(\hat{\mu}, \hat{\sigma}^2|Y) &= \ln\left\{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right]\right\} \\
&= \sum_{i=1}^{N} \ln\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}\ln(2\pi) - \left[\sum_{i=1}^{N}\frac{1}{2}\ln\,\sigma^2 - \frac{1}{2\sigma^2}(Y_i-\mu)^2\right] \qquad (3)
\end{aligned}$$

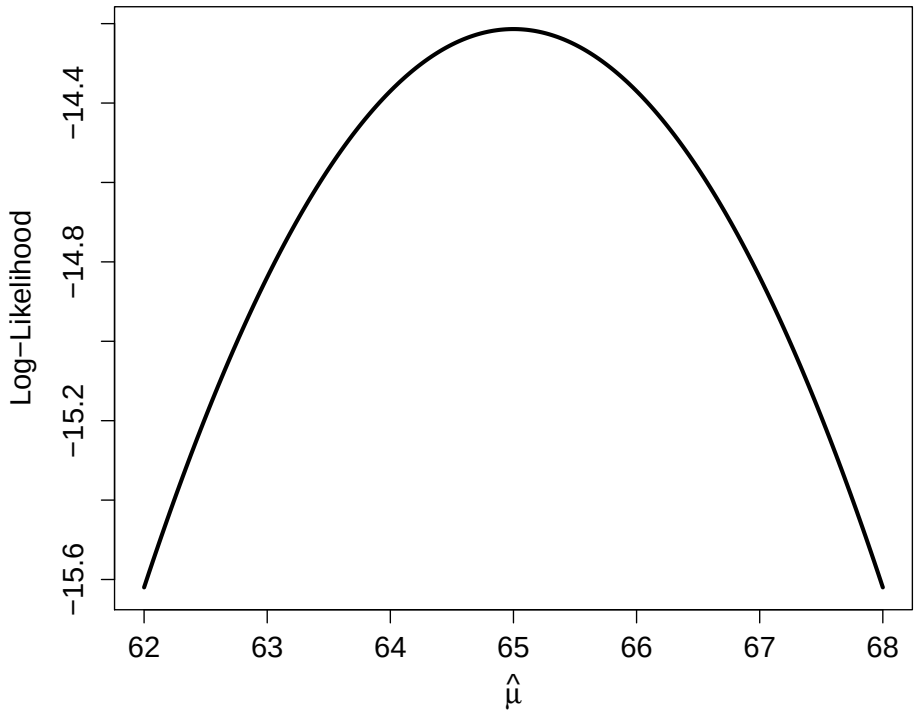
This means that – for $\sigma = 4$ and $\hat{\mu} = 68$ we would get:

$$\begin{aligned}
\ln L &= -2\ln(2\pi) - \ln\left\{\frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(64-68)^2}{32}\right]\right\} + \\
&\quad \ln\left\{\frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(63-68)^2}{32}\right]\right\} + \\
&\quad \ln\left\{\frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(59-68)^2}{32}\right]\right\} + \\
&\quad \ln\left\{\frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(71-68)^2}{32}\right]\right\} + \\
&\quad \ln\left\{\frac{1}{\sqrt{2\pi 16}}\exp\left[-\frac{(68-68)^2}{32}\right]\right\} \\
&= -2.80523 - 3.08648 - 4.83648 - 2.58648 - 2.30523 \\
&= -15.6199
\end{aligned}$$

More generally, if we again fix $\sigma^2 = 16$ and consider this *log-likelihood* the same way we did above, we get the figure below. Notice that the maximum is at the same value as before, but the actual values of the log-likelihood are much more reasonable. Moreover, we get to work with sums rather than products, which makes the math a lot easier.

[1] There are a lot of mathematical reasons for this choice, but the two most compelling are (a) the log transformation turns products into sums, and (b) it simplifies the Taylor series math discussed below.

Figure 2: Log-likelihood of $\hat{\mu}$ for five sample data points



Maximization

Question: *Graphs are nice, but how do we normally find a maximum?*

Answer: Good old **differential calculus**...

What happens when we take the first derivatives of (3) with respect to μ and σ^2?

