Chapter 2 Part 1

04/22/2021

Chapter 2: Using Likelihoods

Learning Objectives

After finishing this chapter, you should be able to:

- Describe the concept of a likelihood, in words.
- Know and apply the Principle of Maximum Likelihood for a simple example.
- Identify three ways in which you can obtain or approximate an MLE .
- Use likelihoods to compare models.
- Construct a likelihood for a simple model.

Chapter Packages

Get These Installed and Called in your R Script

```
# Packages required for Chapter 2

#install.packages("mosaic","xtable")

library(gridExtra)
library(knitr)
library(mosaic)
library(xtable)
library(kableExtra)
library(tidyverse)
```

Beyond

- We want to move beyond the bubble of independent, identically distributed, normal responses (iidN).
  - When have we done this before?
- Lets see an example in R of using LLSR on a Binary Response

Likelihoods

- In this instance we'll use logistic regression instead of linear least squares regression.
- Fitting a logistic regression requires the use of likelihood methods.
- Another setting where likelihood methods come into play is when data is produced from a complex structure which may imply correlation among outcomes.
  - For example, test scores for students who have been taught by the same teacher may be correlated.
  - We'll see that likelihood methods are useful when modeling correlated data.

Likelihoods

- Likelihood methods not only provide a great deal of flexibility in the types of models we can fit, but they also provide ways in which to compare models as well.
- You might find likelihood methods a bit more complicated, but conceptually the approach is straightforward.
- As you go through the material here, worry less about calculus and computational details and focus on the concepts.
- You will have software to help you with computation, but model specification and interpretation will be up to you.

Case Study: Does Sex Run in Families?

- Read the introduction to 2.2 in our book "Case Study: Does Sex Run in Families?"

Research Questions

- We now specify several models related to sex balance in families.
- Our models liken having babies to flipping a coin (heads=boy, tails=girl), of course, recognizing that in truth there is a little more to having babies.
- The baseline model (Model 0) assumes that the probability of a boy is the same as the probability of a girl.
- The first model (Model 1) considers the situation that the coin is loaded and the probability of heads (a boy) is different than the probability of tails (a girl).
- Model 2 is conditioned on the previous number of boys or girls in a family to get at the question of whether sex runs in families.

Research Questions

- Models 0 and 1 assume that having children is like flipping a coin. The gender of each child is independent of the gender of other children and the probability of a boy is the same for each new child.
- Let p_B be the probability a child is a boy.

Research Questions

1. **Model 0: Sex Unconditional Model (Equal probabilities).** Is a child just as likely to be a boy as it is to be a girl; is $p_B = 0.5$?
2. **Model 1: Sex Unconditional Model (Different probabilities).** Is the coin loaded; is $p_B \neq 0.5$?
3. **Model 2: Sex Conditional Model (Sex bias).** Do boys or girls run in families? That is, is there a tendency for families with more boys than girls to be more likely to produce another boy? Is the case the same for girls?
4. **Model 3: Stopping Rule Model (Waiting for a boy).** Is there evidence that couples stop having children once a boy is born?

Goal

Goal: incorporate the family composition data represented as series of coin flips to find the "best" estimate for the probability of having a boy, p_B,

- We need to evaluate the assumptions built into these models.
- We will use likelihood-based methods to compare models!

Starting Off

- While the NLSY data is of interest, we start with a smaller, hypothetical data set of 30 families with a total of 50 children in order to illustrate concepts related to likelihoods
- The data are the frequencies of possible family gender compositions for one-, two-, and three-child families.
- The methods we develop on this small data set will then be applied to the one-, two- and three-family NLSY data.
- Lets set this up in R

Model 0: Sex Unconditional, Equal Probabilities

- For the Sex Unconditional models, having children is modeled using coin flips.
- The result of each flip is independent of results of other flips.
- The chance that a baby is a boy is specified to be $p_B = 0.5$.
- It makes no difference if the first and third children are boys, the probability that the second child is a boy is 0.5; that is, the results for each child are **independent** of the others.
- Under this model you expect to see equal numbers of boys and girls.

