Missing data lecture 4: Asymptotics of MLEs and Some Bayes

MLEs in repeated measure models

Last class we talked about MLEs for a simple repeated measure model:

$$\begin{aligned} y_i \mid X_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \Sigma) \\ \epsilon_i &\perp\!\!\!\perp \epsilon_j \forall i \neq j \end{aligned}$$

Let $y = (y_1^T, y_2^T, \dots, y_n^T)^T$ and let $X = (X_1^T, X_2^T, \dots, X_n^T)^T$, and let $\epsilon = (\epsilon_1^T, \epsilon_2^T, \dots, \epsilon_n^T)^T$. Then the model can be written:

$$\begin{aligned} y \mid X &= X\beta + \epsilon \\ \epsilon &\sim \text{Normal}(0, I_n \otimes \Sigma) \end{aligned}$$

We showed that if Σ were known, we could write the MLE for β as:

$$\hat{\beta} = (\textstyle\sum_i X_i^T \Sigma^{-1} X)^{-1} (\sum_i X_i \Sigma^{-1} y_i)$$

We can also show that the MLE for Σ if β were known is:

$$\Sigma = \frac{1}{n} \sum_i (y_i - X_i\beta)(y_i - X_i)^T$$

Combining these two fact together, we can iteratively maximize the MLE by doing:

$$\begin{aligned} \Sigma^{(t+1)} &= \frac{1}{n} \sum_i (y_i - X_i\beta^{(t)})(y_i - X_i\beta)^T \\ \beta^{(t+1)} &= (\textstyle\sum_i X_i^T (\Sigma^{(t)})^{-1} X_i)^{-1} (\sum_i X_i (\Sigma^{(t)})^{-1} y_i) \end{aligned}$$

Which is similar to what the book has.

Inference for MLEs

We've shown that we can find a point in parameter space $\hat{\theta}$ such that there is evidence against any other point $\theta \neq \hat{\theta}$ being the parameter that generated the data under an assumed statistical model $f_Y(y_i \mid \theta)$.

How do assess the uncertainty in this point estimate? One way to think about this is to consider the distribution of MLEs under different hypothetical datasets.

If we can characterize the distribution, we can build a confidence interval for θ that will contain the true value of θ for some prescribed proportion of our hypothetical datasets.

Consider the model $y_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, where σ^2 is known one-dimensional normal model. Then we know that $\hat{\mu} = \bar{y}$, and that its distribution is $\bar{y} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$. Using this distribution, we can construct a statistic whose distribution doesn't depend on any unknown parameters. This is also known as a *pivotal quantity*.

$$\sqrt{n}(\bar{y} - \mu)/\sigma \sim \text{Normal}(0, 1)$$

We know the cumulative distribution function for $N(0, 1)$, $\Phi(x)$ and we can use this distribution to build a confidence interval for μ:

$$P(z_{\alpha/2} < \sqrt{n}(\bar{y} - \mu)/\sigma < z_{1-\alpha/2})$$

where $z_p = \Phi^{-1}(p)$ and α is usually 0.05. Now we solve the system of inequalities for μ to get our $1 - \alpha$ confidence interval:

$$\begin{aligned} P(z_{\alpha/2} < \sqrt{n}(\bar{y} - \mu)/\sigma < z_{1-\alpha/2}) &= P(\frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \bar{y} - \mu < \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}) \\ &= P(\frac{\sigma}{\sqrt{n}} z_{\alpha/2} - \bar{y} < -\mu < -\bar{y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}) \\ &= P(\bar{y} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} > \mu > \bar{y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}) \end{aligned}$$

Because the distribution is symmetric, $z_{\alpha/2} = -z_{1-\alpha/2}$ which gives us:

$$P(\bar{y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < \mu < \bar{y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2})$$

Then the interval $(\bar{y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2})$ will contain μ in 0.95 of datasets generated under the assumed $N(\mu, \sigma^2)$.

For all but the simplest models, the sampling distribution is intractable, but we can approximate the distribution using asymptotics.