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{N} (Y_i - \mu) \\ \frac{\partial \ln L}{\partial \sigma^2} &= \frac{-N}{2\sigma^2} + \frac{1}{2}\sigma^4 \sum_{i=1}^{N} (Y_i - \mu)^2 \end{aligned}$$

If we set these equal to zero and solve for the two unknowns, we get:

$$\begin{aligned}\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} Y_i \\ \hat{\sigma^2} &= \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2\end{aligned}$$

...which are the basic formulas for mean and variance. That is, our standard versions of $\hat{\mu}$ and $\hat{\sigma^2}$ are the *maximum likelihood estimators* for μ and σ^2.

MLE and the Linear Regression Model

Now suppose we want to set the mean of Y to be a function of some other variable X; that is, you suspect that professors' salaries are a function of, e.g., gender, or something. We write:

$$\begin{aligned}\mathrm{E}(Y) \equiv \mu &= \beta_0 + \beta_1 X_i \\ \mathrm{Var}(Y) &= \sigma^2\end{aligned}$$

We can then just substitute this equation in for the systematic mean part (μ) in the previous equations...

E.g.:

$$L(\beta_0, \beta_1, \sigma^2|Y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right] \tag{4}$$

and:

$$\begin{aligned}\ln L(\beta_0, \beta_1, \sigma^2|Y) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right] \\ &= -\frac{N}{2}\ln(2\pi) - \sum_{i=1}^{N}\left[\frac{1}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \qquad (5)\end{aligned}$$

With respect to the parameters $\{\beta_0, \beta_1, \sigma^2\}$, only the last term is important...

- The first one $\left(-\frac{N}{2}\ln(2\pi)\right)$ is invariant with respect to the parameters of interest, and so can be dropped.
- This is due to something called the *Fisher-Neyman Factorization Lemma*.

Thus, the *kernel* of the log-likelihood is:

$$-\sum_{i=1}^{N}\left[\frac{1}{2}\ln\sigma^2+\frac{1}{2\sigma^2}(Y_i-\beta_0-\beta_1X_i)^2\right] \tag{6}$$

...which is the old familiar sums-of-squared-residuals term, scaled by the variance parameter σ^2.

This leads us to several interesting things:

- The least-squares estimator of the OLS βs *is the maximum likelihood estimator as well.*
- MLE is not confined to models with "ugly" dependent variables (though it *is* highly useful for them).

MLE Properties

Now consider a very general model of the form

$$\Pr(Y)=f(\mathbf{X},\theta). \tag{7}$$

We can write down a general likelihood for this model, which is simply

$$L=\prod_{i=1}^{N}f(Y_i|\mathbf{X}_i,\theta) \tag{8}$$

and

$$\ln L=\sum_{i=1}^{N}\ln f(Y_i|\mathbf{X}_i,\theta). \tag{9}$$

where θ is the "true" parameter vector of length k. Assume for the sake of argument that we've obtained a set of parameter estimates $\hat{\theta}$ that satisfy

$$\ln L(\hat{\theta}|Y,\mathbf{X})=\max_{\theta}\{\ln L(\theta|Y,\mathbf{X})\}$$

that is, that are MLEs. (We'll worry about exactly where those come from a bit later...). To clean things up a bit, let's drop the conditioning on $\mathbf{X}$ and Y, so that we can express the log-likelihood at its maximum as $\ln L(\hat{\theta})$.

Consistency

What can we say about this function? Think first about its first derivative, taken with respect to the estimates of the parameters of interest, which we've seen before:

$$\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \hat{\theta}}. \tag{10}$$

This is the *gradient*; intuitively, this describes the "slope" of the hyperplane "tangent to" the (log-)likelihood surface.

What would we expect of this at a maximum? Of course, we'd expect it to be uniformly zero, since that's the first-order condition for a maximum of a function. In fact, it *has* to be zero at $\hat{\theta}$, something we'll come back to in a bit.

