Estimating parameters

Study Sessions (in addition to office hours)

TA: Zhanyang Sun

Time: Wednesdays 7-8 PM (starting today!)

Location: Kirby 120

Reminder: Lab 1 Due Friday at 12pm (noon)

- Extension policy: Bank of 5 days, you may use 1-2 on an assignment

Goal

Logistic regression model:

$$Y_i \sim Bernoulli(\pi_i) \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

> Given data $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, how do I estimate β_0 and β_1?

Motivating example

$Y_i =$ result of flipping a coin (Heads or Tails)

Is Y_i a random variable?

Motivating example

$Y_i =$ result of flipping a coin (Heads or Tails)

> Is Y_i a random variable?

Yes -- there are two possible outcomes, but we don't know which will happen until we flip the coin.

Motivating example

$=1$ $=0$

Y_i = result of flipping a coin (Heads or Tails)

Let's make a model:

- **Step 1:** Distribution of the response

$P(Y_i = 1)$

$$Y_i \sim Bernoulli(\pi)$$

(assuming for new
π doesn't change)

- **Step 2:** Construct a model for the parameters

$$\pi = ??$$

Motivating example

$Y_i =$ result of flipping a coin (Heads or Tails)

Let's make a model:

- **Step 1:** Distribution of the response

$$Y_i \sim Bernoulli(\pi)$$

- **Step 2:** Construct a model for the parameters

$$\pi = ??$$

Right now, we don't have any information to help us estimate π

Motivating example

$Y_i =$ result of flipping a coin (Heads or Tails)

Suppose your friend estimates that the probability of heads is 0.9

- $Y_i \sim Bernoulli(\pi)$
- $\widehat{\pi} = 0.9$

> How can we assess whether this estimate $\widehat{\pi}$ is reasonable?

Motivating example

$Y_i =$ result of flipping a coin (Heads or Tails)

Suppose your friend estimates that the probability of heads is 0.9

- $Y_i \sim Bernoulli(\pi)$
- $\widehat{\pi} = 0.9$

> How can we assess whether this estimate $\widehat{\pi}$ is reasonable?

See if the estimate fits observed data.

Motivating example

Suppose we flip the coin 5 times, and observe

$$y_1, \ldots, y_5 = T, T, T, T, H$$

What is the probability of (i.e., how *likely* is) getting this string of flips if $\pi = 0.9$? Discuss with your neighbor for 2 minutes, then we will discuss as a class.

$P(H) = 0.9 \Rightarrow P(T) = 0.1$

$P(T, T, T, T, H) = (0.1)(0.1)(0.1)(0.1)(0.9)$

$= (0.1)^4 (0.9)$ $\Leftarrow$ multiplication rule for independent events

$= 0.00009$

Likelihood

Definition: The *likelihood* $L(Model) = P(Data|Model)$ of a model is the probability of the observed data, given that we assume a certain model and certain values for the parameters that define that model.

- Model: $Y_i \sim Bernoulli(\pi)$, and $\widehat{\pi} = 0.9$
- Data: $y_1, \ldots, y_5 = T, T, T, T, H$
- Likelihood: $L(\widehat{\pi}) = P(y_1, \ldots y_5|\pi = 0.9) = 0.00009$

Class Activity, Part I

https://sta214-f22.github.io/class_activities/ca_lecture_5.html

Class Activity

$$L(0.2) = P(y_1, \ldots, y_5 | \pi = 0.2)$$
$$= (0.2)(0.8)(0.8)(0.2)(0.8) = 0.020$$

$$L(0.3) = P(y_1, \ldots, y_5 | \pi = 0.3)$$
$$= (0.3)(0.7)(0.7)(0.3)(0.7) = 0.031$$

> Which value, 0.2 or 0.3, seems more reasonable?

higher likelihood!

Class Activity

Which value of $\widehat{\pi}$ in the table would you pick?

$\hat{\pi} = 0.4$ b/c it has highest likelihood

(also observed proportion of 1's in data)

Maximum likelihood

Maximum likelihood principle: estimate the parameters to be the values that maximize the likelihood

$\widehat{\pi}$	**Likelihood**
0.30	0.031
0.35	0.033
0.40	0.036
0.45	0.033

Maximum likelihood estimate: $\widehat{\pi} = 0.4$

Maximum likelihood

Maximum likelihood principle: estimate the parameters to be the values that maximize the likelihood

Steps for maximum likelihood estimation:

- *Likelihood*: For each potential value of the parameter, compute the likelihood of the observed data
- *Maximize*: Find the parameter value that gives the largest likelihood

Maximum likelihood for logistic regression

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \, Size_i \qquad \pi_i = \frac{\exp\{\beta_0 + \beta_1 \, Size_i\}}{1 + \exp\{\beta_0 + \beta_1 \, Size_i\}}$$

Observed data:

Tumor cancerous	Yes	No	No	Yes	No
Size of tumor (cm)	6	1	0.5	4	1.2

Suppose $\beta_0 = -2, \ \beta_1 = 0.5$. How would I compute the likelihood?

