Chapter 2 Part 2

04/22/2021

Graphically approximating an MLE

- Since we don't all know calculus, we will graphically approximate MLE's
- We are looking for the maximum of the likelihood function/graph
- Lets generate a few in R

Graphs

- In (a), the likelihood for the data set of 50 children. The height of each point is the likelihood and the possible values for p_B appear across the horizontal axis. It appears that our data is most likely when $p_B = 0.6$ as we would expect.
- In (b) the log of the likelihood function is maximized at the same spot: $p_B = 0.6$ as in (a)
  - we will see advantages of using log likelihoods a bit later.
  - Based on these graphs, any ideas why the log likelihood is better?
- In (c) and (d) are also maximized at $p_B = 0.6$
  - less variability and a sharper peak since there is more data
  - (although the same proportions of boys and girls).

Numerically approximating an MLE

- This can serve as an alternative to calculus
- Grid searches can be successful but usually for functions with a single peak - We will need to pick a left hand side and right hand side values between which we believe there is a maximum (based on our analysis)
- The smaller increments we make our grid the accurate the answer will be
- What types of situations, do you think a grid search would have difficultly identifying a maximum?
    - Consider our MLE graphs from before

Grid Search

- A grid search specifies a set of finite possible values for p_B and then the likelihood, $\mathrm{Lik}(p_B)$, is computed for each of the possible values.
- First, we define a relatively coarse grid by specifying 50 values for p_B and then computing how likely we would see our data for each of these possible values.
- The second example uses a finer grid, 1,000 values for p_B, which allows us to determine a better (more precise) approximation of the MLE.
- In addition, most packages, like R, have an optimization function which can also be used to obtain MLEs. Both of these approaches we will illustrate now in R

MLEs using calculus (optional)

- Calculus may provide another way to determine an MLE.
- Check it out in the book if you are interested!
- We maximize the likelihood function (more often the log likelihood actually)
- You can ask yourself, why is it OK to maximize the log likelihood function instead of the regular likelihood function?

How does sample size affect the likelihood?

Consider two hypothetical cases under the Sex Unconditional Model:

Hypothetical Case 1: n = 50 children with 30 boys and 20 girls In previous sections, we found the MLE, $\hat{p}_B = 0.6$.

Hypothetical Case 2: n = 1000 children with 600 boys and 400 girls Our earlier work suggests that the MLE here is also $\hat{p}_B = 0.6$.

- The graphs of the likelihoods and log-likelihoods for these two cases we created earlier give us an idea of how the increase in the sample size affects the precision of our estimates.

Increased Sample Size

- The likelihoods and log-likelihoods for the two sample sizes have similar forms; however, the graphs with the larger sample size are much narrower
- Does a narrow graph lead to increased accuracy or decreased accuracy and why?
  - Recall we are trying to estimate p_B the leads to the highest likelihood

Increased Sample Size

- With only 50 children there is a wide range of p_B values that lead to values of the log-likelihood near its maximum, so it's less clear what the optimal p_B is.
- As we have seen in statistics courses before, a larger sample size will result in less variation in our estimates, thereby affecting the power of hypothesis tests and the width of the confidence intervals.
- How does changing sample size effect CI width?

Summary

- Using likelihoods to find estimates of parameters is conceptually intuitive
  - select the estimate for your parameter where your data is most likely.
- Often MLEs make a lot of intuitive sense in the context of a problem as well
  - for example, here the MLE for the probability of a boy is the observed proportion of boys in the data.
- It may seem like a lot of work for such an obvious result, but MLEs have some nice, useful theoretical properties, and we'll see that many more complex models can be fit using the principle of maximum likelihood.

Summary

- We constructed a likelihood that reflected features of our Sex Unconditional Model
- Then we approximated the parameter value for which our data is most likely using a graph or software
  - Or we determined our optimal parameter value exactly using calculus.
  - You may not be familiar with calculus, yet the concept is clear from the graphs: just find the value of p_B where the likelihood or log likelihood is a maximum.
  - Our "best" estimate for p_B, the MLE, is where our data is most likely to be observed.
- Work to understand the *idea* of a likelihood. Likelihoods are the foundation upon which estimates are obtained and models compared for most of this course. Do not be overly concerned with calculus and computation at this point.