Intuitively, it would be nice to know what this function "looked like" in the vicinity of the true parameter value θ. We can accomplish this through a first-order Taylor series expansion (sometimes called a "linearization") of the function around θ:

$$\frac{\partial \ln L}{\partial \hat{\theta}} = \frac{\partial \ln L}{\partial \theta} + \frac{\partial^2 \ln L}{\partial \theta^2}(\hat{\theta} - \theta). \tag{11}$$

The front half of that second term in the linearization is the "Hessian" we discussed before: the $k \times k$ matrix of second derivatives and cross-partials. Note that, because $\hat{\theta}$ is at a maximum, we know that the term on the left-hand side is equal to zero. This means that we can rewrite (11) as:

$$\begin{aligned} \hat{\theta} - \theta &= \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1} \frac{\partial \ln L}{\partial \theta} \\ &= -\mathbf{H}(\theta)^{-1}\mathbf{g}(\theta) \end{aligned} \tag{12}$$

At this point, it would be great if we could show analytically that the right-hand side of this equation was equal to zero; that would mean our general MLE was an unbiased estimator of θ. But, we can't; instead, we can do the next best thing, which is to show that

$$\text{plim}(\hat{\theta} - \theta) = 0;$$

that is, that the MLE is consistent. To do this, we need to do two things:

1. Multiply the first term by N, and divide the second by N as well (this obviously doesn't change anything, but makes our life easier).
2. *Assume* that the first term $\mathbf{H}(\theta)$ converges to some finite value. This is not a heroic assumption.
3. *Show* that the second (gradient) term goes to zero in expectation as $N \to \infty$

This latter step is not that hard. We can write the expectation of the gradient as

$$\mathrm{E}[\mathbf{g}(\theta)] = \frac{1}{N}\,\mathrm{E}\left(\frac{\partial \ln L_1}{\partial \theta} + \frac{\partial \ln L_2}{\partial \theta} + ... + \frac{\partial \ln L_N}{\partial \theta}\right) \tag{13}$$

which, by the law of large numbers, is

$$\mathrm{E}[\mathbf{g}(\theta)] = \frac{1}{N}\left[\mathrm{E}\left(\frac{\partial \ln L_1}{\partial \theta}\right) + \mathrm{E}\left(\frac{\partial \ln L_2}{\partial \theta}\right) + ...\right] \tag{14}$$

Note that (a) by i.i.d., each of the expectations inside the brackets are all identical, and (b) because the true parameter value is assumed to be a maximum of the likelihood, the expectation of the gradient of the likelihood there is zero. Thus, the second term converges in expectation to zero as $N \to \infty$, and the MLE estimate $\hat{\theta}$ is a consistent estimate of θ.

Consistency means that MLEs are intrinsically "large-sample" estimators. MLEs *will* be biased in small samples, and the degree of that bias can be substantial. As a practical matter, the question of "how much data are needed" is one that we'll address when we get to talking about specific estimators.

Efficiency

To assess the relative efficiency of MLEs, we have to start with the famous Cramer-Rao theorem. That theorem states that – given certain regularity conditions – the variance of *any* estimator of a parameter θ has a minimum possible value (a "lower bound"). Specifically,

$$\mathrm{Var}(\hat{\theta}) \geq \left[-\mathrm{E}\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right)\right]^{-1}$$

Any estimator that attains this lower bound is therefore fully efficient.

We can calculate the variance (that is, the variance-covariance matrix) of the MLE using what we know about variances and our previous Taylor series linearization. The general equation for the variance is:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \\
&= \mathrm{E}\left[\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1} \frac{\partial \ln L}{\partial \theta}\frac{\partial \ln L'}{\partial \theta}\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\right]
\end{aligned} \tag{15}$$

The middle term is known as the "outer product of the gradient" (sometimes "OPG"). Imposing some additional regularity conditions causes the expectation of the OPG to reduce to

$$\mathrm{E}\left[\frac{\partial \ln L}{\partial \theta}\frac{\partial \ln L'}{\partial \theta}\right] = \mathrm{E}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] \tag{16}$$

which in turn means (15) reduces to:

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \left[-\mathrm{E}\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)\right]^{-1} \\ &= [\mathbf{I}(\theta)]^{-1} \end{aligned} \tag{17}$$

where $\mathbf{I}(\theta)$ denotes the Fisher information matrix, the negative expectation of the Hessian of $\ln L$ with respect to θ. (17) illustrates that the MLE achieves the Cramer-Rao lower bound in (15); thus showing that the MLE is asymptotically efficient.

