M-Estimation (or Unbiased Estimating Equations)

If we think back to introductory statistics classes, there are several general principles that we often end up following in order to derive (point) estimators for quantities of interest. The most prevalent is maximum likelihood estimation, where we wish to optimize the likelihood of an observed sample. In order to do this, we derive $L(\theta)$ based on a distributional assumption, we take the logarithm to make it easier to work with ($\ell(\theta)$), then differentiate giving $S(\theta)$. We then solve $S(\widehat{\theta}) = 0$. If we had assumed our density was $f(y;\theta)$, then we can make these quantities specific, notably

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) \\ \ell(\theta) &= \sum_{i=1}^{n} \log f(y_i;\theta) \\ S(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(y_i;\theta). \end{aligned}$$

We know that $E[S(\theta)] = 0$. Moreover, we know that, asymptotically, the $\widehat{\theta}$ which solves $S(\widehat{\theta}) = 0$ is such that (as $n \to \infty$),

$$\widehat{\theta} \dot{\sim} N\left(\theta, I(\theta)^{-1}\right),$$

where $I(\theta) = E[-S'(\theta)]$ is the Fisher Information matrix.

In certain settings, in place of this likelihood derivation, we would have also seen *least squares* estimators (this is particularly relevant, for instance, in regression!). The idea here is that, if $\widehat{\theta}$ is going to be a "good" estimator for θ, then it presumably should be near θ, generally. If we use $\mathcal{L}(\widehat{\theta})$ to represent a **loss** function, which measures the distance between our estimate and the true value, then estimators which have lower expected loss are preferable. Frequently, we would take $\mathcal{L}(\widehat{\theta}) = (\theta - \widehat{\theta})^2$, which we call the **squared** loss function, or **squared error**. If we find the estimator $\widehat{\theta}$ such that $E[\mathcal{L}(\widehat{\theta})]$ is minimized, then we call this estimator the **least squares estimator** (since it minimizes the squared error function!). Now, in practice we are often going to want to minimize these values in a particular sample! For instance, if we consider β in the linear regression context, then we are suggesting that $E[Y_i|X_i] = X_i'\beta$, and the least squares estimator is going to minimize the MSE empirically. That is

$$\begin{aligned} \mathcal{L}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 \\ \mathcal{L}'(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}(Y_i - X_i'\beta)^2, \end{aligned}$$

and then our estimators solves $\mathcal{L}'(\widehat{\beta}) = 0$.

Generally, if $\widehat{\theta}$ is a least squares estimator, then we would find that it is consistent for θ (that is, it converges in probability) and also, as $n \to \infty$

$$\widehat{\theta} \dot{\sim} N(\theta, \text{var}(\widehat{\theta})).$$

In the event of linear regression we find that the likelihood estimator for β (assuming normality) and the least squares estimator (where no distribution is assumed) are **exactly the same**. This is a curious result: one process made a strict distributional assumption, and using properties of that distribution gets to the estimator and related properties; the other makes no such assumptions, but gets to the same point. Upon further investigation, there is a core similarity between these two processes: they are both optimization procedures.

For likelihood theory we want to maximize $L(\theta)$ and for least squares estimation we want to minimize $\mathcal{L}(\theta)$, which is equivalent to the maximization of $-\mathcal{L}(\theta)$. In both cases, these functions take the form $\sum_{i=1}^{n} \rho_i(Y_i, \theta)$ for a particular ρ_i, so that when we use our standard calculus results to optimize them, our estimator $\widehat{\theta}$ ends up solving

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \rho_i(Y_i, \theta) = 0.$$

It turns out that estimators of this form come up **incredibly often** and are incredibly useful!

Definition: An estimator, $\widehat{\theta}$ is said to be an **M-estimator**, if $\widehat{\theta}$ is the solution to $U(\widehat{\theta}) = 0$, for some $U(\theta)$ which takes the form

$$U(\theta) = \sum_{i=1}^{n} \Psi(\theta; Y_i).$$

We call $U(\theta)$ an **estimating function**, and the equations specified by $U(\theta) = 0$ are said to be **estimating equations**. We call $\widehat{\theta}$ an M-estimator where the M stands for *maximum*. Both likelihood estimators **and** least squares estimators are specific kinds of M-estimators.