We'll need the multivariate central limit theorem, which we'll just take as a given:

Theorem 1. Multivariate CLT

Suppose that X_i are random vectors in R^d with common mean $\mathbb{E}[X_i] = \mu$ and covariance $\Sigma = \mathbb{E}\left[(X_i - \mu)(X_i - \mu)^T\right]$. Let $\bar{X}$ be the sample average of the X_i with $\bar{X}_j = \sum_i X_{ij}/n$. Then as $n \to \infty$:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \text{Normal}(0, \Sigma)$$

We can use this idea to get a pivotal quantity that involves the MLE for θ and the asymptotic variance covariance matrix of the MLE. We'll start with some key assumptions:

1. The MLE is consistent for θ, which means that as we collect more samples the MLE with converge in probability to θ.

This will rule out the Neyman-Scott problem.

2. y_i are iid with density $f_Y(y_i \mid \theta)$, $\theta \subseteq \mathbb{R}^d$
3. The support of the random variable y_i doesn't depend on θ.

In our earlier presentation of how things can go wrong with MLE, having support that depends on the value of the parameter can make things go awry, so we'll assume that we're not in that scenario (like the $y_i \sim \text{Uniform}(0, \theta)$).

4. The true parameter θ is in the interior of the parameter space. This ensures that the gradient of the likelihood value at the maximizer $\widehat{\theta}$ will be zero.

There are some more conditions on the gradient and hessian of the log-likelihood, but we'll assume that those are satisfied.

The gradient of the log-likelihood evaluated at the MLE $\hat{\theta}$ can be expanded around the true parameter value $\theta^\dagger$:

$$(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\hat{\theta}} = \nabla_\theta \ell_Y(\theta \mid y) \mid_{\theta=\theta^\dagger} + \nabla^2_\theta \ell_Y(\theta \mid y) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)} (\hat{\theta} + \theta)$$

where $\tilde{\theta}_j$ lies on the cord between $\hat{\theta}$ and $\theta^\dagger$ and may differ by the index j of the vector $\nabla_\theta \ell_Y(\theta \mid y) \mid_{\theta=\hat{\theta}}$.

To be precise, we have used the mean-value theorem on each coordinate j of the gradient of the log-likelihood with respect to θ evaluated at $\widehat{\theta}$: $\nabla_\theta \ell_Y(\theta \mid y) \mid_{\theta=\hat{\theta}}$:

$$\left.\frac{\partial \ell_Y(\theta \mid y)}{\partial \theta_j}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial \ell_Y(\theta \mid y)}{\partial \theta_j}\right|_{\theta=\theta^\dagger} + \nabla_\theta \left.\frac{\partial \ell_Y(\theta \mid y)}{\partial \theta_j}\right|_{\theta=\tilde{\theta}_j} (\hat{\theta} - \theta)$$

and then constructed the matrix $\nabla^2_\theta \ell_Y(\theta \mid y) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)}$ so that row j of this matrix is the vector $\nabla_\theta \left.\frac{\partial \ell_Y(\theta|y)}{\partial \theta_j}\right|_{\theta=\tilde{\theta}_j}$.

Multiply both sides by $n^{-1/2}$:

$$n^{-1/2}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\hat{\theta}} = n^{-1/2}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\theta^\dagger} + n^{-1/2}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)} (\hat{\theta}+\theta)$$

Note that $\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\hat{\theta}} = 0$:

$$0 = n^{-1/2}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\theta^\dagger} + \frac{\sqrt{n}}{n}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)} (\hat{\theta}+\theta)$$

$$0 = n^{-1/2}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\theta^\dagger} + \frac{1}{n}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)} \sqrt{n}(\hat{\theta}+\theta)$$

Solving for $\sqrt{n}(\hat{\theta}-\theta)$:

$$\sqrt{n}(\hat{\theta}+\theta) = \left(-\frac{1}{n}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)}\right)^{-1} \frac{\sqrt{n}}{n}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\theta^\dagger}$$

Now that we have an expression for $\sqrt{n}(\hat{\theta}-\theta)$, if we can derive an asymptotic distribution for the right-hand side, we'll have some hope of generating confidence sets for θ. Because the data are iid we can write the second term on the left-hand side as:

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(\nabla_\theta \ell_Y(\theta \mid y_i)) \mid_{\theta=\theta^\dagger}$$

This is just the average of the gradient of the log-likelihood scaled by $\sqrt{n}$. It turns out that $\mathbb{E}\left[(\nabla_\theta \ell_Y(\theta \mid y_i)) \mid_{\theta=\theta^\dagger}\right] = 0$, under $y_i \sim f_Y(y_i \mid \theta)$, so

