Using Likelihoods

Prof. Maria Tackett

01.19.22

STA 310

Click for PDF of slides

STA 310

Announcements

- Homework 01 due Wednesday at 11:59pm
- Week 03 reading: BMLR: Chapter 3 - Distribution Theory
- See syllabus for office hours schedule
  - Office hours online this week
- Team lab tomorrow - introducing Mini-Project 01
  - Find a paper using a GLM in their analysis
  - Evaluate the analysis in the paper
  - "Replicate" the analysis the same or similar data
  - Present results in a presentation and short write up
  - More details in lab!

In-person learning 😷

Attendance in lectures and labs is expected as long as you're healthy and not in quarantine

Lectures

- If you're unable to attend, you can watch the recording of the lecture on Panopto (link in Sakai)
- Ask questions on GitHub Discussions or in office hours

Labs

- Labs are not recorded
- On weeks with teamwork: If you are unable to attend lab but are able to participate remotely, work with your teammates to set up a Zoom call

STA 310

Class Q&A Forum: GitHub Discussions

- Class Q&A forum on GitHub Discussions
  - Place for questions about course content, assignments, etc.
  - Only use email for personal questions (e.g., grades, illness, etc.)
  - Let Prof. Tackett know if you do not have access to the forum

Demo

Homework 01

- Notes on variable transformations
- Exercise 5

> *This question will be graded based on*
>
> *The quality of the model selection process, including the exploratory data analysis. A high quality model selection process is accurate, comprehensive, and strategic (e.g., trying all possible interaction terms will not receive full credit).*
>
> *The quality of the summary. A high quality summary is accurate, comprehensive, answers the primary analysis question, and tells a cohesive story (e.g., a list of interpretations will not receive full credit).*

Using Likelihoods

Learning goals

- Describe the concept of a likelihood
- Construct the likelihood for a simple model
- Define the Maximum Likelihood Estimate (MLE) and use it to answer an analysis question
- Identify three ways to calculate or approximate the MLE and apply these methods to find the MLE for a simple model
- Use likelihoods to compare models (next week)

What is the likelihood?

A likelihood is a function that tells us how likely we are to observe our data for a given parameter value (or values).

- Unlike Ordinary Least Squares (OLS), they do not require the responses be independent, identically distributed, and normal (iidN)
- They are not the same as probability functions
    - Probability function: Fixed parameter value(s) + input possible outcomes $\Rightarrow$ probability of seeing the different outcomes given the parameter value(s)
    - Likelihood: Fixed data + input possible parameter values $\Rightarrow$ probability of seeing the fixed data for each parameter value

STA 310

Fouls in college basketball games

The data includes 30 randomly selected NCAA men's basketball games played in the 2009 - 2010 season.

We will focus on the variables **`foul1`**, **`foul2`**, and **`foul3`**, which indicate which team had a foul called them for the 1st, 2nd, and 3rd fouls, respectively.

- **`H`**: Foul was called on the home team
- **`V`**: Foul was called on the visiting team

We are focusing on the first three fouls for this analysis, but this could easily be extended to include all fouls in a game.

The dataset was derived from **`basektball0910.csv`** used in [BMLR Section 11.2]

STA 310

Fouls in college basketball games

```
refs <- read_csv("data/04-refs.csv")
refs %>% slice(1:5) %>% kable()
```

game	date	visitor	hometeam	foul1	foul2	foul3
166	20100126	CLEM	BC	V	V	V
224	20100224	DEPAUL	CIN	H	H	V
317	20100109	MARQET	NOVA	H	H	H
214	20100228	MARQET	SETON	V	V	H
278	20100128	SETON	SFL	H	V	V

We will treat the games as independent in this analysis.

Different likelihood models

Model 1 (Unconditional Model): What is the probability the referees call a foul on the home team, assuming foul calls within a game are independent?

