Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 11:
Matched Sampling (Correlated Response);
Ordered Logistic Regression; Current Status Data

November 20, 2014

1

Lecture Outline

- Case-Control Studies
  - Matching
  - Conditional Logistic Regression
- Ordered Logistic Regression
- Analysis of Current Status Data

2

Case-Control Studies

3

Common Study Designs

- Cross-sectional studies (surveys)
- Cohort studies
  - Interventional studies
- Case-control studies

4

Cross-sectional Studies

- Surveys of subjects sampled from a population
- Constrained sample sizes
  - None: "Poisson sampling"
  - Total N: "Multinomial sampling"
- Real or event time
  - "Real time" = "calendar time"
  - "Event time" = when some event happens
    - birth, marriage, diagnosis, treatment, death
- Efficient for examining
  - Common outcomes and risk factors
  - Associations (not cause and effect)
  - Can estimate prevalence of risk factors and outcomes
    - Overall and within groups

5

Cohort Studies

- Groups defined by risk factor
  - Identified prospectively or retrospectively
- Constrained sample sizes
  - Risk factor strata: "Binomial sampling"
- Followed longitudinally for outcome(s)
  - Prospectively into the future, or
  - Retrospectively since some defining event
    - e.g., since being born in a particular hospital in a particular year
- Efficient for examining
  - Common outcomes
  - Many different outcomes for same exposure
  - Associations (not cause and effect)
  - Estimate incidence within risk factor groups
    - Cannot estimate prevalence of risk factor

6

Interventional Studies

- Subjects assigned to some intervention
  - Ideally controlled, randomized
- Followed longitudinally for some outcome
  - So a special case of a cohort study
- Efficient for examining
  - Common outcomes
  - Cause and effect

7

Case-Control Studies

- Groups defined by some outcome event
  - E.g., death, diagnosis of disease
- Constrained sample sizes
  - Outcome strata: "Binomial sampling"
- Characterize prior exposures
  - Longitudinal study into the past
  - How to handle time in exposure?
    - ever / never exposed, cumulative exposure, time since exposure
- Efficient for examining
  - Rare outcomes
  - Many different risk factors for same outcome
  - Associations (not cause and effect)
  - Estimate prevalence of exposure by disease
    - Cannot estimate prevalence of disease

8

Detecting Associations

- Consider random variables
  - *D* be the disease state with values *(d_1, d_2, …)*
  - *R* be a risk factor with values *(r_1, r_2, …)*
- We consider the "statistical questions" that can be answered by study designs
  - Cross-sectional
  - Cohort
  - Case-control

9

Detecting Cause and Effect

- Demonstrated rigorously only through randomized studies
  - An association between variables in a randomized study
- A characteristic of study design
- There is nothing in the data that can distinguish between randomized studies and observational studies

10

Detecting Associations

- θ is some summary measure of a probability distribution
  - Mean, median, geometric mean, proportion, odds, hazard, …
  - Categorical data: Proportion (mean) or odds
- Cross-sectional surveys show
  - $\theta(D \mid R = r_1) \neq \theta(D \mid R = r_2)$, OR
  - $\theta(R \mid D = d_1) \neq \theta(R \mid D = d_2)$
- Cohort studies sample within risk factors so only consider
  - $\theta(D \mid R = r_1) \neq \theta(D \mid R = r_2)$
- Case-control studies sample within disease so only consider
  - $\theta(R \mid D = d_1) \neq E(\theta \mid D = d_2)$

11

Binary Response: Regression Models

- **Cohort studies** sample within risk factors so for arbitrary summary measure of binary outcome distribution consider

$$D_i \mid R_i, W_{1i}, W_{2i} + \ldots \sim F_D(y; \theta_i) \text{independent}$$

$$g\left(\theta_i\left(D_i \mid X_i, \vec{W}_i\right)\right) = \beta_0 + \beta_1 \times R_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

- **Case-control studies** sample within disease so for arbitrary summary measure of binary outcome distribution

$$R_i \mid D_i, W_{1i}, W_{2i} + \ldots \sim F_R(y; \theta_i) \text{independent}$$

$$g\left(\theta_i\left(R_i \mid D_i, \vec{W}_i\right)\right) = \beta_0 + \beta_1 \times D_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