Model 1: Sex Unconditional, Unequal Probabilities

- You may want your model to allow for the probability of a boy, p_B, to be something different than 0.5.
- With this version of the Sex Unconditional model, $p_B > 0.5$ or $p_B < 0.5$ or $p_B = 0.5$
  - you expect to see more boys than girls or fewer boys than girls or equal numbers of boys and girls, respectively.
- We retain the assumption of independence; that is, the probability of a boy, p_B, is the same for each child.
  - Seeing a boy for the first child will not lead you to change the probability that the second child is a boy
  - this would not imply that "sex runs in families."

What Is a Likelihood?

- As is often the case in statistics, our objective is to find an estimate for a model parameter using our data
- Currently the parameter to estimate is the probability of a boy, p_B
- The data is the sex composition for each family.
- One way in which to interpret probability is to imagine repeatedly producing children.
- The probability of a boy will be the overall proportion of boys as the number of children increases.

What Is a Likelihood?

- With likelihood methods, conceptually we consider different possible values for our parameter(s), p_B, and determine how likely we would be to see our observed data in each case, $\text{Lik}(p_B)$.
- We'll select as our estimate the value of p_B for which our data is most likely.
- A **likelihood** is a function that tells us how likely we are to observe our data for a given parameter value, p_B. For a single family which has a girl followed by two boys, GBB, the likelihood function looks like:

$$\text{Lik}(p_B) = P(G)P(B)P(B) = (1 - p_B)p_B^2$$

- Lets visualize this function in R!

Understanding This Graph

- When $p_B = 0.3$ we see a family of a girl followed by two boys 6.3% ($0.7 \cdot 0.3^2$) of the time.
- However, it indicates that we are much more likely to see our data if $p_B = 0.6$ where the likelihood of GBB is $0.4 \cdot 0.6^2$ or 14.4%.
- If the choice was between 0.3 and 0.6 for an estimate of p_B, we'd choose 0.6.
- Where is the best estimate?
- How would we find it?

Best Estimate

- The "best" estimate of p_B would be the value where we are most likely to see our data from all possible values between 0 and 1, which we refer to as the **maximum likelihood estimate** or MLE.
- We can approximate an MLE using graphical or numerical approaches. Graphically, here it looks like the MLE is just above 0.6.
- In many, but not all, circumstances, we can obtain an MLE exactly using calculus.
- In this simple example, the MLE is 2/3. This is consistent with our intuition since 2 out of the 3 children are boys.

Add More Data

- Suppose another family consisting of three girls is added to our data set.
- We've already seen that the Sex Unconditional Model multiplies probabilities to construct a likelihood because children are independent of one another.
- Extending this idea, families can be assumed to be independent of one another so that the likelihood for both families can be obtained by multiplication. With two families (GBB and GGG) our likelihood is now:

$$\begin{aligned}\text{Lik}(p_B) &= P(GBB)P(GGG) \\ &= [(1-p_B)p_B^2][(1-p_B)^3] \\ &= (1-p_B)^4 p_B^2\end{aligned}$$

- Now we generate a new likelihood in R!

New Plot

- This is right skewed with an MLE at approximately 0.3.
- Using calculus, we can show that the MLE is precisely $1/3$ which is consistent with intuition given the 2 boys and 4 girls in our data.

Now to 30 Families

- Turning now to our hypothetical data with 30 families who have a total of 50 children, we can create the likelihood contribution for each of the family compositions.
- The likelihood function for the hypothetical data set can be found by taking the product of the entries in the last column of our 30 family table before and simplifying.

$$\begin{aligned} \text{Lik}(p_B) &= p_B^6(1-p_B)^7 p_B^{10} \cdots \\ &= p_B^{30}(1-p_B)^{20} \end{aligned} \tag{1}$$

- We will calculate the likelihood (via multiplication) and plot in R

General Form

- The general equation

$$\text{Lik}(p_B) = p_B^{n_{\text{Boys}}}(1 - p_B)^{n_{\text{Girls}}}$$

- As we asserted before, the MLE will be the (number of boys)/(number of kids) or 30/50 here.