Key Theoretical Results

That was a lot of words to say that an M-estimator is an estimator which solves $U(\theta) = 0$, where $U(\theta)$ is just a summation over an IID sample. Why do we care at all? Suppose that θ_0 is a value for θ such that $E[U(\theta_0)] = 0$. Then, for free[1] we get that

1. $\widehat{\theta}$ is a consistent estimator for θ_0.
2. $\widehat{\theta}$ has an asymptotic distribution (as $n \to \infty$) given by

$$\widehat{\theta} \dot{\sim} N(\theta_0, \mathbf{A}(\theta_0)^{-1}\mathbf{B}(\theta_0)\mathbf{A}(\theta_0)^{-1'}),$$

where $\mathbf{A}(\theta_0) = E\left[-\frac{\partial}{\partial \theta} U(\theta)\right]$ evaluated at $\theta = \theta_0$ and $\mathbf{B}(\theta_0) = E[U(\theta_0)U(\theta_0)']$.

[1] In actual fact, we require a few more "regularity" assumptions on $U(\theta)$. These are going to hold for most situations that come up in your STAT courses, and are more a technical requirement than an informative point, so you can feel free to ignore them!

Note that we have not made any assumption about where $U(\theta)$ is coming from. We have not made any assumption about the distributions we are working with. We have not made any assumption except for the fact that, when we take the expected value of our function, $U(\theta)$ at $\theta = \theta_0$, we get 0! When we have such a $U(\cdot)$, not only do we have a consistent estimator for θ_0, but we have one which is asymptotically normal, meaning that we can proceed with all of the standard inference tools that we are used to! **What's more** the asymptotic distribution of our estimator only relies on the form of $U(\theta)$ that we have.

We refer to the variance matrix as a **sandwich variance estimator** since (if you're really hungry) you could view $\mathbf{A}(\theta_0)^{-1}$ as two slices of bread, and $\mathbf{B}(\theta_0)$ as the contents of the sandwich[2]. Now, the nice part of these two matrices is that (often) we are able to reliably estimate them from the data at hand. The reason is that, since they are both simple expectations, we can take

$$\widehat{\mathbf{A}}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}U(\theta, Y_i)\bigg|_{\theta=\widehat{\theta}}$$
$$\widehat{\mathbf{B}}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}U(\widehat{\theta}, Y_i)U(\widehat{\theta}, Y_i)'.$$

This provides us with valid asymptotic inference (so long as n is sufficiently large!).

The Score Equation as an M-Estimator

To re-emphasize these ideas, let's consider an example that we know well: the Score function! We saw above that

$$S(\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(y_i;\theta) = \sum_{i=1}^{n}\frac{f'(y_i;\theta)}{f(y_i;\theta)}.$$

Now, we also know (and have used in our discussion of GLMs) the fact that, if $\theta = \theta_0$, then $E[S(\theta_0)] = 0$, and so we know that the Score function is unbiased! As a result, the estimator $\widehat{\theta}$ which solves $S(\widehat{\theta}) = 0$ will be such that

1. $\widehat{\theta}$ is consistent for the true value θ_0.
2. $\widehat{\theta}$ is asymptotically normal. The relevant matrices are

$$\mathbf{A}(\theta_0) = E\left[-\frac{\partial}{\partial\theta}S(\theta)\bigg|_{\theta=\theta_0}\right] = I(\theta_0)$$
$$\mathbf{B}(\theta_0) = E\left[S(\theta_0)S(\theta_0)'\right] = \text{var}(S(\theta_0)) = I(\theta_0).$$

   Here, we used the property (which we also used during the GLM review) that the variance of the Score function is the expected information. As a result, we have that $\mathbf{A}(\theta_0) = \mathbf{B}(\theta_0)$ and so the asymptotic variance is simply going to be $\mathbf{A}(\theta_0)^{-1} = I(\theta_0)^{-1'}$, which is **exactly** the result stated above!

[2]No one ever said statisticians were *good* at naming things.

More Examples

Of course, we already knew the properties of likelihood and least squares estimators and so M-estimation is not the most useful for these situations. The use case is more that M-estimation unifies these related estimation procedures, and brings along with it *anything* that can be framed in these terms!