In addition, in some instances an estimator exists that is fully (rather than just asymptotically) efficient; such an estimator is often called the "minimum variance unbiased estimator" (MVUE). If such an estimator exists, one can show that the MLE will be it. Thus, MLE's efficiency will always be maximal: they are fully efficient when such an estimator exists, and asymptotically so if not. As a result, "more of the ML estimates that could result across hypothetical experiments are concentrated around the true parameter value than is the case for any other estimator in this class" (King 1998, 80).

Asymptotic Normality

We can combine the results for consistency and efficiency to show that MLEs asymptotic distribution are multivariate normal. In particular, if we consider the limiting distribution of

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathbf{I}(\theta)^{-1}}}$$

as $N \to \infty$, a few things become apparent:

1. The fact of consistency means that the limiting distribution will be centered around zero.
2. The fact that the second term in (12), as expressed in (14), is a series of i.i.d. random variates means (thanks to the central limit theorem) that their distribution will be Normal.
3. The fact that the inverse of the information matrix is the asymptotic variance of the MLE.

Taken together, these things mean that:

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathbf{I}(\theta)^{-1}}} \sim N(\mathbf{0}, \mathbf{1}) \tag{18}$$

or, alternatively,

$$\hat{\theta} \sim N(\theta, \mathbf{I}(\theta)^{-1})$$

which in turn means that we can use more or less standard approaches to testing, inference, etc. (more on that next semester...).

Invariances

MLEs have a couple of useful "invariances." What this means is that they provide the same estimates irrespective of some relatively arbitrary changes or differences in conditions.

The first of these is *invariance to reparameterization*. This is means that any one-to-one reparameterization of an MLE will yield the same maxima. Thus, rather than estimating a parameter θ, we can instead estimate some function of it $g(\theta)$, and then recover an estimate of θ (that is, $\hat{\theta}$) from $g(\theta)$.

As an example, consider the Normal distribution we discussed previously. We know that the kernel of the traditional $N(\mu, \sigma^2)$ log-likelihood is:

$$\ln L(\hat{\mu}, \hat{\sigma}^2) = -\left[\sum_{i=1}^{N} \frac{1}{2} \ln\, \sigma^2 - \frac{1}{2\sigma^2}(Y_i - \mu)^2\right].$$

Suppose we reparameterize the Normal so that the variance is decreasing in the relevant parameter, rather than increasing; we can do this by defining:

$$\phi^2 = 1/\sigma^2$$

so that the "new Normal" is now $N(\mu, \phi^2)$. The kernel of this new log-likelihood is thus

$$\ln L(\hat{\mu}, \hat{\phi}^2) = -\left[\sum_{i=1}^{N} \frac{1}{2} \ln\, \phi^2 - \frac{1}{2\phi^2}(Y_i - \mu)^2\right]. \tag{19}$$

with the score equation

$$\frac{\partial \ln L}{\partial \phi^2} = \frac{-N}{2\phi^2} + \frac{1}{2}\phi^4 \sum_{i=1}^{N}(Y_i - \mu)^2. \tag{20}$$

Setting this equal to zero, multiplying both sides by ϕ^4, and solving yields

$$\begin{aligned} \hat{\phi}^2 &= \frac{N}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \\ &= \frac{1}{\hat{\sigma}^2} \end{aligned} \tag{21}$$

The second invariance is *invariance to sampling plans*. This means that ML estimators are the same irrespective of the rule used for determining sample size. Formally, information in the data only affect the estimation through the likelihood. This seems like a minor thing, but in fact its very useful, since it means that you can (e.g.) "pool" data and get "good" estimates without regard for how the sample size was chosen.

Summary

To summarize, MLEs:

- Maximize $L(\theta|Y, \mathbf{X})$
- Are consistent in N
- Are asymptotically efficient
- Are asymptotically Normal
- Are invariant to (injective) transformations and varying sampling methods

Next time: Optimization...

