Multinomial logistic regression

Agenda

- No homework this week
- Extra credit opportunity: department seminar
  - Dr. Mine Cetinkaya-Rundel
  - Monday, September 26, 12pm - 1pm in Kirby 120
- Today: multinomial logistic regression

Motivation

- When the response is binary, we use *logistic regression*
- What happens when the response is categorical, but has MORE than 2 categories?
- We use *multinomial logistic regression* (aka *multinomial regression*)

Motivation

Question: What is the relationship between age and contraceptive use for women in Indonesia?

Data: 1473 Indonesian couples, with variables

- Y_i = contraceptive method used (1 = no use, 2 = long-term, 3 = short-term)
- X_i = Wife's age (numeric)

The response variable

Contraception	Freq
Long	511
None	629
Short	333

- $n_{None} = 629$ (this is 42.7% of the couples)
- $n_{Long} = 511$ (this is 34.7% of the couple)
- $n_{Short} = 333$ (this is 22.6% of the couples)

The response variable

Y_i = contraceptive method used (1 = no use, 2 = long-term, 3 = short-term)

> What type of variable is Y?

Parametric model building

What are our two steps in building a parametric model?

Building a distribution

Y_i = contraceptive method used (1 = no use, 2 = long-term, 3 = short-term)

> What notation might we use for the probability of no contraceptive use?

Building a distribution

Y_i = contraceptive method used (1 = no use, 2 = long-term, 3 = short-term)

- $\pi_{i(None)} = P(Y_i = None)$
- $\pi_{i(Short)} = P(Y_i = Short)$
- $\pi_{i(Long)} = P(Y_i = Long)$

> What must be true of the three probabilities?

The Categorical distribution

Definition: Let Y_i be an **unordered** categorical variable with J levels $j = 1, \ldots, J$. Let $\pi_j = P(Y_i = j)$, where $\pi_j \in [0, 1]$ for all j, and $\sum_{j=1}^{J} \pi_j = 1$.

Then we say $Y_i \sim Categorical(\pi_1, \ldots, \pi_J)$.

- We can use this distribution as the first step in our modeling process!

What distribution does our response (contraceptive use) have?

Parametric model building

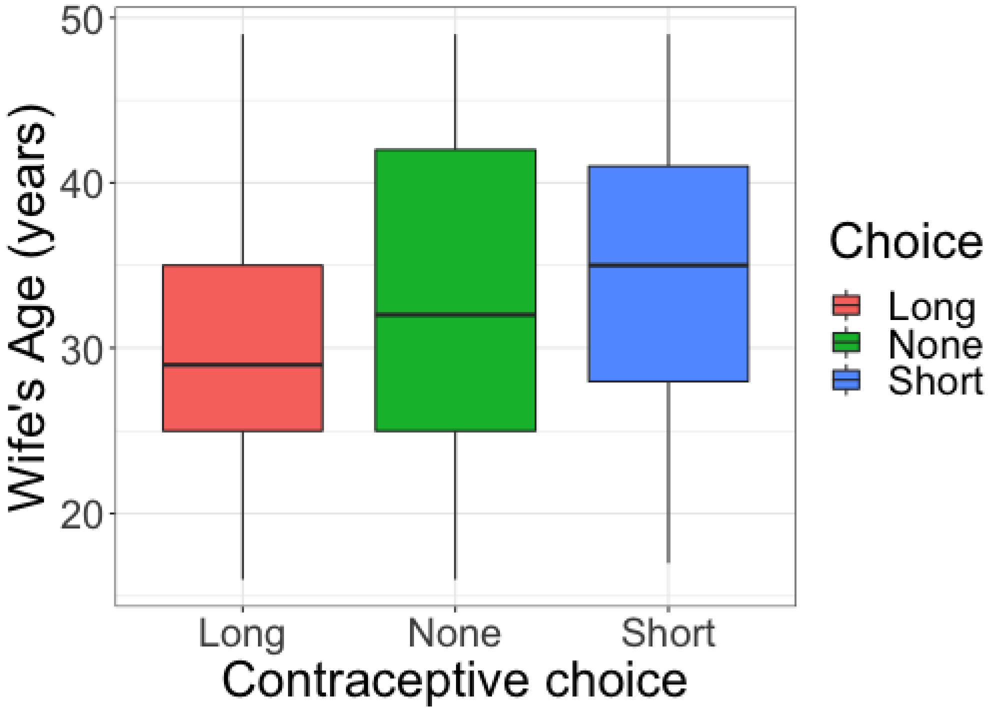
Step 1: Choose a reasonable distribution for Y