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(\nabla_\theta \ell_Y(\theta \mid y_i)) \mid_{\theta=\theta^\dagger} \xrightarrow{d} \text{Normal}(0, \mathbb{E}\left[(\nabla_\theta \ell_Y(\theta \mid y_i)) \mid_{\theta=\theta^\dagger} (\nabla_\theta \ell_Y(\theta \mid y_i)) \mid^T_{\theta=\theta^\dagger}\right])$$

That expectation is a $d \times d$ matrix called the Fisher information of $f_Y(y_i \mid \theta)$: $\mathcal{I}(\theta^\dagger)$. It also turns out that, given conditions on the Hessian of the log-likelihood, $\left(-\frac{1}{n}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)}\right)^{-1}$ converges to probability to $\mathcal{I}(\theta^\dagger)^{-1}$, or the inverse of the Fisher information matrix.

If we have $X \sim \text{Normal}(0, C)$, and we left-multiply X by a matrix A, we'll get: $AX \sim \text{Normal}(0, ACA^T)$.

This is because any linear combination of normal random variables is again normal, and the normal distribution is a function of the mean and covariance only. Expectation is linear:

$$\mathbb{E}\left[AX\right] = A\mathbb{E}\left[X\right] = 0$$

and covariance is $\mathbb{E}\left[(X-\mu)(X-\mu)^T\right] = \mathbb{E}\left[(XX^T\right] = C$, so

$$\mathbb{E}\left[(AX)(AX)^T\right] = A\mathbb{E}\left[XX^T\right]A^T = ACA^T$$

We'll also need the fact that for $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{d} Y$ that $X_n Y_n \xrightarrow{d} XY$.

Putting all this together gives:

$$\left(-\frac{1}{n}(\nabla^2_\theta \ell_Y(\theta \mid y)) \mid_{\theta=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)}\right)^{-1} \frac{\sqrt{n}}{n}(\nabla_\theta \ell_Y(\theta \mid y)) \mid_{\theta=\theta^\dagger} \xrightarrow{d} \text{Normal}(0, \mathcal{I}(\theta^\dagger)^{-1}\mathcal{I}(\theta^\dagger)\mathcal{I}(\theta^\dagger)^{-1})$$

the righthand side of which simplifies to:

$$\text{Normal}(0, \mathcal{I}(\theta^\dagger)^{-1})$$

Finally, putting everything together, we have the

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \text{Normal}(0, \mathcal{I}(\theta^\dagger)^{-1})$$

Let's say we want to get a confidence interval for a single parameter θ_1. Then we can build an asymptotic confidence interval using the pivotal quantity strategy we had above:

$$P(\bar{y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < \mu < \bar{y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2})$$

But with $\bar{y}$ equaling $\hat{\theta}_1$ and $\sigma = \sqrt{\mathcal{I}(\hat{\theta})^{-1}_{1,1}}$. It is sometimes hard to calculate the Fisher information because it involves taking expectations of the negative Hessian of the log-likelihood function. If we'd prefer, we can instead use an estimator for $\mathcal{I}(\hat{\theta})$ as

$$\mathcal{I}(\hat{\theta}) = -\frac{1}{n}\sum_i \nabla^2_\theta \ell_Y(\theta \mid y_i)_{\theta=\hat{\theta}}$$

The $\mathcal{I}$ should have a hat over it, but I can't get the math to compile correctly when I add the `\hat` over it.

There are ways to build multivariate confidence intervals, but I won't go over those right now, though they are covered in the book in Chapter 6.

Bayes

The machinery for Frequentist inference often relies on asymptotic arguments for complex models. Bayesian inference, on the other hand, does not, and gives exact finite sample inference. There are caveats though, which we'll cover.

MLEs are concerned with finding a single point that, in some sense agrees with the dataset at hand, though our inference depends on hypothetical replications of the experiment that would generate alternative datasets. Bayesian inference requires that we characterize the distribution of parameters that agree with the dataset at hand.

The idea of Bayesian inference starts with Bayes rule. Given a prior distribution $p(\theta)$ we can combine that with the observational density of data $f_Y(y \mid \theta)$ to give us an updated distribution $p(\theta \mid y)$ given the dataset at hand:

$$p(\theta \mid y) = \frac{f_Y(y \mid \theta)p(\theta)}{\int_{\Omega_\theta} f_Y(y \mid \theta)p(\theta)d\theta}$$

The nice thing about Bayesian inference is that we get a distribution over θ given the dataset that we can now use to make probability statements about θ. The statement With 0.95 probability θ is in the interval C is just a manipulation of the posterior density: $P(\theta \in C \mid y) = \int_C p(\theta \mid y)d\theta$.

