MLE (cont. from Lecture 03)

Censored data

- Same numbers as last lecture
  - 30 devices tested
  - 21 devices failed before 7 months
  - 9 have survived longer than 7 months; this is the censored data
- Recall the MLE: $\hat{\lambda}_{MLE} = \dfrac{21}{21\bar{t} + 63} = \dfrac{1}{\bar{t} + 3}$
  - $\bar{t}$ is the sample mean of the 21 observed failure times
- What's the MLE of μ?
  - By properties of the Exponential, $\mu = \lambda^{-1}$
  - Is it true that $\hat{\mu}_{MLE} = \hat{\lambda}_{MLE}^{-1} = \bar{t} + 3$?
- Alternate estimators
  - $\hat{\mu} = \bar{t}$ would be naïve and an underestimate because we're throwing out our censored data
  - Replacing censored data as 7
    - This would be $\frac{\sum_i t_i + 63}{30}$ (where i indexes the observed values)
  - Instead we have $\frac{\sum_i t_i + 63}{21}$, which makes the estimate larger
    - Intuition: we're accounting for not observing the censored data
    - So our numerator is the total time, including the cutoff, and the denominator is the count of observed data
- Property: **Invariance of the MLE**
  - So it is true $\hat{\mu}_{MLE} = \hat{\lambda}_{MLE}^{-1}$
- Contrast: Bayesian analysis
  - Let $\hat{\lambda}_{\text{Bayes}} = \mathbb{E}[\lambda \mid \text{data}]$
    - $\hat{\mu}_{\text{Bayes}} = \mathbb{E}[\mu \mid \text{data}] = \mathbb{E}[\lambda^{-1} \mid \text{data}]$
    - But wait! **Jensen's inequality** strikes again!
    - $\hat{\mu}_{\text{Bayes}} = \mathbb{E}[\lambda^{-1} \mid \text{data}] > \frac{1}{\hat{\lambda}_{\text{Bayes}}}$

Invariance of the MLE

Thm (Invariance under reparameterization). Let $\psi = g(\theta)$, where g is invertible (one-to-one). Then, $L(\psi; y) = L(\theta; y)$.

Proof: $L(\psi; y) = f(y; \psi) = f(y; \theta) = L(\theta; y)$. When evaluating individual values, we have that $L(\psi; y) = L(g(\theta); y), L(g^{-1}(\psi); y) = L(\theta; y)$.

Corollary: The MLE is also invariant. $\hat{\psi} = g(\hat{\theta})$.

- What if g is not invertible?
  - Then we have problems with the model, which is bad
- But "invariance is so nice we make it hold by definition even when g is *not* invertible" (Joe's words, not mine)

Examples

Ex (MLE of normal parameters). We have $\theta = (\mu, \sigma^2)$. We then have $(\hat{\mu}, \hat{\sigma}^2) = (\bar{Y}, \frac{1}{n} \sum_i (Y_i - \bar{Y})^2)$. We can use invariance to get $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.

- Convenient property for TeXing
  - $\hat{\sigma}^2 = \widehat{\sigma^2}$, so we don't need to be too picky about examining the height of our hats

Ex (Logit function). $Y \sim \mathrm{Bin}(n, p), \theta = \mathrm{logit}(p) = \log \frac{p}{1-p}$ (log-odds).

We have $L(p) = p^y (1-p)^{n-y}$. We have log-likelihood $\ell(p) = y \log p + (n-1) \log(1-p)$. The first derivative w.r.t. p is $\ell'(p) = \frac{y}{p} - \frac{n-y}{1-p}$. Setting this equal to zero and solving for p gives us $\hat{p} = \frac{Y}{n}$ (make sure to put a nice hat on it) (capital Y is estimator, lowercase y is estimate).

By invariance, $\hat{\theta} = \mathrm{logit}(\hat{p})$.

- But this estimator *sucks*. Why?
  - It is undefined
  - With positive probability, $Y = n$

Ex (Bad unbiased estimators). $Y \sim \mathrm{Pois}(\lambda)$. Our estimand is $\theta = e^{-3\lambda}$.

Consider the estimator $\hat{\theta} = (-2)^y$. Claim: this is unbiased.

$$\mathbb{E}[\hat{\theta}] = \left(\lambda \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \right) = e^{-\lambda} e^{-2\lambda} = \theta.$$

This is the only unbiased estimator! But it sucks because we'll regularly estimate invalid values (i.e., outside the range $[0, 1]$).

We also have that $\hat{\lambda}_{MLE} = Y$. By invariance, $\hat{\theta}_{MLE} = e^{-3Y}$. This is biased but better.

Method of moments (MoM)

- More of a principle than an estimator
  - Not unique, unlike the MLE
  - The MLE is a uniquely specified procedure (although there may be more than one peak)

Procedure

1. Write estimand in terms of theoretical moments
2. Replace estimand by estimator, replace theoretical moments w/ sample moments

3. Solve for the estimator

Examples

Ex (MoM of Poisson). Let $Y_1, \dots, Y_n \overset{i.i.d.}{\sim} \mathrm{Pois}(\theta)$. Find a method of moments estimator.

Using the first moment. We have $\theta = \mathbb{E}[Y_1]$. We then have $\hat{\theta}_{MoM} = \bar{Y}$. Note this is the same as the MLE. This estimator is unbiased.

Using the second moment. We can write the variance $\theta = \mathbb{E}(Y_1^2) - \mathbb{E}Y_1^2$. We then have $\hat{\theta}_{MoM2} = \frac{1}{n}\sum_{j=1}^n Y_j^2 - \bar{Y}^2$. This estimator is biased.

Which is better in terms of MSE? We can solve this with theoretical calculations, simulation, or asymptotics. The first is "better" in terms of MSE.

Ex (MoM of paired data). Let (X_j, Y_j) be i.i.d. pairs (independence across pairs, not within pairs) and let $j = 1, \dots, n$. Let us define an estimand

$$\beta = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y}{\mathbb{E}(X^2) - \mathbb{E}X^2}.$$

(This is an important estimand in linear regression; more discussion is available on Ed. We will also discuss this later in the course).

We can then make a MoM estimator

$$\hat{\beta}_{MoM} = \frac{\frac{1}{n}\sum_{j=1}^n X_j Y_j - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{j=1}^n X_j^2 - \bar{X}^2}.$$

Properties of MoM

- Requires very few assumptions
- Fairly simple procedure
  - Though you might need some arithmetic to solve for estimator
  - E.g., system of equations

MoM vs MLE

- Can be the same
- When they differ, MLE tends to have nice properties
  - Especially in asymptotics
- MoM tends to be easy to calculate
  - MLE may not have closed-form solution
    - Though iterative methods exist
    - E.g., Newton-Raphson (Newton's), expectation maximization (EM) algorithm
      - EM may not be covered in this class
      - EM was created in Harvard Stats department

- MoM can often be used to initialize Newton's method