$$Y_i \sim Categorical(\pi_{i(None)}, \pi_{i(Short)}, \pi_{i(Long)})$$

Step 2: Choose a model for any parameters

- Need to relate our probabilities to $X = Age$

EDA



50
40
30
20
Wife's Age (years)
Long
None
Short
Contraceptive choice
Choice
Long
None
Short

EDA

- Boxplots show there may be some differences with age, but don't let us model the relationship
- We want something like an empirical log odds plot

> Can we use the log odds here?

Relative risk

- If Y_i is *binary*, the odds $\dfrac{\pi_i}{1-\pi_i}$ compare the probabilities of the two possible outcomes
- If Y_i has more than two outcomes, we need to generalize the odds
- The *relative risk* compares the probabilities of two potential outcomes

Relative risk of long term vs. no contraceptive use:

Relative risk of short term vs. no contraceptive use:

Example

Consider the 48 twenty-one year old wives in our data:

- Long: 23
- Short: 6
- None: 19

> For a 21 year old, what is the *empirical* relative risk of using long term vs. short term contraceptives?

Relative risk

Definition: Let Y_i be a categorical variable with J levels $j = 1, \ldots, J$. Let $\pi_j = P(Y_i = j)$. Then the relative risk of level j vs. level k is

$$\frac{\pi_{ij}}{\pi_{ik}}$$

Class activity, Part I

https://sta214-f22.github.io/class_activities/ca_lecture_14.html

Class activity

Speed Range	Slow	Good	Fast	Total
$(50, 51)$	5	1	0	6
$(51, 52)$	5	5	3	13
$(52, 53)$	6	12	2	20
$(53, 54)$	5	31	4	40

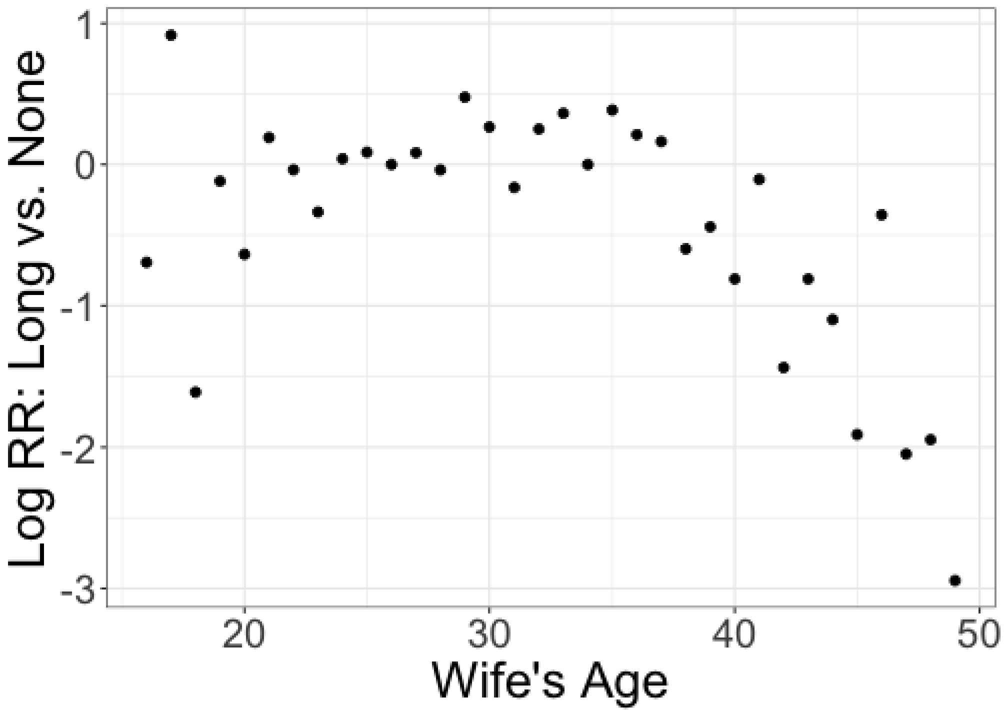
> What is the relative risk of Good vs. Slow for the $(52, 53)$ speed group?