Model 2 (Conditional Conditional Model):

- Is there a tendency for the referees to call more fouls on the visiting team or home team?
- Is there a tendency for referees to call a foul on the team that already has more fouls?

Ultimately we want to decide which model is better.

Exploratory data analysis

```
refs %>%
count(foul1, foul2, foul3) %>% kable()
```

foul1	foul2	foul3	n
H	H	H	3
H	H	V	2
H	V	H	3
H	V	V	7
V	H	H	7
V	H	V	1
V	V	H	5
V	V	V	2

There are

- 46 total fouls on the home team
- 44 total fouls on the visiting team

Model 1: Unconditional model

What is the probability the referees call a foul on the home team, assuming foul calls within a game are independent?

STA 310

Likelihood

Let p_H be the probability the referees call a foul on the home team.

The likelihood for a single observation

$$Lik(p_H) = p_H^{y_i}(1 - p_H)^{n_i - y_i}$$

Where y_i is the number of fouls called on the home team.

(In this example, we know $n_i = 3$ for all observations.)

Example

For a single game where the first three fouls are H, H, V, then

$$Lik(p_H) = p_H^2(1 - p_H)^{3-2} = p_H^2(1 - p_H)$$

Model 1: Likelihood contribution

Foul1	Foul2	Foul3	n	Likelihood Contribution
H	H	H	3	p_H^3
H	H	V	2	$p_H^2(1 - p_H)$
H	V	H	3	$p_H^2(1 - p_H)$
H	V	V	7	A
V	H	H	7	B
V	H	V	1	$p_H(1 - p_H)^2$
V	V	H	5	$p_H(1 - p_H)^2$
V	V	V	2	$(1 - p_H)^3$

02:00

Fill in A and B.

STA 310

Model 1: Likelihood function

Because the observations (the games) are independent, the likelihood is

$$Lik(p_H) = \prod_{i=1}^{n} p_H^{y_i}(1 - p_H)^{3-y_i}$$

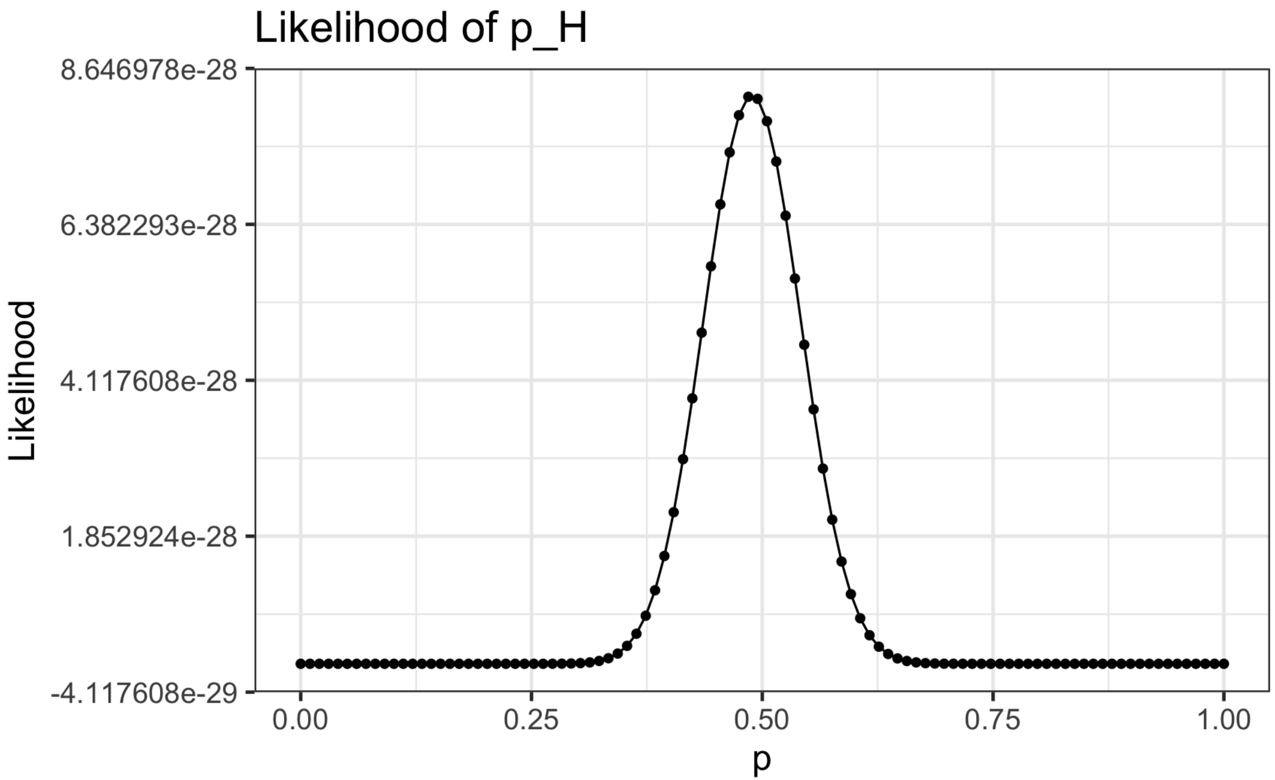
We will use this function to find the maximum likelihood estimate (MLE). The MLE is the value between 0 and 1 where we are most likely to see the observed data.