Sample Means

Consider the estimation of the mean of some distribution, Y, which we denote with θ_0. We know that we want some function $\Psi(\theta, Y)$, such that $E[\Psi(\theta_0, Y)] = 0$. Consider what happens if we simply take $\Psi(\theta, Y) = Y - \theta$. Well clearly, for the true θ_0 we would find

$$E[\Psi(\theta_0, Y)] = E[Y] - \theta_0 = \theta_0 - \theta_0 = 0.$$

As a result, if we take

$$U(\theta) = \sum_{i=1}^{n} Y_i - \theta,$$

then this defines an estimator for the mean of Y_i which we know will be consistent and asymptotically normal! In this case, we can actually solve this for $\widehat{\theta}$ in closed form,

$$\begin{aligned} U(\widehat{\theta}) = 0 &\iff \sum_{i=1}^{n} Y_i - \widehat{\theta} = 0 \\ &\iff n\widehat{\theta} = \sum_{i=1}^{n} Y_i \\ &\iff \widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \end{aligned}$$

This is just the sample mean, $\overline{Y}$! Recall that we *know* from the WLLN and the CLT that $\overline{X}$ is consistent for $E[X]$ and that $\overline{X} \dot{\sim} N(\theta_0, n^{-1}\operatorname{var}(Y))$. We could also have gotten that from this M-estimation approach!

$$\begin{aligned} \mathbf{A}(\theta_0) &= E\left[-\frac{\partial}{\partial\theta}U(\theta)\Big|_{\theta=\theta_0}\right] = n \\ \mathbf{B}(\theta_0) &= E\left[U(\theta_0)U(\theta_0)'\right] \\ &= \operatorname{var}\left[\sum_{i=1}^{n} Y_i - \theta_0\right] \\ &= \sum_{i=1}^{n} \operatorname{var}(Y_i - \theta_0) \\ &= n\operatorname{var}(Y) \\ \implies \mathbf{A}(\theta_0)^{-1}\mathbf{B}(\theta_0)\mathbf{A}(\theta_0)^{-1'} &= n^{-1}\operatorname{var}(Y). \end{aligned}$$

Joint Mean and Variance

Consider if instead of an estimator for the mean, we want to joint estimate the mean *and* the variance. Consider the estimating equation given by

$$U(\theta) = \sum_{i=1}^{n} \begin{pmatrix} Y_i - \theta_1 \\ (Y_i - \theta_1)^2 - \theta_2 \end{pmatrix}.$$

Here we are taking $\theta = (\theta_1, \theta_2)$ to be a 2D vector of the parameters. Now, if Y_i is such that $E[Y_i] = \mu$ and $\text{var}(Y_i) = \sigma^2$, then consider $E[U(\mu, \sigma^2)]$.

$$\begin{aligned} E[U(\mu, \sigma^2)] = \sum_{i=1}^{n} E\left\{ \begin{pmatrix} Y_i - \mu \\ (Y_i - \mu)^2 - \sigma^2 \end{pmatrix} \right\} &= \begin{pmatrix} E[Y_i] - \mu \\ E[(Y_i - \mu)^2] - \sigma^2 \end{pmatrix} \\ &= \begin{pmatrix} \mu - \mu \\ \text{var}(Y) - \sigma^2 \end{pmatrix} = \mathbf{0}. \end{aligned}$$

You could of course still work out the $\mathbf{A}(\theta_0)$ and $\mathbf{B}(\theta_0)$, but I will leave that as an exercise if you care: for now, we can say that because this is an unbiased estimating equation, the resulting estimator will be consistent for (μ, σ^2).

Quasi-Likelihood Estimators

When we were considering quasi-likelihood estimators for β, we defined them such that $\widehat{\beta}$ solves

$$U(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{Y_i - \mu_i}{\phi V(\mu_i)} = 0.$$

We claimed that as long as μ_i was correctly specified then this process results in consistent estimators for β (even when $V(\mu_i)$ is incorrect). We also claimed that the variance of our estimator can be estimated consistently even when $V(\mu_i)$ is incorrect. The reason is (you guessed it), QMLEs are M-estimators! We can see that $E[U(\beta)] = 0$ is true, so long as μ_i is correctly specified, since the only random component is the Y_i and $E[Y_i] - \mu_i = 0$. Using the sandwich variance estimators give us a technique for estimating the variance of our estimator, regardless of the true underlying distribution, which is what makes M-estimation so powerful!

Conclusions

Through the remainder of the course we will see M-estimators come up several times. Throughout your other statistics courses (and particularly, if you continue into graduate school) you will (likely) see several other estimators which are M-estimators. Despite their prevalence, it is not often that (at the undergraduate level) M-estimation is presented as a general theory. I think that this does a disservice as it forces you to conceptualize several different estimators (for instance, likelihood and least squares estimators) as distinct, whereas they share the same underlying theory. There is also a special connection to the

University of Waterloo (look up VP Godambe, if you're interested!). Still, M-estimation is quite theoretical, and even though I have done my best to simplify the results, this note may still seem confusing **and that's okay**. In order to be successful in this course, all you need to know is that M-estimators are a special type of estimator, which, based on some underlying theory (which we did not cover) have very desirable properties!