Conditional Logistic Regression

Dr Wan Nor Arifin

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

Last update: Jun 24, 2024

Expected outcomes

- Understand the concept of conditional logistic regression
- Perform conditional logistic regression for 1-1 and 1-M matching
- Perform model assessment
- Present and interpret results

Outlines

- Introduction
- Conditional 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: <u>binary</u> categorical variable
  - Independent variables: numerical, categorical variables, **stratum** variable
- Matching of case-control by **stratum** using variables believed to be associated to the outcome, e.g. age and gender – allows controlling for the effect of these variables
- Matched case-control study – 1:1 to 1:*M* design

Introduction

- Model the relationship

binary outcome = numerical predictors +
categorical predictors +
stratum variable

Introduction

- Analytical challenge in analyzing matched case-control:
  - 1:1 matching – two subjects per stratum
  - n case-control pairs (i.e. sample size = $2n$), p covariates
  - Need to estimate $n + p$ coefficients in this fully stratified analysis!
  - Biased, large number of parameters to be estimated
- Requires analysis by conditional likelihood estimation – to get rid of stratum specific parameters

Conditional Logistic Regression Model

Stratum-specific Logit Function

- For a stratum-specific binary logistic regression with k stratum, the logit function is given as:

$$g_k(\boldsymbol{x}) = \alpha_k + \boldsymbol{\beta}'\boldsymbol{x}$$

where α_k indicates stratum specific intercepts

- For a conditional logistic regression model, there are too many intercepts as there are many strata (case-control pairs)
- So the conditional model is developed so as to remove these intercepts

Conditional Likelihood

- Conditional likelihood for the *k*th stratum is the probability of the observed data relative to the probability of the data for all possible assignments of n_{1k} cases and n_{0k} controls to $n_k = n_{1k} + n_{0k}$ subjects

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} \mathrm{P}(\mathbf{x}_i | y_i = 1) \prod_{i=n_{1k}+1}^{n_k} \mathrm{P}(\mathbf{x}_i | y_i = 0)}{\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \mathrm{P}\left(\mathbf{x}_{ji_j} | y_{i_j} = 1\right) \prod_{i_j=n_{1k}+1}^{n_k} \mathrm{P}(\mathbf{x}_{ji_j} | y_{i_j} = 0) \right\}}$$

Conditional Likelihood

- The number of possible assignments of case status to n_{1k} subjects among n_k subjects is given by the binomial coefficient:

$$c_k = {}^{n_k}C_{n_{1k}} = \binom{n_k}{n_{1k}} = \frac{n_k!}{n_{1k}!(n_k - n1k)!}$$

Conditional Likelihood

- Then, the full conditional likelihood is given as:

$$l(\boldsymbol{\beta}) = \prod_{k=1}^{K} l_k(\boldsymbol{\beta})$$

Conditional Likelihood

- The conditional likelihood can also be simplified as:

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\sum_{j=1}^{c_k} \prod_{i_j=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_{ji_j}}}$$

- This likelihood form is similar to the one used for proportional hazards model for survival analysis. (Faraway, 2016)

Conditional Likelihood

- For 1:1 matching, this is simplified as:

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}} + e^{\boldsymbol{\beta}'\mathbf{x}_{0k}}}$$

Given values of $\boldsymbol{\beta}$, $\mathbf{x}_{1k}$ and $\mathbf{x}_{0k}$, it is the <u>probability</u> that the subject identified as the case is in fact the case, within *k* stratum

- For 1:3 matching, this is given as:

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k2}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k3}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k4}}}$$

Given values of $\boldsymbol{\beta}$, it is the probability that the subject with data $\mathbf{x}_{1k}$ is the case relative to three controls with data $\mathbf{x}_{2k}$ to $\mathbf{x}_{4k}$, within *k* stratum

Conditional vs Unconditional Likelihood

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} \mathrm{P}(\mathbf{x}_i | y_i = 1) \prod_{i=n_{1k}+1}^{n_k} \mathrm{P}(\mathbf{x}_i | y_i = 0)}{\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \mathrm{P}\left(\mathbf{x}_{ji_j} | y_{i_j} = 1\right) \prod_{i_j=n_{1k}+1}^{n_k} \mathrm{P}(\mathbf{x}_{ji_j} | y_{i_j} = 0) \right\}}$$

Conditional:
- when sample size <u>smaller than</u> number of parameters
- only estimates β coefficients

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Unconditional:
- when sample size <u>larger than</u> number of parameters
- estimates both α intercepts and β coefficients

Odds Ratios

- The odds ratio for a covariate x_i are calculated in the same way as the binary logistic regression as follows:

$$\mathrm{OR}(x_i) = e^{\beta_i}$$

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*:

Log Likelihood of model withOUT x variable(s) –
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
   - Other numerical issues
     - Discordant pairs – check for dichotomous covariates
     - Multicollinearity – check SE relative to coefficient

   → 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 – Difficult and not available in packages / software
   - Regression diagnostics – Not available in packages / software

→ Final model

References

- Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models (2nd ed.). Boca Raton, FL: CRC press.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Hoboken, NJ: John Wiley & Sons Inc.
- Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). New York, USA: Springer.