www.epilepsy.va.gov/Statistics

Statistics in Evidence Based Medicine (2014)

Lecture 5: Logistic Regression for Matched Data

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

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, ✓
matched data analysis

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

Today's Lecture

- Review of previous lectures
- Continuation of previous topic
  - Fitness of good statistics
  - How many independent variables we can add
  - Assumptions of logistic regression
- Evaluation of research papers
- Matched data analysis

What Have We Learnt So Far?

- Comparison of odds ratio and relative risk (Lecture 1)
- Meaning of confounding and statistical interaction (Lecture 2)
- Introduction of logistic regression (Lecture 3)
  - Interpretation of coefficients in term of odds ratios
  - Probabilities can be computed from odds
- Checking statistical significance of coefficients (Lecture 4)
- Checking overall fit of logistic model(Lecture 4)

Review Overall Fit of Logistic Model

Null Hypothesis: A smaller model with intercept only is better than a larger model with independent variables

- Likelihood Ratio Test
- Wald Test
- Similarity between the two test: **A larger difference from critical value (low p value) means reject the null hypothesis**
- Other tests

Goodness of Fit Statistics

How well does the final model fit the data against actual outcomes?

- Descriptive statistics
  - Pseudo R^2
  - c index
- Inferential tests
  - Chi Square
  - Hosmer &Lemeshow

Descriptive Statistics R^2

- How well we can predict the dependent variable from the independent variables
- Similar to Pearson R^2 for linear regression: Proportion of variance explained by the model
- Different versions
- Use as a supplementary help

c Index(Basic Idea)

- Fit a logistic model and compute estimated probabilities from the model
- If the estimated probability is higher than a cutoff value say 0.5 consider it a yes, otherwise consider it a no.
- Make a two by two table for observed positives and negatives.

c Index (Basic Idea)

	Observed positive	Observed negative
Predicted positive	a	b
Predicted negative	c	d

Sensitivity=$\frac{a}{a+c}$,   Specificty=$\frac{d}{b+d}$

Higher sensitivity and specificity indicate a better fit

c Index

- Extend the idea of two by two table
- Consider many tables with different cutoff values
- Compute sensitivity and specificity for every table
- Draw a graph of sensitivity vs. 1-specificity known as ROC curve
- **Area under the ROC curve provides an overall measure of fit of the model**

c Index

- c provides an estimate of the probability that a randomly selected pair (one true positive and one true negative) will be correctly ordered by the test
- By correctly ordered we mean that a true positive will have a higher predicted probability of the event compared to a negative subject
- Higher c statistic means a better fit

Goodness of Fit Inferential Tests

- Null Hypothesis: Model is correct
- Alternative Hypothesis: Model is not correct
- **A p value greater than 0.05 means that a model is a good fit. A low p value means reject the model**

Chi Square Test for Grouped Data

- Compute residuals

  r_i=actual outcome-predicted probability

- Standardize residuals
- Compute a test statistic from standardized residuals

Chi Square=sum of standardized residuals

- Compute probability of chi square statistic
- P value > 0.05 means can't reject the null hypothesis, model appears reasonable
- Another similar test is deviance test

Hosmer & Lemeshow Test for Individual Level Data

- Make equal groups of cases based on values of the predicted probabilities
- Compute observed and expected number of events and non events in each group
- Compute a test statistic
- Requirements
  - Needs a large sample size
  - None of the group can have an expected value less than 1

Assumptions for Logistic Regression

- Events are independent
- Linear relationship between logit Y and independent variable X

$$\text{Logit}(Y) = \log_e(\text{odds of } Y) = \alpha + \beta X$$

  - Linear relationship can be checked graphically

Minimum Number of Events per Variable

- We can't add as many independent variables as we want without considering the sample size
- A rule of thumb is that there should be at least 10 events per variable in the model.

Evaluating a Research Paper

What can you understand and evaluate?

1. Why logistic regression was used ✓
2. Interpretation of coefficients in logistic model ✓
3. Statistical accuracy of coefficients with hypothesis tests and confidence interval ✓
4. Overall fit of the model ✓
5. Goodness of fit of the model ✓

Matched Pair Analysis

We can match on variables like age, sex, race etc.

Matching leads to more efficient statistical results

- Prospective and Case control study
  - Each matching pair is called a strata
- Crossover trial
  - Strata consists of two binary measurements on the same subject

Conditional Logistic Regression

- There may be many independent variables other than primary variable of interest which influence the outcome
- Conditional logistic regression model allows to add not matched independent variables in the model

Example of the Endometrial Cancer Data

- Subset of data from Breslow and Day Case Control Study

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect069.htm

- Outcome variable endometrial cancer (yes, no), prognostic factors gall bladder disease (yes, no variable) and hypertension (yes, no variable)
- The goal of the case-control analysis was to determine the relative risk of endometrial cancer for gall bladder disease, controlling for the effect of hypertension.

Only Gall Bladder Status as Independent Variable

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9555	0.5262	3.2970	0.0694

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gall	2.600	0.927	7.293

Gall Bladder and Hypertension as Covariates

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9704	0.5307	3.3432	0.0675
Hyper	1	0.3481	0.3770	0.8526	0.3558

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gall	2.639	0.933	7.468
Hyper	1.416	0.677	2.965

References for Further Studies

- Logistic Regression in the Medical Literature

  Bagley, White & Golomb

  http://www.aliquote.org/cours/2012_biomed/biblio/Bagley2001.pdf

- An introduction to Logistic Regression Analysis and Reporting

  Peng, Lee & Ingersoll

  http://sta559s11.pbworks.com/w/file/fetch/37766848/IntroLogisticRegressionPengEducResearch.pdf

Final Points

- Interpretation of logistic regression model depends upon the coding scheme
- Different statistical packages may yield different results
- **Not covered in the course:** Logistic regression for multinomial and ordinal response variable and correlated data
- For binary outcome variable logistic regression is NOT the only choice

www.epilepsy.va.gov/Statistics

Questions/Comments

Rizwana.Rehman@va.gov

(919) 286-0411 ext: 5024

Thank you for being patient !