36-463/663: Multilevel & Hierarchical Models

Review of Estimation

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Preview

- Discussion/example of multilevel glm (generalized mixed-effects models, Hierarchical glm…)
- We will start moving into Bayesian interpretations of multilevel models
  - The recommended text by Lynch will be useful.
  - You can buy it online
  - I will make selected chapter available for reading
    - Chapters 2 and 3 online now.
      - Review of probability theory and MLE, and intro to Bayes
    - Chapters 4 and 9 a little later on.
      - MCMC and Bayesian multi-level modeling
  - Gelman & Hill, Chapter 18, give a more brisk introduction, oriented toward hierarchical / multilevel modeling

Discussion/Example of Multilevel GLMs

- Blackboard (the one in the classroom!)
- Hosp.r / hosp.txt

Review of Estimation…

- Parameter, Statistic, Point Estimator, Point Estimate
- How Good is the Estimator?
- Uncertainty & Standard Error
- Methods of Estimation
- Maximum Likelihood Estimators (MLE’s)
- Standard Error for MLE’s

Parameter, Statistic, Point Estimator, Point Estimate

- Let $X_1, \ldots, X_n$ be a sample (usually iid) from $f_X(x)$.
- A ***parameter*** is a free variable that characterizes $f_X(x)$, e.g:
  - The mean μ or variance σ^2 of a Normal r.v.
  - The mean λ of a Poisson r.v.
  - The lower and upper interval endpoints, A and B, of a Unif(A,B) r.v.
- A ***statistic*** is any quantity that can be calculated from a sample, e.g.:
  - The sample average $\bar{X}$ or the sample variance S^2
  - The minimum or maximum of all the X's
- Often we devise (and calculate) a statistic to estimate a parameter.

- A ***point estimate*** for θ is a single number that can be regarded as a reasonable value for θ
- A ***point estimator*** for θ is a statistic that gives the formula for computing the point estimate for θ
- **Example**: Randomly sample 5 people on CMU campus and ask: married? Let X = # of yes's. Now do the survey n times and let $X_1, \ldots, X_n$ be the number who say they are married, in each sample of 5.
  - $X_1, \ldots, X_n$ are an iid sample from a Binomial(5,p) distribution
  - A ***parameter*** of interest is p = P[person on campus is married]
  - A reasonable ***point estimator*** for p is $\frac{1}{3}(X_1/5 + X_2/5 + X_3/5) = \frac{1}{3}\sum_{i=1}^{3} X_i/5$
  - If we do this survey n=3 times and we get $X_1=2$, $x_2=4$, $x_3=0$, then the ***point estimate*** is

$$\frac{1}{3}\sum_{i=1}^{3} x_i/5 = \frac{1}{3}(2/5 + 4/5 + 0/5) = 0.40$$

How Good is the Estimator?

- Let θ be a parameter of $f_X(x)$, and let T be an estimator of θ, from a sample of size n. Some criteria for a "good" estimator are:
  - *Consistency*: T is *consistent* for θ if

$$\lim_{n\to\infty} P[|T-\theta| > \epsilon] = 0, \text{ for all } \epsilon > 0$$

  - *Unbiasedness*: T is *unbiased* for θ if

$$E[T] = \theta$$

  - *Efficiency*: We prefer estimators with low variance
  - *Asymptotic normality*: We prefer estimators for which

$$\frac{T - E[T]}{\sqrt{Var(T)}} \sim N(0,1), \quad \text{as } T \to \infty$$

- Usually not all criteria can be met at once!

Uncertainty & Standard Error

- **Whenever** we report a point estimate T for a parameter θ, **we should also** report a measure of *uncertainty* about our estimate
- Measures of uncertainty:
  - Standard error (SE) = SD(T) = $\sqrt{Var(T)}$
  - "Usual" confidence interval: $[T - 2\cdot SE(T), T + 2 \cdot SE(T)]$
  - Other intervals or sets such as $[T_{0.025}, T_{0.975}]$ from a simulation based on the data

Methods of Estimation – How can we systematically construct "good" estimators?

- Several methods have proven useful:
  - *Method of Moments (MoM)*: The k^{th} moment of X is $E[X^k]$. MoM estimators combine unbiased estimates of moments of X.
  - *Least Squares (LS)*: Obtained by minimizing squared error $\sum_{i=1}^{n}(Y_i - E[Y_i])^2$. Ordinary linear regression!
  - *Maximum likelihood (ML)*: The likelihood is the probability of the data we observed. ML estimators (MLE's) choose parameter values that maximize the likelihood.
  - *Bayesian Estimation (Bayes)*: Treat the parameters as random variables, and use Bayes' rule to pick the parameter value most likely, given the data (the reverse of ML!)
- *Notation*: Usually we write $\hat{\theta}$ rather than T, for an estimate/estimator.