STA 310

Visualizing the likelihood

Plot

Code



Visualizing the likelihood

Plot

Code

```
p <- seq(0,1, length.out = 100) #sequence of 100 values betw
lik <- p^44 *(1 -p)^46

x <- tibble(p = p, lik = lik)
ggplot(data = x, aes(x = p, y = lik)) +
  geom_point() +
  geom_line() +
  labs(y = "Likelihood",
       title = "Likelihood of p_H")
```

STA 310

What is your best guess for the MLE, $\hat{p}_H$?

A. 0.489

B. 0.500

C. 0.511

D. 0.556

Click here to submit your response.

02:00

Finding the maximum likelihood estimate

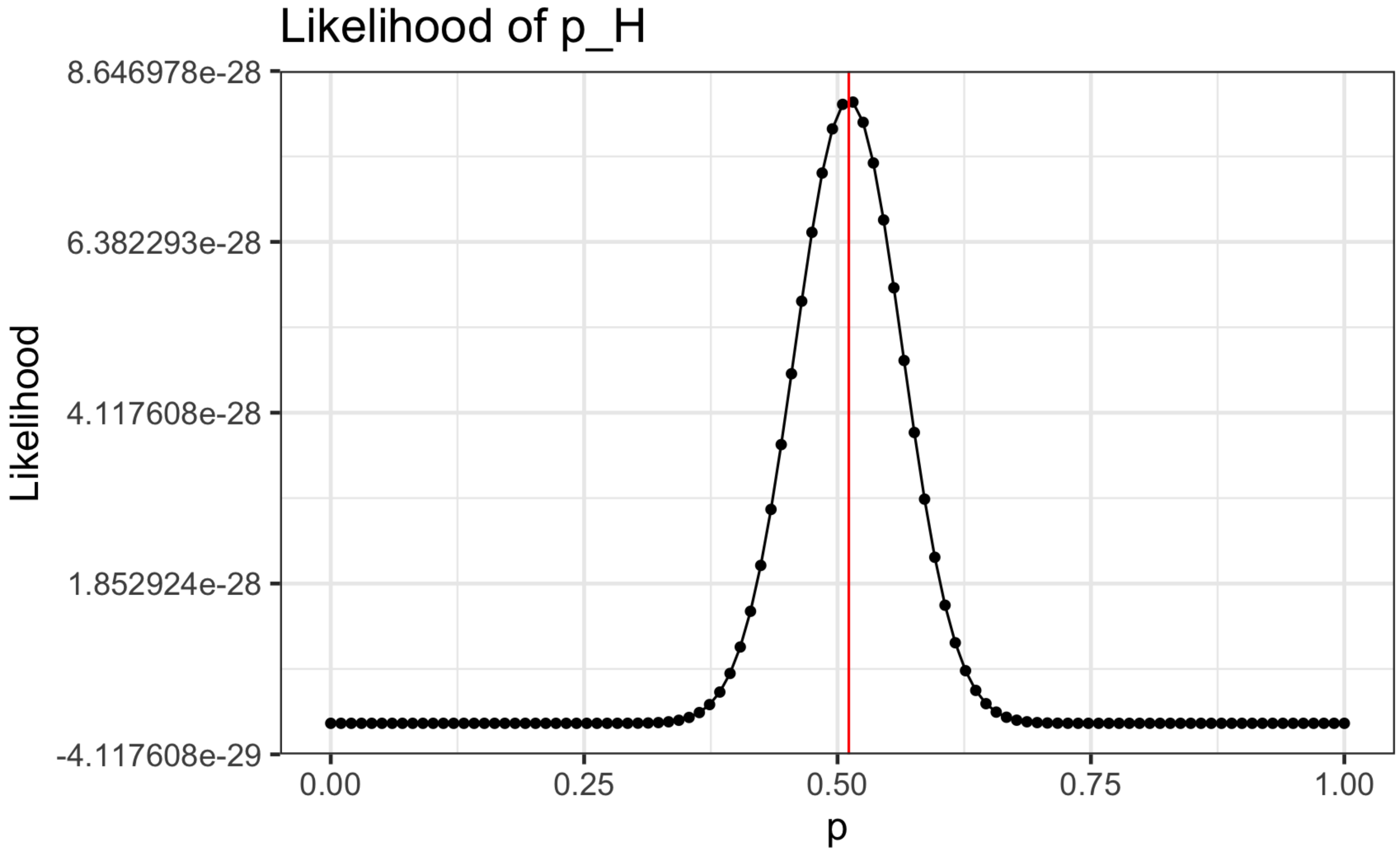
There are three primary ways to find the MLE

✅ Approximate using a graph

✅ Numerical approximation

✅ Using calculus

Approximate MLE from a graph



Find the MLE using numerical approximation

Specify a finite set of possible values the for p_H and calculate the likelihood for each value

```
# write an R function for the likelihood
ref_lik <- function(ph) {
  ph^46 *(1 - ph)^44
}
```

```