1) compute $\hat{\pi}_i$ for each person, using $\beta_0 = -2$, $\beta_1 = 0.5$

2) multiply probabilities!

Class Activity, Part II

https://sta214-f22.github.io/class_activities/ca_lecture_5.html

Class Activity

$$\widehat{\pi}_i = \frac{\exp\{-2 + 0.5\ Size_i\}}{1 + \exp\{-2 + 0.5\ Size_i\}}$$

Tumor cancerous	Yes	No	No	Yes	No
Size of tumor (cm)	6	1	0.5	4	1.2
$\widehat{\pi}_i$	0.731	0.182	0.148	0.5	0.197

Likelihood = $(0.731)(1-0.182)(1-0.148)(0.5)(1-0.197)$

$= 0.204$

Maximum likelihood for logistic regression

Likelihood:

- For estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, $\widehat{\pi}_i = \dfrac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\}}$
- $L(\widehat{\beta}_0, \widehat{\beta}_1) = P(Y_1, \dots, Y_n | \widehat{\beta}_0, \widehat{\beta}_1)$

1) Do by hand (tedious)
2) use R!
3) Calculus!

Maximize:

- Choose $\widehat{\beta}_0, \widehat{\beta}_1$ to maximize $L(\widehat{\beta}_0, \widehat{\beta}_1)$

> So far, we only considered a few values for β_0 and β_1. How should we check other values, to make sure our estimates actually maximize likelihood?

Computing likelihood in R

Observed data: T, T, T, T, H

- We are going to consider several different potential values for $\widehat{\pi}$:

$$0, 0.1, 0.2, 0.3, \ldots, 0.9, 1$$

- For each potential value, we will compute the likelihood:

$$L(\widehat{\pi}) = (1 - \widehat{\pi})^4(\widehat{\pi})$$

- We then see which value has the highest likelihood.
- Is this all possible values? No, but let's start here.

R code

```
# List the values for pi hat
pi_hat <- seq(from = 0, to = 1, by = 0.1)
```

"sequence"

0 0.1 0.2 ..... 0.9 1.0

(values of $\hat{\pi}$ to try)

R code

```
# List the values for pi hat
pi_hat <- seq(from = 0, to = 1, by = 0.1)

# Create a space to store the likelihoods
likelihood <- rep(0,length(pi_hat))
```

"repeat"

"

0 0 0 . . . 0 0

for storage

will fill it in as we calculate

likelihoods

Java: i = 0; i < n; i++

R code

```
# List the values for pi hat
pi_hat <- seq(from = 0, to = 1, by = 0.1)

# Create a space to store the likelihoods
likelihood <- rep(0,length(pi_hat))

# Compute and store the likelihoods
for( i in 1:length(pi_hat) ){
  likelihood[i] <- pi_hat[i]*(1-pi_hat[i])^4
}
```

for loop

Do

i = 1   pi_hat[1] * (1-pi_hat[1])^4 $= 0(1-0)^4$

i = 2   pi_hat[2] * (1-pi_hat[2])^4 $= 0.1(1-0.1)^4$

i = 3 $0.2(1-0.2)^4$

⋮

i = 11 $1(1-1)^4$

R code

```r
# List the values for pi hat
pi_hat <- seq(from = 0, to = 1, by = 0.1)

# Create a space to store the likelihoods
likelihood <- rep(0,length(pi_hat))

# Compute and store the likelihoods
for( i in 1:length(pi_hat) ){
  likelihood[i] <- pi_hat[i]*(1-pi_hat[i])^4
}
```

> Run this code in your R console. Which value of $\widehat{\pi}$ gives the highest likelihood?

Results

pi_hat	likelihood
0.0	0.00000
0.1	0.06561
0.2	0.08192
0.3	0.07203
0.4	0.05184
0.5	0.03125
0.6	0.01536
0.7	0.00567
0.8	0.00128
0.9	0.00009
1.0	0.00000

$= \frac{1}{5}$