Maximum Likelihood Estimators (MLE's)

- Let $X_1, \ldots, X_n$ be an iid sample from $f_X(x;\theta)$, $x_1, \ldots, x_n$ are the observed values
- The *likelihood* of the sample is the joint density

$$\begin{aligned} L(\theta) &= f(x_1, \ldots, x_n;\ \theta) = f(x_1;\ \theta) f(x_2;\ \theta) \cdots f(x_n;\ \theta) \\ &= \prod_{i=1}^{n} f(x_i;\ \theta) \end{aligned}$$

- The *maximum likelihood estimate* $\hat{\theta}_{MLE}$ maximizes $L(\theta)$:

$$L(\hat{\theta}_{MLE}) \geq L(\theta) \quad \forall\ \theta$$

- Strategy: It's usually (but not always) easier to work with the *log likelihood*

$$LL(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\ \theta)\ .$$

Example: MLE of Exponential Distrib.

- Prof Smedley noticed that different students take different amounts of time to learn to make a boxplot. He taught boxplots at the beginning of the semester and then gave a short "boxplot quiz" each day after that, and noted how many days it was before each student got 100% on the daily quiz. For three randomly-chosen students, the number of days before their first 100% was

  $x_1 = 3$, $x_2 = 10$, and $x_3 = 8$
- If the waiting time 'till 100% is an exponential r.v., what is the MLE for the success rate λ?

- The pdf of an exponential rv is $f_x(x;\lambda) = \lambda e^{-\lambda x}$
- The likelihood for an iid sample of size n is

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

- The log-likelihood is then

$$LL(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

- to find the MLE, we differentiate and set to zero

$$0 \stackrel{set}{=} \frac{d}{d\lambda} LL(\lambda) = \frac{d}{d\lambda}\left[n \log \lambda - \lambda \sum_{i=1}^{n} x_i \right] = n/\lambda - \sum_{i=1}^{n} x_i$$

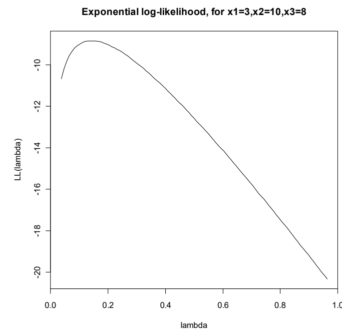
- Solving for λ, we get

$$\hat{\lambda}_{MLE} = n \Big/ \sum_{i=1}^{n} x_i = 1/\bar{X} = \hat{\lambda}_{MoM}$$

- In our example, n=3 and x_1 = 3, x_2 = 10, x_3 = 8, so

$$LL(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i = 3 \log \lambda - \lambda(3 + 10 + 8) = 3 \log \lambda - 21\lambda$$

- Here is the log-likelihood:



- The MLE is

$$1/\bar{X} = 1/((3 + 10 + 8/3) = 1/7$$

Standard Error for MLE's

- Let $\hat{\theta}$ be the MLE for θ, using an iid sample from $f_X(x;\theta)$.
- The *observed information* for estimating θ is

$$I(\theta) = -\frac{d^2}{d\theta^2} LL(\theta) = -LL''(\theta)$$

- If $f_X(x;\theta)$ is a smooth function of θ, then

$$\frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \approx N(0,1)$$

where $Var(\hat{\theta}) \approx 1/I(\hat{\theta})$, so $SE(\hat{\theta}) \approx \sqrt(1/I(\hat{\theta})$

Exponential Example, Continued…

- For the exponential distribution we saw

$$\begin{aligned} LL(\lambda) &= n \log \lambda - \lambda \sum_{i=1}^{n} X_i \\ \hat{\lambda} &= 1/\bar{X} \end{aligned}$$

- The information function is

$$I(\lambda) = -LL''(\lambda) = -\frac{d}{d\lambda}\left[n/\lambda - Y\right] = n/\lambda^2$$

- and so the standard error is

$$SE(\hat{\lambda}) = \sqrt{1/I(\hat{\lambda})} = \sqrt{\hat{\lambda}^2/n} = \hat{\lambda}/\sqrt{n}$$

In our example…

- The MLE and SE are:

$$\begin{aligned} \hat{\lambda} &= 1/7 \\ I(\lambda) &= 3/\lambda^2 \\ SE(\hat{\theta}) &= \hat{\lambda}/\sqrt{3} = (1/7)/\sqrt{3} \end{aligned}$$

- and so a rough 95% confidence interval for λ would be

  [1/7 – 2(1/7)/sqrt(3), 1/7 + 2(1/7)/sqrt(3)],

  or [-0.022, 0.308]

(notice that the left endpoint is not great here!)

Summary

- Discussion/example of multilevel glm
- Review of Estimation
  - Parameter, Statistic, Point Estimator, Point Estimate
  - How Good is the Estimator?
  - Uncertainty & Standard Error
  - Methods of Estimation
  - Maximum Likelihood Estimators (MLE’s)
  - Standard Error for MLE’s