# use the optimize function to find the MLE
optimize(ref_lik, interval = c(0,1), maximum = TRUE)
```

```
## $maximum
## [1] 0.5111132
##
## $objective
## [1] 8.25947e-28
```

STA 310

Find MLE using calculus

- Find the MLE by taking the first derivative of the likelihood function.
- This can be tricky because of the Product Rule, so we can maximize the log(Likelihood) instead. The same value maximizes the likelihood and log(Likelihood)

STA 310

Find MLE using calculus

$$Lik(p_H) = \prod_{i=1}^{n} p_H^{y_i}(1 - p_H)^{3-y_i}$$

$$\log(Lik(p_H)) = \sum_{i=1}^{n} y_i \log(p_H) + (3 - y_i)\log(1 - p_H)$$

$$= 46\log(p_H) + 44\log(1 - p_H)$$

STA 310

Find MLE using calculus

$$\frac{d}{dp_H}\log(Lik(p_H)) = \frac{46}{p_H} - \frac{44}{1-p_H} = 0$$

$$\Rightarrow \frac{46}{p_H} = \frac{44}{1-p_H}$$

$$\Rightarrow 46(1-p_H) = 44p_H$$

$$\Rightarrow 46 = 90p_H$$

$$\hat{p}_H = \frac{46}{90} = 0.511$$

😐

STA 310

Model 2: Conditional model

Is there a tendency for the referees to call more fouls on the visiting team or home team?

