POLYTOMOUS MODELS: MULTINOMIAL REGRESSION

Henna Vartiainen

February 27th, 2023

OUTLINE/KEY CONCEPTS

- Basics
- Assumptions
- Different types
  - Stratified
  - Multinomial logistic regression
    - Reference category
- Model simplification
- Questions

BASICS: MULTINOMIAL MODELS

- Your dependent variable is **categorical/nominal** with at least 3 levels
  - DV is not ordered or ranked
- You want to use a variable/variables to **predict** another variable
- At least one of your predictor variables is not continuous

EXAMPLE

- Henna opens an ice cream kiosk that sells 4 kinds of ice cream
- She wants to find out which variables can predict which ice cream they will buy

EXAMPLE

- Henna opens an ice cream kiosk that sells 4 kinds of ice cream
- She wants to find out which variables can predict which ice cream they will buy

- Predictors: Age, Gender, Hair color, Height (cm), Weight (kg), Annual income ($)
- DV: type of ice cream purchased
  - Rum raisin (A), Vanilla (B), Chocolate (C), Strawberry (D)

ASSUMPTIONS

- No outliers
- Independence of observations
- No multicollinearity

ASSUMPTIONS

- No outliers
- Independence of observations
- No multicollinearity

- Does not assume normality or homoscedasticity!

OPTION 1: STRATIFIED MODEL

- Looks at each DV choice and treats it as an independent binomial logistic regression model
- You only wish to make inferences about the choice of specific categories
- Keeps one category as-is, groups the rest of them together

OPTION 1: STRATIFIED MODEL

- Looks at each DV choice and treats it as an independent binomial logistic regression model
- You only wish to make inferences about the choice of specific categories
- Keeps one category as-is, groups the rest of them together
  - Why would anyone buy rum raisin ice cream over the other options?
  - A vs [B, C, D]
  - Rum raisin vs [vanilla, chocolate, strawberry]

OPTION 1: STRATIFIED MODEL

Coefficient	Estimate	Std error	Z-value	P-value
(Intercept)				
Age	**0.239814**			.001
GenderFemale	**-2.32432**			.000
HairBlonde				.999
Height				.998
Weight				.996
Income				.064

1. Coefficient represents the increase in log odds of choosing rum raisin associated with each unit increase
2. Coefficients can be converted from log odds to odds ratio by applying the exponent function
3. Odds ratio > 1 = increase per unit; Odds ratio < 1 = decrease per unit

OPTION 1: STRATIFIED MODEL

Coefficient	Estimate	Std error	Z-value	P-value
(Intercept)				
Age	**0.239814**			.001
GenderFemale	**-2.32432**			.000
HairBlonde				.999
Height				.998
Weight				.996
Income				.064

1. Age Exp(0.239814) = 1.2710127
2. All else being equal, every additional year of age is associated with a 27% increase in the odds of choosing rum raisin over the other ice cream options
3. GenderFemale Exp(-2.32432) = 0.09784996
4. All else being equal, being female reduces the odds of selecting rum raisin by 90 %

OPTION 2: MULTINOMIAL LOGISTIC REGRESSION MODEL

- A series of binomial models comparing the **reference category** to each of the other categories
- Runs a generalized linear model on the log-odds of each category versus the reference category
  - Reference category = Rum raisin (A)
  - [A vs B], [A vs C], [A vs D]
  - [Rum raisin vs Vanilla], [Rum raisin vs Chocolate], [Rum raisin vs Strawberry]

CHANGING THE REFERENCE

- What if you want to look at other categories?
  - *E.g.* What are the odds ratios of Vanilla (B) relative to Chocolate (C)?
- Two options

  1. Change the reference category and rerun your analyses!

  2. Calculate the difference between the coefficients of B and C against A

MODEL SIMPLIFICATION

- Gradual process of elimination of variables
  - Ensures that significant variables that confound each other are not accidentally removed

1. Remove the variable with the least significant p-value
2. Run the model without the model without the variable
3. Check the coefficients;
   They should not change by more than 20-25%
4. Stop when all non-significant variables have been tested

QUESTIONS

- With both types of models, you have to run multiple significance tests
  - Controlling for Type I error?
- Assumptions – some sources say that linearity is required, some do not
- Longitudinal options?