Using likelihoods to compare models

Prof. Maria Tackett

01.24.22

[Click for PDF of slides]

Announcements

- Week 03 reading:
  - BMLR: Chapter 3 - Distribution Theory (for reference)
  - BMLR: Chapter 4 - Poisson Regression
- Quiz 01 Tue, Jan 25 at 9am - Thu, Jan 27 at 3:30pm (start of lab)

Quiz 01

- Open Jan 25 at 9am and must be completed by Thu, Jan 27 at 3:30pm
    - The quiz is not timed and will be administered in Gradescope.
- Covers
    - Syllabus
    - BMLR Chapters 1 - 2
    - Jan 05 - Jan 24 lectures
- Fill in the blank, multiple choice, short answer questions
- Open book, open note, open internet (not crowd sourcing sites). You <u>cannot</u> discuss the quiz with anyone else. Please email me if you have questions.

STA 310

Learning goals

- Use likelihood to compare models
- Activity: Exploring the response variable for mini-project 01

Recap

Fouls in college basketball games

The data set [`04-refs.csv`] 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

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.

Likelihoods

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

Model 1 (Unconditional Model)

- p_H: probability of a foul being called on the home team

Model 2 (Conditional Model)

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

STA 310

Likelihoods

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

Model 1 (Unconditional Model)

$$Lik(p_H) = p_H^{46}(1 - p_H)^{44}$$

Model 2 (Conditional Model)

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

STA 310

Maximum likelihood estimates

The **maximum likelihood estimate (MLE)** is the value between 0 and 1 where we are most likely to see the observed data.

Model 1 (Unconditional Model)

- $\hat{p}_H = 46/90 = 0.511$

Model 2 (Conditional Model)

- $\hat{p}_{H|N} = 25/48 = 0.521$
- $\hat{p}_{H|HBias} = 8/20 = 0.4$
- $\hat{p}_{H|VBias} = 13/22 = 0.591$

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

MLEs for Model 2

Click here for details on finding MLEs for Model2

Model comparison

STA 310

Model comparisons

- Nested models
- Non-nested models

Comparing nested models

Nested Models

Nested models: Models such that the parameters of the reduced model are a subset of the parameters for a larger model

Example:

$$
\begin{aligned}
\text{Model A: } y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
\text{Model B: } y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon
\end{aligned}
$$

Model A is nested in Model B. We could use likelihoods to test whether it is useful to add x_3 and x_4 to the model.

$$
\begin{aligned}
H_0 &: \beta_3 = \beta_4 = 0 \\
H_a &: \text{ at least one } \beta_j \text{ is not equal to } 0
\end{aligned}
$$

Nested models

Another way to think about nested models: Parameters in larger model can be equated to get the simpler model or if some parameters can be set to constants

Example:

$$\text{Model 1: } p_H$$
$$\text{Model 2: } p_{H|N}, p_{H|HBias}, p_{H|VBias}$$

Model 1 is nested in Model 2. The parameters $p_{H|N}$, $p_{H|HBias}$, and $p_{H|VBias}$ can be set equal to p_H to get Model 1.

$$\begin{aligned} H_0 &: p_{H|N} = p_{H|HBias} = p_{H|VBias} = p_H \\ H_a &: \text{At least one of } p_{H|N}, p_{H|HBias}, p_{H|VBias} \text{ differs from the others} \end{aligned}$$

STA 310

Steps to compare models

1️⃣ Find the MLEs for each model.

2️⃣ Plug the MLEs into the log-likelihood function for each model to get the maximum value of the log-likelihood for each model.

3️⃣ Find the difference in the maximum log-likelihoods

4️⃣ Use the Likelihood Ratio Test to determine if the difference is statistically significant

Steps 1 - 2

Find the MLEs for each model and plug them into the log-likelihood functions.