Class activity

How would you interpret the relative risk of Good vs. Slow for the $(52, 53)$ speed group?

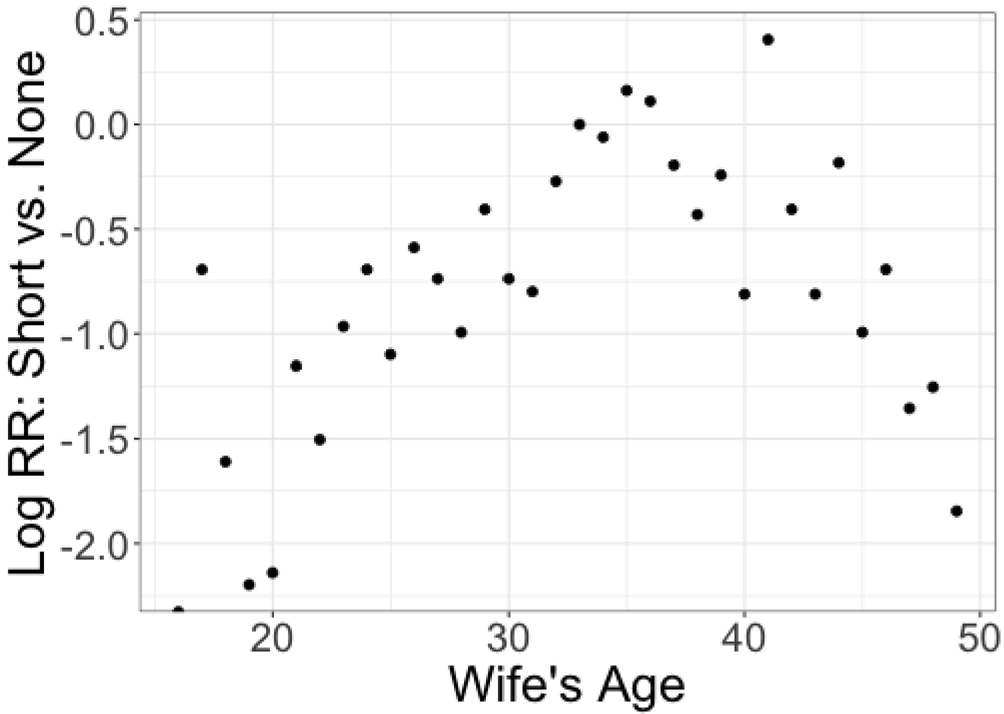
Log relative risk

Instead of modeling the log odds, we can model the *log relative risk*



Log relative risk

Instead of modeling the log odds, we can model the *log relative risk*



Multinomial regression model

Step 1: Choose a reasonable distribution for Y

$$Y_i \sim Categorical(\pi_{i(None)}, \pi_{i(Short)}, \pi_{i(Long)})$$

Step 2: Choose a model for any parameters

$$\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)} + \beta_{1(Long)} Age_i$$

$$\log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)} + \beta_{1(Short)} Age_i$$

- Pick a *reference* or *baseline* category to compare to (here it is None)

