

L6.3: Conditional Logistic Regression

BIOS 6612

Julia Wrobel

Overview

Today, we cover:

- Likelihood derivation for conditional logistic regression

Optional readings:

- Agresti: 7.3, 11.2

Modeling data from matched studies

The model for matched data is

$$\text{logit}\{p_k(X_{ik})\} = \alpha_k + \beta_1 X_{ik1} + \beta_2 X_{ik2} + \cdots + \beta_p x_{ikp}$$

- $p_k(X_{ik})$ represent the probability that the i th person in the k th matched set has disease
- $k \in (1, 2, 3, \dots, K)$ indicates the current matched set
- $i \in (0, 1, 2, \dots, M + 1)$ indicates individual in the current matched set
 - $i = 0$ is the case
- \mathbf{X}_{ik} is vector of covariates for the i th person in the k th matched set
 - $\mathbf{X}_{ik} = X_{ik1}, X_{ik2}, \dots, X_{ikp}$

Constructing the conditional likelihood

How do we construct a likelihood function that allows us to find estimates for the parameters of interest, β ?

- 1. Find the conditional likelihood for the k th stratum/matched set
- 2. Combine likelihoods to obtain likelihood over all strata

Constructing the conditional likelihood

We need the probability that the individual in the study whose vector of covariates is \mathbf{X}_{0k} is actually the case, conditional on the observed covariate values $\mathbf{X}_{ik}, i = 0, 1, \dots, M$ for all individuals in the k th matched set. Define:

- $P(\mathbf{X}_{ik}|Y = 1)$ be the probability that a person with disease in the k th matched set has covariate vector \mathbf{X}_{ik}
- $P(\mathbf{X}_{ik}|Y = 0)$ be the probability that a person without disease in the k th matched set has covariate vector \mathbf{X}_{ik}

Then the joint probability that \mathbf{X}_{0k} corresponds to the case and $\mathbf{X}_{ik}, i = 0, 1, \dots, M$ to the controls is

$$P(\mathbf{X}_{0k}|Y = 1) \prod_{i=1}^M P(\mathbf{X}_{ik}|Y = 0)$$

Constructing the conditional likelihood

The probability that one of the $M + 1$ subjects in the k th matched set is the case and the remainder are controls is the union of the probabilities that:

- Person with \mathbf{X}_{0k} has disease and the rest are disease-free
- Person with \mathbf{X}_{1k} has disease and the rest are disease-free
- ...
- Person with \mathbf{X}_{Mk} has disease and the rest are disease free

This is:

$$P(\mathbf{X}_{0k}|Y = 1) \prod_{i=1}^M P(\mathbf{X}_{ik}|Y = 0) + P(\mathbf{X}_{1k}|Y = 1) \prod_{i \neq 1} P(\mathbf{X}_{ik}|Y = 0) + \cdots + P(\mathbf{X}_{Mk}|Y = 1) \prod_{i \neq M} P(\mathbf{X}_{ik}|Y = 0)$$

Which can also be expressed as

$$\sum_{l=0}^M P(\mathbf{X}_{lk}|Y = 1) \prod_{r \neq l} P(\mathbf{X}_{rk}|Y = 0)$$

Constructing the conditional likelihood

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Then, the conditional probability of interest is the ratio

$$L_k = \frac{P(\mathbf{X}_{0k}|Y = 1) \prod_{i=1}^M P(\mathbf{X}_{ik}|Y = 0)}{\sum_{l=0}^M P(\mathbf{X}_{lk}|Y = 1) \prod_{r \neq i}^M P(\mathbf{X}_{rk}|Y = 0)}.$$

This is the conditional likelihood for the k th stratum/matched set.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Constructing the conditional likelihood

We can use Bayes' Theorem to substitute:

$P(\mathbf{X}_{ik}|Y=1)$ with $\frac{P(Y=1|\mathbf{X}_{ik})P(\mathbf{X}_{ik})}{P(Y=1)}$ and $P(\mathbf{X}_{ik}|Y=0)$ with $\frac{P(Y=0|\mathbf{X}_{ik})P(\mathbf{X}_{ik})}{P(Y=0)}$

So the likelihood becomes

$$L_k = \frac{\frac{P(Y=1|\mathbf{X}_{0k})P(\mathbf{X}_{0k})}{P(Y=1)} \prod_{i=1}^M \frac{P(Y=0|\mathbf{X}_{ik})P(\mathbf{X}_{ik})}{P(Y=0)}}{\sum_{l=0}^M \frac{P(Y=1|\mathbf{X}_{lk})P(\mathbf{X}_{lk})}{P(Y=1)} \prod_{r \neq i}^M \frac{P(Y=0|\mathbf{X}_{rk})P(\mathbf{X}_{rk})}{P(Y=0)}}$$

Constructing the conditional likelihood

The factor $\frac{\prod_i^M P(X_{ik})}{P(Y=1)P(Y=0)^M}$ appears in both the numerator and denominator and will cancel out, so the the likelihood reduces to

$$L_k = \frac{P(Y = 1|\mathbf{X}_{0k}) \prod_{i=1}^M P(Y = 0|\mathbf{X}_{ik})}{\sum_{l=0}^M P(Y = 1|\mathbf{X}_{lk}) \prod_{r \neq i}^M P(Y = 0|\mathbf{X}_{rk})}$$

Constructing the conditional likelihood in terms of Betas

We have a logistic model for disease in the i th person in the k th stratum:

$$\text{logit}P(Y_i = 1|\mathbf{X}_{ik}) = \alpha_k + \beta_1 X_{ik1} + \beta_2 X_{ik2} + \cdots + \beta_p X_{ikp}$$

Now we can substitute

$$P(Y_i = 1|\mathbf{X}_{ik}) = \frac{e^{\alpha_k + \mathbf{X}_{ik}\boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{X}_{ik}\boldsymbol{\beta}}} \text{ and } P(Y_i = 0|\mathbf{X}_{ik}) = \frac{1}{1 + e^{\alpha_k + \mathbf{X}_{ik}\boldsymbol{\beta}}}$$

$\leftarrow 1 - P(Y_i = 1|\mathbf{X}_{ik})$

Constructing the conditional likelihood in terms of Betas

I'm skipping a couple algebraic steps, but you end up with a stratum specific likelihood where the intercept terms cancel. The intercept terms are said to have been "conditioned out". Effects of matching variables cannot be estimated!

$$L_k(\beta) = \frac{e^{\alpha_k + \mathbf{X}_{ik}\beta}}{\sum_{l=0}^M e^{\alpha_k + \mathbf{X}_{ikl}\beta}} = \frac{\cancel{e^{\alpha_k}} e^{\mathbf{X}_{ik}\beta}}{\cancel{e^{\alpha_k}} \sum_l e^{\mathbf{X}_{ikl}\beta}}$$

Overall likelihood is the product of the likelihood for each stratum:

$$L(\beta) = \prod_{k=1}^K L_k(\beta)$$

Conditional likelihood for matched case-control data

Since we can't estimate the intercepts, disease probabilities $p(\mathbf{X}_i \mathbf{k})$ are not estimable either.

- We can still include interaction terms between matching variables and exposure factors to determine whether the effect of the exposure variable is consistent across different values of the matching variable
- The log odds ratios are independent of the intercepts, so they CAN be estimated

This conditional likelihood behaves much like an ordinary likelihood!

- Likelihood is maximized to get estimates of the coefficients
- Can use likelihood ratio tests to compare nested models and conduct hypothesis tests for parameter estimates