Multinomial Logistic Regression

Dr Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Jun 11, 2023

Expected outcomes

- Understand the concept of multinomial logistic regression
- Perform multinomial logistic regression
- Perform model assessment
- Present and interpret results

Outlines

- Introduction
- Multinomial logistic regression model
- Model building:
  - Variable selection
  - Variable assessment
  - Interaction term assessment
  - Model fit assessment

Introduction

Introduction

- A regression method to model relationship between:
  - Outcome: multinomial categorical variable
  - Independent variables: numerical, categorical variables
- Multinomial i.e. multilevels, > than two levels
- Other names:
  - Discrete choice model; polychotomous/polytomous logistic regression model; baseline logit model

Introduction

- Multinomial <u>measurement scale</u>
  - Nominal categorical variable
  - No order
  - Examples:
    - Diabetic treatment: Diet control, Oral hypoglycemic agent, Insulin
    - Birth: Spontaneous vaginal delivery, Assisted vaginal delivery, Caesarean delivery
    - Cancer subtypes etc.
- Versus ordinal categorical variable → Ordinal logistic regression

Introduction

- Model the relationship

multinomial outcome = numerical predictors +
categorical predictors

Introduction

- For a three-level outcome (0, 1, 2), it can be split into two binary outcomes:

binary outcome 1 = numerical predictors + categorical predictors

binary outcome 2 = numerical predictors + categorical predictors

where, treating 0 as reference category

binary outcome 1: 1 vs 0
binary outcome 2: 2 vs 0

Multinomial Logistic Regression Model

Logit Functions

- Extending binary logistic regression, these are specified as two logit functions g_1 and g_2:

$$g_1(\boldsymbol{x}) = ln\left[\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = ln\left(\frac{p_1}{p_0}\right)$$ Compare 1 to 0

$$= \beta_{10} + \beta_{11}x_1 + \beta_{11}x_2 + \cdots + \beta_{1p}x_p$$

$$g_2(\boldsymbol{x}) = ln\left[\frac{P(Y=2|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = ln\left(\frac{p_2}{p_0}\right)$$ Compare 2 to 0

$$= \beta_{20} + \beta_{21}x_1 + \beta_{21}x_2 + \cdots + \beta_{2p}x_p$$

for a vector $\boldsymbol{x}$ comprising of p covariates and a constant term $x_0 = 1$

Odds Ratios

- Odds ratios for a covariate x_i are calculated as follows:

$$\mathrm{OR}_1(x_i) = e^{\beta_{1i}}$$

$$\mathrm{OR}_2(x_i) = e^{\beta_{2i}}$$

Conditional Probabilities

- The calculation for conditional probabilities is as follows:

$$P(Y=0\mid \boldsymbol{x})=\frac{1}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}}$$

$$P(Y=1\mid \boldsymbol{x})=\frac{e^{g_1(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}}$$

$$P(Y=2\mid \boldsymbol{x})=\frac{e^{g_2(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}}$$

Testing Significance

- Wald test, W
- Likelihood ratio test, G

Testing Significance

- Wald test, W:

$$W = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}$$

then, two-tailed P-value is $P(|z| > W)$, as W follows standard normal distribution.

- More suitable for testing a single variable.

Testing Significance

- Likelihood ratio test, *G*:

L0: Log Likelihood of model withOUT x variable(s) –
L1: Log Likelihood of model with x variable(s)

$$G = -2(L_0 - L_1) \text{ OR}$$

$$G = D_0 - D_1$$

D = Deviance =
-2 Log Likelihood of model

then, *P*-value is $P[\chi^2(\text{df}) > G]$, as *G* follows standard normal distribution, and df = difference in number of parameters between the models.

- Suitable for testing single/many variables.

Model Building

Model-building Steps

1. Variable selection
   - Univariable
   - Multivariable

   → Preliminary main effects model
2. Variable assessment
   - Linearity in logit – numerical variable, from separate binary logistic models
   - Other numerical issues
     - Small cell counts
     - Multicollinearity

   → Main effects model

Model-building Steps

3. Interaction term assessment
   - Two-way between selected variables – clinically sensible

→ Preliminary final model

4. Model fit assessment
   - Goodness-of-fit
     - Multinomial Hosmer-Lemeshow Test
   - Pseudo-R^2
   - Regression diagnostics – from separate binary logistic models

→ Final model