Let's look at a specific example: the standard $y_i \overset{\text{iid}}{\sim}$ Bernuolli(θ) with $\theta \sim$ Beta(α, β).

The likelihood is:

$$\prod_i \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}$$

The prior for θ will be:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The numerator the posterior is the product of these two expressions, where we let $s = \sum_i y_i$ for convenience:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}$$

Integrating over θ will give us the denominator of our expression:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+s)\Gamma(\beta+n-s)}{\Gamma(\alpha+\beta+n)}$$

The ratio of the numerator and the denominator gives us:

$$\frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+s)\Gamma(\beta+n-s)}{\Gamma(\alpha+\beta+n)}}$$

which simplifies to:

$$\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+n-s)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}$$

This is just the Beta distribution with updated coefficients.

The prior mean is $\frac{\alpha}{\alpha+\beta}$, while the posterior mean is $\frac{s+\alpha}{n+\alpha+\beta}$. We can rewrite this to get a better understanding of the posterior mean represents in this circumstance:

$$\frac{s+\alpha}{n+\alpha+\beta} = \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{n+\alpha+\beta} + \left(1 - \frac{\alpha+\beta}{n+\alpha+\beta}\right)\frac{s}{n}$$

This shows that the posterior mean is a weighted average of the prior mean and the data mean. This sort of exemplifies what we would hope for from Bayesian inference, some adjudication between the prior and the data. The posterior distribution is a Beta distribution, so we can get probability statements easily by using `qbeta` in R.

We didn't have to go through all of the marginalization above. We could have noticed that the *kernel* of the posterior, namely the expression that depends on the unknown parmaeters, had a familiar form:

$$p(\theta \mid s) \propto \theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}$$

This is the kernel of the beta distribution, so we could have just stopped here and said that

$$p(\theta \mid s) \equiv \text{Beta}(\alpha+s, \beta+n-s)$$

This procedure is aided by conjugate priors, which match the likelihood in a way; the functional form of the prior slots into the way the parameters are expressed in the likelihood to yield a family of posteriors that are in the same family as the prior.

These probabilities are "right" under our prior assumption. It may not be true for alternative realizations of the data, or for alternative draws from the prior if the prior that generated the data does not match $p(\theta)$.

There is another downside to using a prior. MLEs have a nice property called invariance, namely that if $\hat{\theta}$ is the MLE then the MLE for a function of θ, say $g(\theta)$, is just $g(\hat{\theta})$. This isn't true in Bayesian inference generally, because we're now dealing with distributions rather than points. So usually the posterior for θ won't be the same as the posterior for $g(\theta)$: Let $\eta = g(\theta)$, and assume for simplicity's sake that g is one-to-one. Then $\theta = g^{-1}(\eta)$. If θ has posterior $p(\theta \mid y)$, the posterior for $g(\theta)$ is:

$$p(g^{-1}(\eta) \mid y) \det \nabla_\eta g^{-1}(\eta)$$

In fact, this is true of priors too! Given $p(\theta)$ and a transformation $\eta = g(\theta)$ the prior for η is:

$$p(\eta) = p(g^{-1}(\eta)) \det \nabla_\eta g^{-1}(\eta)$$

This is somewhat problematic if you think about how to represent ignorance. Let's say you want to learn about a parameter $\theta \in [0,1]$. The simplest prior for this is the flat prior $p(\theta) \propto 1$. That implies that θ is equally likely to be anywhere in the interval of $[0,1]$. But what does that imply about $\eta = \theta^2$? Well on $[0,1]$, $g(x) = x^2$ is one-to-one, and the inverse is $\sqrt{g(\eta)} = \theta$. The derivative of $\sqrt{\eta}$ is proportional to $\eta^{-1/2}$, which implies a downward sloping distribution on $[0,1]$. But why would you have knowledge of θ^2 without knowledge of θ? It seems counterintuitive.

The dyed-in-the-wool Bayesian would argue that there is no such thing as true ignorance, and that the problem that you face will have consequences for where you expect your parameter to lie. Suppose you're modeling the proportion of Corvallis residents with synovial sarcoma, which is a very rare cancer. You're probably going to use a prior that favors small values of θ.

What if instead you're looking at the proportion of rainy days in Corvallis from November to April? You'd probably use a prior that at the very least avoided θ near 0.