Model 2: Sex Conditional

- Our first research question involves determining whether sex runs in the family.
- Do families who already have boys tend to have more additional boys than expected by chance, and do families who already have girls tend to have more additional girls than expected by chance?
- What do you think? And how could we use a statistical model to investigate this phenomenon? There are a number of different ways to construct a model for this question. Here's one possibility.
- Remember our model so far specifically assumed independence between babies, that is, that sex does not run in the family, each baby can be born any sex at equal probability

Model Specification

- Unlike the previous model, the p_B in a Sex Conditional Model *depends* on existing family compositions.
- We introduce **conditional probabilities** and conditional notation to make the dependence explicit.
- One way to interpret the notation $P(A|B)$ is the "probability of A *given* B has occurred."
- Another way to read this notation is the "probability of A *conditional* on B."

Getting Notation Set Up

- Here, let $p_{B|N}$ represent the probability the next child is a boy given that there are equal numbers of boys and girls (sex-neutral) in the existing family. -Let $p_{B|\mathrm{BBias}}$ represent the probability the next child is a boy if the family is boy-biased; i.e., there are more boys than girls prior to this child. Similarly
- Let $p_{B|\mathrm{GBias}}$ represent the probability the next child is a boy if the family is girl-biased; i.e., there are more girls than boys prior to this child.

Understanding These in Terms of Our Example

- Let's think about how these conditional probabilities can be used to describe sex running in families.
- While we only had one parameter, p_B, to estimate in the Sex Unconditional Model, here we have three parameters: $p_{B|N}$, $p_{B|\mathrm{BBias}}$, and $p_{B|\mathrm{GBias}}$.
- What would it mean to us, asking the research question, if all three of these were equal?

Understanding These in Terms of Our Example

- A conditional probability $p_{B|\mathrm{BBias}}$ that is larger than $p_{B|N}$ suggests families with more boys are more likely to produce additional boys in contrast to families with equal boys and girls.
- This finding would support the theory of "boys run in families."

Understanding These in Terms of Our Example

- An analogous argument holds for girls.
- In addition, comparisons of $p_{B|\mathrm{BBias}}$ and $p_{B|\mathrm{GBias}}$ to the parameter estimate p_B from the Sex Unconditional Model may be interesting and can be performed using likelihoods.

Table 1: Family contributions to the likelihood for a Sex Conditional Model using a hypothetical data set of n=50 children from 30 families.

Composition	Likelihood contribution	Prior Status	Number of families
B	$p_{B\|N}$	neutral	6
G	$(1-p_{B\|N})$	neutral	7
BB	$(p_{B\|N})(p_{B\|\mathrm{BBias}})$	neutral, boy bias	5
BG	$(p_{B\|N})(1-p_{B\|\mathrm{BBias}})$	neutral, boy bias	4
GB	$(1-p_{B\|N})(p_{B\|\mathrm{GBias}})$	neutral, girl bias	5
GGB	$(1-p_{B\|N})(1-p_{B\|\mathrm{GBias}})(p_{B\|\mathrm{GBias}})$	neutral, girl bias, girl bias	1
GBB	$(1-p_{B\|N})(p_{B\|\mathrm{GBias}})(p_{B\|N})$	neutral, girl bias, neutral	2
Total			30

Application to Hypothetical Data

- Using the family composition data for 50 children in the 30 families that appears in this table, we construct a likelihood.
- The six singleton families with only one boy contribute $p_{B|N}^6$ to the likelihood and the seven families with only one girl contribute $p_{G|N}^7$ or $(1 - p_{B|N})^7$.
  - Why do we use $1 - p_{B|N}$ instead of $p_{G|N}$?
- There are five families with two boys each with probability $(p_{B|N})(p_{B|\mathrm{BBias}})$ contributing $(p_{B|N})(p_{B|\mathrm{BBias}})]^5$