Multinomial regression model

Step 1: Choose a reasonable distribution for Y

$$Y_i \sim Categorical(\pi_{i(None)}, \pi_{i(Short)}, \pi_{i(Long)})$$

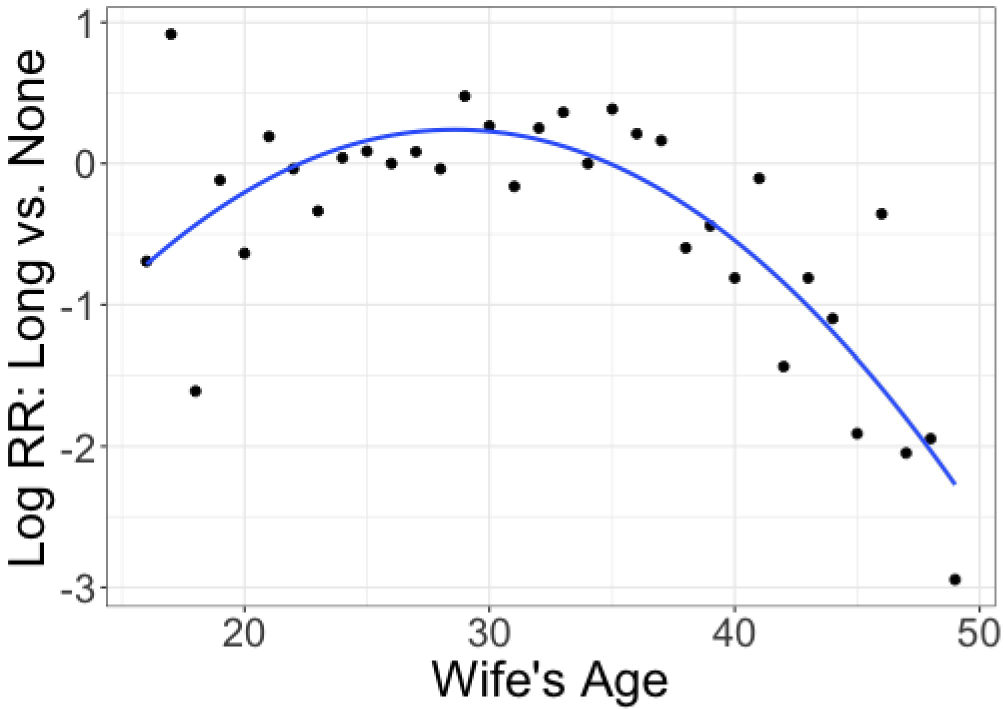
Step 2: Choose a model for any parameters

$$\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)} + \beta_{1(Long)} Age_i$$

$$\log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)} + \beta_{1(Short)} Age_i$$

> From the empirical log relative risk plots, did it look like the log relative risk was a linear function of Age?

Log relative risk



Multinomial regression model

Step 1: Choose a reasonable distribution for Y

$$Y_i \sim Categorical(\pi_{i(None)}, \pi_{i(Short)}, \pi_{i(Long)})$$

Step 2: Choose a model for any parameters

$$\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)} + \beta_{1(Long)} Age_i + \beta_{2(Long)} Age_i^2$$

$$\log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)} + \beta_{1(Short)} Age_i + \beta_{2(Short)} Age_i^2$$

Estimated model

$$\log\left(\frac{\widehat{\pi}_{i(Long)}}{\widehat{\pi}_{i(None)}}\right) = -5.07 + 0.37 Age_i - 0.0063 Age_i^2$$

$$\log\left(\frac{\widehat{\pi}_{i(Short)}}{\widehat{\pi}_{i(None)}}\right) = -8.21 + 0.46 Age_i - 0.0065 Age_i^2$$

> What is the predicted relative risk of long term vs. none for a woman age 30?

Class activity, Part II

https://sta214-f22.github.io/class_activities/ca_lecture_14.html

Class activity

> Write down the population multinomial regression model, using Slow as the reference category, and assuming that the log relative risk is a linear function of Speed.

Class activity

$$\log\left(\frac{\widehat{\pi}_{i(Good)}}{\widehat{\pi}_{i(Slow)}}\right) = -39.68 + 0.77\,\text{Speed}_i$$

> Calculate the predicted relative risk of Good vs. Slow for a race where the winning speed was 52.5 mph.

Class activity

> From this information, can you calculate the predicted *probability* that the condition was Good? If not, what more information do you need?