Appendix I: Approximating a Function Via a Taylor Series Expansion

Sometimes it is useful to approximate the shape of a function in the "neighborhood" of some point a. We can do so by something called a "Taylor series expansion," which makes use of the fact that a continuous, infinitely differentiable function can be approximated arbitrarily well by a high-order polynomial. Formally:

$$\begin{aligned} f(X) &\approx \sum_{i=0}^{\infty} \frac{f^{[i]}(a)}{i!}(x-a)^i \\ &= f(a) + \frac{f'(a)}{1!}(x-a)^1 + \frac{f''(a)}{2!}(x-a)^2 + ... \end{aligned}$$

Notice that the approximation is expressed as a series of polynomial functions of X; the first term contains x^0, the second x^1, the third x^2, etc. For example, suppose we want to approximate the function $f(X) = -2\exp(2X)$ in the area around $X = 0$. The first-order approximation is:

$$f(x) \approx -2\exp[2(0)] - \frac{4\exp[2(0)]}{1!}(X-0)^1 = -2 - 4X$$

and the second-order is:

$$\begin{aligned} f(x) &\approx -2\exp[2(0)] - \frac{4\exp[2(0)]}{1!}(X-0)^1 - \frac{8\exp[2(0)]}{2!}(X-0)^2 \\ &= -2 - 4X - 4X^2 \end{aligned}$$

and so forth.

We can do a multivariate Taylor series expansion in a similar way, using the gradients and Hessians we discussed above:

First-order: $$f(\mathbf{X}) \approx f(\mathbf{a}) + \mathbf{g}(\mathbf{a})^\top(\mathbf{x}-\mathbf{a})$$

Second-order: $$f(\mathbf{X}) \approx f(\mathbf{a}) + \mathbf{g}(\mathbf{a})^\top(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{H}(\mathbf{x}-\mathbf{a})$$

Etc. This in turn leads us to yet another application...

The Delta Method

The "delta method" is a means of approximating the mean and variance of a nonlinear function $f(X)$ of a random variable X with known mean and variance. We often need to use this because, as a rule, $\mathrm{E}[f(X)] \neq f[\mathrm{E}(X)]$ and $\mathrm{Var}[f(X)] \neq f[\mathrm{Var}(X)]$.

For some X where $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$,

$$\begin{aligned} f(X) &\approx f(\mu) + (x - \mu)f'(\mu) \\ &= [f(\mu) - \mu f'(\mu)] + xf'(\mu) \\ &= \quad a \quad + \quad xb. \end{aligned}$$

That is, the first-order (linear) Taylor series approximation yields a form for the function that is linear in X. Recalling that

$$\mathrm{E}(a + bX) = a + b\mathrm{E}(X)$$

and

$$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X),$$

this means that

$$\begin{aligned} \mathrm{E}[f(X)] &\approx [f(\mu) - \mu f'(\mu)] + \mathrm{E}(X)f'(\mu) \\ &\approx f(\mu) - \mu f'(\mu) + \mu f'(\mu) \\ &\approx f(\mu) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[f(X)] &\approx [f'(\mu)]^2 \times \mathrm{Var}(X) \\ &\approx [f'(\mu)]^2 \times \sigma^2 \end{aligned}$$

This allows for straightforward expressions for the mean and variance of a function.

Appendix II: R Code for Figures

```
data<-c(64,63,59,71,68) # data
muhat<-seq(62,68,by=0.01) # parameter space
nrow<-c(1:length(muhat))
ncol=c(1:length(data))

Ls.4<-matrix(nrow=length(muhat),ncol=length(data))

for (i in nrow) {
  for (j in ncol){
    Ls.4[i,j]<-dnorm(data[j],mean=muhat[i],sd=4) # Likelihoods for each observed Y | mu
  }
}

L.4<-apply(Ls.4,1,prod) # Product of the individual likelihoods

# Loops AND apply()s. Prof is cray-cray!
# Likelihood plot

plot(muhat,L.4,t="l",lwd=3,ylab="Likelihood",xlab=expression(hat(mu)))

# Log-likelihood plot

LnLs.4<-matrix(nrow=length(muhat),ncol=length(data))

for (i in nrow) {
  for (j in ncol){
    LnLs.4[i,j]<-log(dnorm(data[j],mean=muhat[i],sd=4)) # Individual log-likelihoods
  }
}

LnL.4<-apply(LnLs.4,1,sum) # Sum of the log-likelihoods

plot(muhat,LnL.4,t="l",lwd=3,ylab="Log-Likelihood",xlab=expression(hat(mu)))
```