Is there a tendency for referees to call a foul on the team that already has more fouls?

Model 2: Likelihood contributions

- Now let's assume fouls are <u>not</u> independent within each game. We will specify this dependence using conditional probabilities.
    - Conditional probability: $P(A|B) =$ Probability of A given B has occurred

Define new parameters:

- $p_{H|N}$: Probability referees call foul on home team given there are equal numbers of fouls on the home and visiting teams

- $p_{H|HBias}$: Probability referees call foul on home team given there are more prior fouls on the home team

- $p_{H|V Bias}: Probability referees call foul on home team given there are more prior fouls on the visiting team

Model 2: Likelihood contributions

Foul1	Foul2	Foul3	n	Likelihood Contribution
H	H	H	3	$(p_{H\|N})(p_{H\|HBias})(p_{H\|HBias}) = (p_{H\|N})(p_{H\|HBias})^2$
H	H	V	2	$(p_{H\|N})(p_{H\|HBias})(1 - p_{H\|HBias})$
H	V	H	3	$(p_{H\|N})(1 - p_{H\|HBias})(p_{H\|N}) = (p_{H\|N})^2(1 - p_{H\|HBias})$
H	V	V	7	A
V	H	H	7	B
V	H	V	1	$(1 - p_{H\|N})(p_{H\|VBias})(1 - p_{H\|N}) = (1 - p_{H\|N})^2(p_{H\|VBias})$
V	V	H	5	$(1 - p_{H\|N})(1 - p_{H\|VBias})(p_{H\|VBias})$
V	V	V	2	$(1 - p_{H\|N})(1 - p_{H\|VBias})(1 - p_{H\|VBias})$ $= (1 - p_{H\|N})(1 - p_{H\|VBias})^2$

STA 310

Likelihood function

$$
\begin{aligned}
Lik(p_{H|N}, p_{H|HBias}, p_{H|VBias}) = [(p_{H|N})^{25}(1 - p_{H|N})^{23}(p_{H|HBias})^{8} \\
(1 - p_{H|HBias})^{12}(p_{H|VBias})^{13}(1 - p_{H|VBias})^{9}]
\end{aligned}
$$

(Note: The exponents sum to 90, the total number of fouls in the data)

$$
\begin{aligned}
\log(Lik(p_{H|N}, p_{H|HBias}, p_{H|VBias})) &= 25\log(p_{H|N}) + 23\log(1 - p_{H|N}) \\
&+ 8\log(p_{H|HBias}) + 12\log(1 - p_{H|HBias}) \\
&+ 13\log(p_{H|VBias}) + 9\log(1 - p_{H|VBias})
\end{aligned}
$$

STA 310

If fouls within a game are independent, how would you expect $\hat{p}_H$, $\hat{p}_{H|HBias}$ and $\hat{p}_{H|VBias}$ to compare?

a. $\hat{p}_H$ is greater than $\hat{p}_{H|HBias}$ and $\hat{p}_{H|VBias}$

b. $\hat{p}_{H|HBias}$ is greater than $\hat{p}_H$ and $\hat{p}_{H|VBias}$

c. $\hat{p}_{H|VBias}$ is greater than $\hat{p}_H$ and $\hat{p}_{H|VBias}$

d. They are all approximately equal.

If there is a tendency for referees to call a foul on the team that already has more fouls, how would you expect $\hat{p}_H$ and $\hat{p}_{H|HBias}$ to compare?

a. $\hat{p}_H$ is greater than $\hat{p}_{H|HBias}$

b. $\hat{p}_{H|HBias}$ is greater than $\hat{p}_H$

c. They are approximately equal.

Click here to submit your response.

Next time

- Using likelihoods to compare models
- Chapter 3: Distribution theory

Acknowledgements

These slides are based on content in [BMLR Chapter 2 - Beyond Least Squares: Using Likelihoods]