www.epilepsy.va.gov/Statistics

Statistics in Evidence Based Medicine (2014)

Lecture 4: Special Cases of Logistic Models

Rizwana Rehman, PhD
Regional Statistician
Southeast Epilepsy Center of Excellence
Durham VA Medical Center, Durham NC
Rizwana.Rehman@va.gov
(919)286-0411 ext: 5024

Epilepsy Center of Excellence
SOUTHEAST

Course Outline

Understanding logistic regression in five lectures

Difference between relative risk and odds ratio ✓,

marginal and conditional odds ratios, ✓

terminology and interpretation of logistic regression

Suggested Book: Logistic Regression A Self-Learning Text by Kleinbaum & Klein

Third Edition Springer

Today's Lecture

- Review of previous lectures
- Model fit statistics
  - Significance of coefficients in the logistic regression
- Special cases
- Summary

Odds (*o*)

- The odds (*O*) of an event are the likelihood of an event occurring **divided** by the likelihood of event not occurring
- For a 2×2 table divide the counts of occurrence of an event by counts of non occurrence of an event

Odds can lie between zero and infinity

Odds are ratios of proportions

Relationship between Odds and Probability

- To calculate the odds (o) from Probability (p)

$$\text{Odds} = \frac{p}{1-p}$$

- To calculate the probability from Odds

$$\text{Probability} = \frac{o}{1+o}$$

Odds Ratio

- Ratio of two odds is called an odds ratio.
- It is a measure of association between two variables. Odds Ratio=1 means that there is no association between two variables.
  - Example: Association between Heart Disease (HD) and Blood Pressure (BP)
    - Compute odds of HD among BP group
    - Compute odds of HD among No PB group
    - Divide the odds to get the Odds Ratio

Odds Ratios and Relative Risk

- Relative risk is the ratio between two probabilities
- Odds Ratio can’t be interpreted as relative risk for a common disease in a case control study
- For a rare disease odds ratio and relative risk are approximately equal

Logistic Regression and Odds

- Logistic regression is used for a (binary) outcome variable
- Logistic regression applies logit transformation to the dependent variable to produce a linear relationship.

Estimated natural log (Odds)

$$\text{Logit (Y)} = \log_e\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

- If coefficient b_1 is positive, then large values of X are associated with large value of logit of Y and small values of X are associated with small values of logit of Y
- If coefficient b_1 is negative, then large values of X are associated with small values of logit of Y and small values of X are associated with large values of logit of Y

Interpretation of b_1 for a Binary X

- X coded as 0, 1
- $\log_e$(Odds Ratio) = b_1

The estimated regression coefficient b_1 is the natural log of the odds ratio. This is the change in log odds of Y when X changes from 0 to 1.

Odds Ratio=change in log odds $= e^{b_1}$

Special Case: Binary (0,1) Xs

- Suppose we have many independent variables in a logistic model
- We can obtain an **adjusted odds ratio** for each (0,1) **X** variable in the logistic model by exponentiating the coefficient corresponding to that variable

Example: Special Case

- Outcome variable Coronary Heart Disease CHD (0,1)
- X_1=Catecholamine level **CAT (0=low,1=high)**
- X_2=**Age** in years (continuous)
- X_3=**ECG (0=normal,1=abnormal)**

Logit(Y)=0.1023+0.652×CAT+0.029×AGE+0.342×ECG

Example Continued

- $\mathbf{b_1}$ for CAT (O, 1) =0.652
- b_2 for AGE Continuous=0.029
- b_3 for ECG (0,1) =0.342

The odds of CHD for people with a high catecholamine level were $e^{0.652}$= 1.919 times the odds for people with low catecholamine level while controlling for Age and ECG.

Example Continued

- b_1 for CAT (O, 1) =0.652
- b_2 for AGE Continuous=0.029
- $\mathbf{b_3}$ for ECG (0,1) =0.342

The odds of CHD among people with an abnormal EEG were $e^{0.342}$= 1.408 times the odds for people with normal EEG while controlling for Age and Catecholamine level.

Example Continued

- b_1 for CAT (O, 1) =0.652
- $\mathbf{b_2}$ for AGE Continuous=0.029
- b_3 for ECG (0,1) =0.342

With a one year change in age the log odds of CHD change by 0.029 adjusting for Age and Catecholamine level

Model Fit: Statistical Significance of Coefficients

- **Specify the null Hypothesis: A coefficient b_i is zero**
- Decide the significance level
- Compute a test statistic
- Compare results with a chi square distribution
- If the value of test statistic is greater than the value of chi square, then reject the null hypothesis; **contribution of variable is significant**

For **complete understanding of p value and confidence intervals check "Understanding P Value and Confidence Intervals " from 2013 lecture series on statistics website**

Model Fit: The Likelihood Ratio Test

- Inclusion of an explanatory variable in the model tells us more about the outcome than a model which does not include that variable.
- Based upon likelihood functions
- Measures the discrepancy between the observed value and predicted values