Model 1:

- $\hat{p}_H = 46/90 = 0.511$

```r
loglik1 <- function(ph){
 log(ph^46 * (1 - ph)^44)
}
loglik1(46/90)
```

```
## [1] -62.36102
```

Model 2

- $\hat{p}_{H|N} = 25/48 = 0.521$
- $\hat{p}_{H|HBias} = 8/20 = 0.4$
- $\hat{p}_{H|VBias} = 13/22 = 0.591$

```r
loglik2 <- function(phn, phh, phv) {
  log(phn^25 * (1 - phn)^23 * phh^8 *
        (1 - phh)^12 * phv^13 * (1 - phv)^9)
}
loglik2(25/48, 8/20, 13/22)
```

```
## [1] -61.57319
```

Step 3

Find the difference in the log-likelihoods

```
(diff <- loglik2(25/48, 8/20, 13/22) - loglik1(46/90))
```

```
## [1] 0.7878318
```

Is the difference in the maximum log-likelihoods statistically significant?

Likelihood Ratio Test

Test statistic

$$LRT = 2[\max\{\log(Lik(\text{larger model}))\} - \max\{\log(Lik(\text{reduced model}))\}]$$

$$= 2\log\left(\frac{\max\{(Lik(\text{larger model})\}}{\max\{(Lik(\text{reduced model})\}}\right)$$

LRT follows a χ^2 distribution where the degrees of freedom equal the difference in the number of parameters between the two models

STA 310

Step 4

```r
(LRT <- 2 * (loglik2(25/48, 8/20, 13/22) - loglik1(46/90)))
```

```
## [1] 1.575664
```

The test statistic follows a χ^2 distribution with 2 degrees of freedom. Therefore, the p-value is $P(\chi^2 > LRT)$.

```r
pchisq(LRT, 2, lower.tail = FALSE)
```

```
## [1] 0.4548299
```

The p-value is very large, so we fail to reject H_0. We do not have convincing evidence that the conditional model is an improvement over the unconditional model. Therefore, we can stick with the unconditional model.

STA 310

Comparing non-nested models

STA 310

Comparing non-nested models

AIC = -2(max log-likelihood) + 2p

```
(Model1_AIC <- 2 * loglik1(46/90) + 2 
```

```
## [1] -122.722
```

```
(Model2_AIC <-2 * loglik2(25/48, 8/20,
```

```
## [1] -117.1464
```

BIC = -2(max log-likelihood) + plog(n)

```
(Model1_BIC <- 2 * loglik1(46/90) + 1 
```

```
## [1] -121.3208
```

```
(Model2_BIC <-2 * loglik2(25/48, 8/20,
```

```
## [1] -112.9428
```

Choose Model 1, the unconditional model, based on AIC and BIC

STA 310

Looking ahead

- Likelihoods help us answer the question of how likely we are to observe the data given different parameters

- In this example, we did not consider covariates, so in practice the parameters we want to estimate will look more similar to this

$$p_H = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

- Finding the MLE becomes much more complex and numerical methods may be required.
    - We will primarily rely on software to find the MLE, but the conceptual ideas will be the same

STA 310

Response variable in mini-project 01

Activity Instructions

The goal of this activity is for your team to start exploring the response variable for your mini-project 01 analysis. The properties explored in this activity are ones you will consider throughout the semester as you decide which GLM is most appropriate for a given data set. Use Table 3.1 in BMLR for reference.

Write the following for the primary response variable in your analysis:

- What is the response variable? What is its definition?
- Is the response variable discrete or continuous?
- What possible values can it take? (not necessarily just the values in the data set)
- What is the name of the distribution the variable follows?
- What is/are the parameter(s) for this distribution? Estimate the parameters from the data.
- Visualize the distribution of the response variable. Is this what you expected? Why or why not?

Activity Instructions

[Click here]() to put the answers on your team's slide.

You can add any analysis to the bottom of the **`proposal.Rmd`** document in your team's project repo.

Looking ahead

- Review Chapter 3 - Distribution Theory
  - Use this chapter as a reference throughout the semester
- For next time - Chapter 4 - Poisson Regression

STA 310

Acknowledgements

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