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

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

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





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

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,...,M for all individuals in the kth matched set. Define:

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

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

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

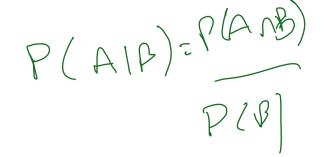
The probability that one of the M+1 subjects in the kth matched set is the case and the remainder are controls is the union of the probabilities that:

- 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

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

Which can also be expressed as

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



Then, the conditional probability of interest is the ratio

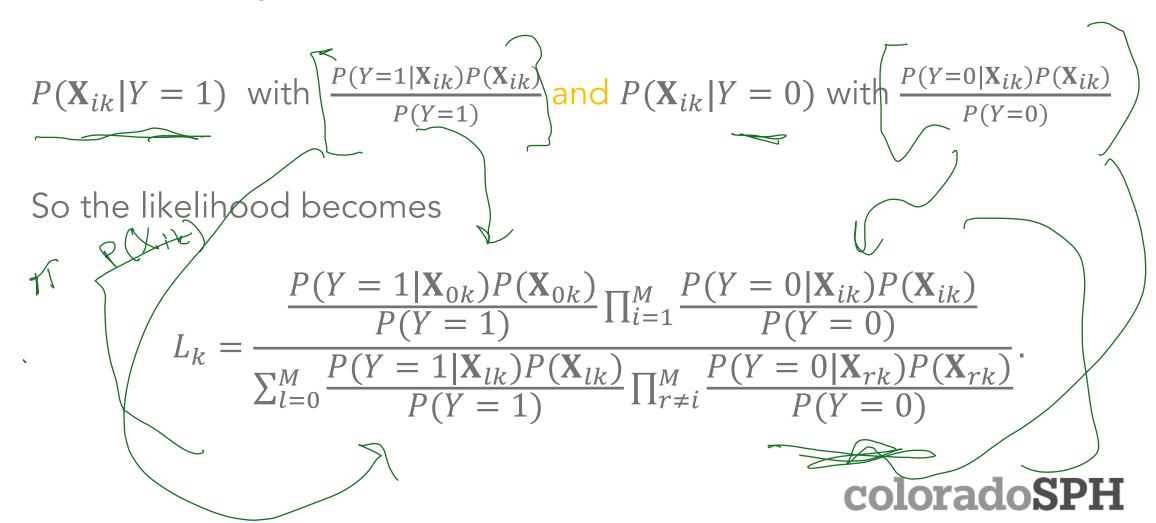
$$L_{k} = \frac{P(\mathbf{X}_{0k}|Y=1) \prod_{i=1}^{M} P(\mathbf{X}_{ik}|Y=0)}{\sum_{l=0}^{M} P(\mathbf{X}_{lk}|Y=1) \prod_{r\neq i}^{M} P(\mathbf{X}_{rk}|Y=0)}.$$

This is the conditional likelihood for the kth stratum/matched set.

P(AIB) = P(BIN)P(A) ihood P(B)

Constructing the conditional likelihood

We can use Bayes' Theorem to substitute:



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 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 ith person in the kth stratum:

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

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

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!

$$L_{k}(\beta) = \frac{e^{\alpha_{k} + \mathbf{X}_{ik} \beta}}{\sum_{l=0}^{M} e^{\alpha_{k} + \mathbf{X}_{ikl} \beta}} = \frac{e^{\alpha_{k} + \mathbf{X}_{ikl} \beta}}{e^{\alpha_{k} + \mathbf{X}_{ikl} \beta}}$$

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

1 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