Likelihood Ratio Test

- Logit (Y)= b_0------------------------1
- Logit (Y)= $b_0+b_1X_1$----------------2

Is $b_1 = 0$

- For each model calculate likelihood function's values
- Take the difference 1-2; Difference is called likelihood ratio statistic
- Compare with a chi square distribution with 1 degree of freedom
- If log likelihood ratio is bigger than Chi Square, then reject null hypothesis

Example: Smoking and Lung Cancer

Male Lung Cancer & Smoking (Doll and Hill 1950)

	Lung cancer (Case)	Control
Smokers	647	622
Non-smokers	2	27

$$\text{Odds Ratio} = \frac{647 \times 27}{2 \times 622} = 14.04$$

The odds of lung cancer in smokers were 14 times the odds of lung cancer in non-smokers

Logistic Regression for Association between Lung Cancer and Smoking

$$\textbf{Logit(Y)} = \log_e\left(\frac{p}{1-p}\right) = -2.6025 + 2.6419 \times \textbf{Smoking}$$

- 2.6419 is the increment to the log odds for smokers
- Moving from non smokers to smokers increases the log odds of lung cancer by 2.6419

$\log_e$(odds ratio)=2.6419

- Estimated odds of lung cancer among smokers are $e^{2.6419}$=14.04 times the odds of lung cancer among non smokers

Using Likelihood Ratio Test

- We first consider intercept only model

  Logit (Y)= b_0 (intercept only)----1799.41
- Then we consider model with smoking
- Logit (Y)= $b_0+b_1\times$ Smoking----1773.27
- Difference= 26.14
- Compare with a Chi Square distribution with one degree of freedom =3.84 at 95% significance level
- P value < 0.001 <0.05 reject hull hypothesis
- Inclusion improves the model. **Smoking is associated with Lung cancer**

Using likelihood Ratio Test for Overall Evaluation of a Logistic Model

- **A logistic model is better fit if it is an improvement on intercept only model**
- We have k predictors variables

Logit (Y)= $b_0+b_1X_1 + b_1X_1$------$+b_kX_k$ -----------2

- **Intercept only** Logit (Y)= b_0 ----------1
- **Compute likelihood functions for both and proceed as described before**

Model Fit: The Wald Test

- Null Hypothesis: b_i is zero
- For a coefficient b_i compute the Wald test statistic $[b_i/\text{standard error of } b_i]^2$
- Compare with a chi square distribution
- If the value of Wald test statistic is greater than the chi square, then reject null hypothesis

Wald Test: Results for smoking

Null Hypothesis: b_1= 2.6419 is zero

- Wald statistic=$[2.6419/0.7349]^2$= 12.922
- Compare with chi square=3.84 with one degree of freedom
- P = 0.0003 < 0.05
- Reject null hypothesis

Confidence Intervals

- Difficult but possible for likelihood ratio test!
- We can compute Wald confidence intervals for coefficients

$$b_i \pm 1.96 \times \text{standard error } b_i$$

- By taking the exponents of the lower and upper limits of confidence intervals we can obtain a confidence interval of the **odds ratio**

Computing Wald Confidence Intervals

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6025	0.7328	12.6139	0.0004
smoke	1	2.6419	0.7349	12.9229	0.0003

Smoking coefficient=2.6419

Lower limit =2.6419-1.96×0.7349= 1.2015

Upper limit= 2.6419-1.96×0.7349 = 4.0823

Odds Ratio =1.515

Lower limit= $e^{1.2015}$= 3.325

Upper limit= $e^{4.0823}$=59.282

Computer (sas) Output

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
smoke	14.04	3.325	59.281

Special Case: Which test to use?

- Wald test can also be used for an overall evaluation of model
- In most situations both tests yield similar inferences
- **When both provide conflicting results then likelihood ratio test is more accurate**
- While reading a paper pay attention to the test used and significance of coefficients and overall fit of model

Special Case: Interaction

- Suppose a research paper reports

$\text{Logit}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$

- Also suppose that b_3 is significant
- Above is an example of significant statistical interaction
- Regression coefficients of a variable correspond to the change in log odds and its exponentiated form provides odds ratio

Interaction (Effect Modification)

- The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variable.
- Distinct from confounding

$$\text{logit}(Y) = b_0 + b_1X_1 + b_2X_2$$

- Confounding can be adjusted in statistical analysis by estimating one common odds ratio
- Confounding is common; interaction is rare

Interpretation of Coefficients

$$\text{Logit }(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

- Interpretation of $b_{1,}$ b_2 changes
- Depends upon nature of predictors X_1 and X_2 (continuous, dichotomous or ordinal)
- Detailed discussion is out of scope of our lecture

Summary

What have we learnt

- Use of logistic regression for binary data
- Meaning of confounding and statistical interaction
- Interpretation of coefficients
- Checking statistical significance of coefficients
- Checking overall fit of logistic model
- Beware of interactions

www.epilepsy.va.gov/Statistics

Questions/Comments

Rizwana.Rehman@va.gov

(919) 286-0411 ext: 5024

Thank you for being patient !