Application to Hypothetical Data

We construct the likelihood using data from all 30 families assuming families are independent to get:

$$\begin{aligned}\text{Lik}(p_{B|N},\ p_{B|\text{BBias}},\ p_{B|\text{GBias}}) = \big[(p_{B|N})^{17}(1-p_{B|N})^{15}(p_{B|\text{BBias}})^{5} \\ (1-p_{B|\text{BBias}})^{4}(p_{B|\text{GBias}})^{8}(1-p_{B|\text{GBias}})\big]\end{aligned} \tag{1}$$

A Couple of Points are Worth Noting

- There are 50 factors in the likelihood corresponding to the 50 children in these 30 families.
- In the Sex Unconditional example we only had one parameter, p_B; here we have three parameters.
  - This likelihood does not simplify like the Sex Unconditional Model to one that is a product of only two powers: one of p_B and the other of $1 - p_B$.
- The basic idea we discussed regarding using a likelihood to find parameter estimates is the same.

Getting the MLE

- To obtain the MLEs, we need to find the combination of values for our three parameters where the data is most likely to be observed.
- Conceptually, we are trying different combinations of possible values for these three parameters, one after another, until we find *the* combination where the likelihood is a maximum.
- It will not be as easy to graph this likelihood and we will need multivariable calculus to locate the optimal combination of parameter values where the likelihood is a maximum. In this text, we do not assume you know multivariable calculus, but we do want you to retain the concepts associated with maximum likelihood estimates. In practice, we use software to obtain MLEs.

Getting the MLE

- With calculus, we can take partial derivatives of the likelihood with respect to each parameter assuming the other parameters are fixed.
- Differentiating the log of the likelihood often makes things easier. This same approach is recommended here. Set each partial derivative to 0 and solve for all parameters simultaneously.

Switching to Log

- Knowing that it is easier to work with log-likelihoods, let's take the log of the likelihood we constructed.

$$\log(\text{Lik}(p_{B|N},\, p_{B|\text{BBias}},\, p_{B|\text{GBias}})) = 17\log(p_{B|N}) + 15\log(1-p_{B|N}) + 5\log(p_{B|\text{BBias}}) + 4\log(1-p_{B|\text{BBias}}) + 8\log(p_{B|\text{GBias}}) + 1\log(1-p_{B|\text{GBias}})$$

Taking a partial derivative with respect to $p_{B|N}$ yields

$$\frac{17}{p_{B|N}} - \frac{15}{1-p_{B|N}} = 0$$

$$\hat{p}_{B|N} = \frac{17}{32}$$

$$= 0.53$$

Making Sense of it (without us doing the calculus)

- This estimate follows naturally. First consider all of the children who enter into a family with an equal number of boys and girls.
- From our table, we can see there are 32 such children (30 are first kids and 2 are third kids in families with 1 boy and 1 girl).
- Of those children, 17 are boys.
- Given that a child joins a sex-neutral family, the chance they are a boy is 17/32.

Making Sense of it (without us doing the calculus)

Similar calculations for $p_{B|\text{BBias}}$ and $p_{B|\text{GBias}}$ yield:

$$\hat{p}_{B|N} = 17/32 = 0.53$$
$$\hat{p}_{B|\text{B Bias}} = 5/9 = 0.56$$
$$\hat{p}_{B|\text{G Bias}} = 8/9 = 0.89$$

Summary of our Model

- If we anticipate any "sex running in families" effect, we would expect $p_{B|\text{B Bias}}$ to be larger than the probability of a boy in the neutral setting, $p_{B|N}$.
- In our small hypothetical example, $\hat{p}_{B|\text{B Bias}}$ is slightly greater than 0.53, providing light support for the "sex runs in families" theory when it comes to boys.
- What about girls? Do families with more girls than boys tend to have a greater probability of having a girl?
- We found that the MLE for the probability of a girl in a girl-biased setting is 1-0.89=0.11.