12

Special Case: Odds Ratios

- In the case where θ corresponds to the odds, with a log link we can (in some sense) blur the distinction between analyses of cohort studies and case-control studies

$$OR_{R|D} = \frac{\Pr(R_i=1|D_i=1,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0|D_i=1,\vec{W}_i=\vec{w}_i)} \Big/ \frac{\Pr(R_i=1|D_i=0,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0|D_i=0,\vec{W}_i=\vec{w}_i)}$$

$$= \frac{\Pr(R_i=1,D_i=1,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0,D_i=1,\vec{W}_i=\vec{w}_i)} \Big/ \frac{\Pr(R_i=1,D_i=0,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0,D_i=0,\vec{W}_i=\vec{w}_i)}$$

$$= \frac{\Pr(R_i=1,D_i=1,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=1,D_i=0,\vec{W}_i=\vec{w}_i)} \Big/ \frac{\Pr(R_i=0,D_i=1,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0,D_i=0,\vec{W}_i=\vec{w}_i)}$$

$$= \frac{\Pr(D_i=1|R_i=1,\vec{W}_i=\vec{w}_i)}{\Pr(D_i=0|R_i=1,\vec{W}_i=\vec{w}_i)} \Big/ \frac{\Pr(D_i=1|R_i=0,\vec{W}_i=\vec{w}_i)}{\Pr(D_i=0|R_i=0,\vec{W}_i=\vec{w}_i)}$$

$$= OR_{D|R}$$

13

Special Case: Logistic Regression

- We can compare the two logistic regression models that might be considered:

$$\log\left(\frac{\Pr(R_i=1|D_i,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0|D_i,\vec{W}_i=\vec{w}_i)}\right) = \beta_0 + \beta_1 \times D_i + \beta_2 \times w_{1i} + \cdots$$

$$\log\left(\frac{\Pr(D_i=1|R_i,\vec{W}_i=\vec{w}_i)}{\Pr(D_i=0|R_i,\vec{W}_i=\vec{w}_i)}\right) = \alpha_0 + \alpha_1 \times R_i + \alpha_2 \times w_{1i} + \cdots$$

- Correspondences between these models must account for the relative likelihood that a case is sampled and a control is sampled

14

Special Case: Logistic Regression

- Let Z indicate that a given subject was sampled (so Z=1 for all subjects in our sample)

$$\log\left(\frac{\Pr(D_i=1|R_i,\vec{W}_i,Z_i=1)}{\Pr(D_i=0|R_i,\vec{W}_i,Z_i=1)}\right)$$

$$= \log\left(\frac{\Pr(D_i=1,R_i,\vec{W}_i,Z=1)}{\Pr(D_i=0,R_i,\vec{W}_i,Z=1)}\right)$$

$$= \log\left(\frac{\Pr(Z=1|D_i=1,R_i,\vec{W}_i)\Pr(D_i=1,R_i,\vec{W}_i)}{\Pr(Z=1|D_i=0,R_i,\vec{W}_i)\Pr(D_i=0,R_i,\vec{W}_i)}\right)$$

$$= \log\left(\frac{\Pr(Z=1|D_i=1,R_i,\vec{W}_i)}{\Pr(Z=1|D_i=0,R_i,\vec{W}_i)}\right) + \log\left(\frac{\Pr(D_i=1|R_i,\vec{W}_i)}{\Pr(D_i=0|R_i,\vec{W}_i)}\right)$$

15

Special Case: Correspondences

- Only the intercept differs between the two models
  - And the difference relates to the probability of being sampled for the cases and the controls

$$\log\left(\frac{\Pr(R_i=1|D_i,\vec{W}_i=\vec{w}_i)}{\Pr(R_i=0|D_i,\vec{W}_i=\vec{w}_i)}\right) = \beta_0 + \beta_1 \times D_i + \beta_2 \times w_{1i} + \cdots$$

$$\log\left(\frac{\Pr(D_i=1|R_i,\vec{W}_i=\vec{w}_i)}{\Pr(D_i=0|R_i,\vec{W}_i=\vec{w}_i)}\right) = \alpha_0 + \alpha_1 \times R_i + \alpha_2 \times w_{1i} + \cdots$$

$$\alpha_0 = \log\left(\frac{\Pr(Z=1|D_i=1,R_i,\vec{W}_i)}{\Pr(Z=1|D_i=0,R_i,\vec{W}_i)}\right) + \beta_0$$

$$\alpha_j = \beta_j \quad j \geq 1$$

16

Special Case: Implications

- With case-control sampling, we may fit logistic regression modeling the odds of event conditional on the risk factor
- HOWEVER:
  - We should only use those models for inference about associations
  - We cannot use those models to provide estimates of event probabilities or odds for any specified levels of the covariates
- When trying to understand or describe the fitted model, it should be shown as the *odds ratios* relative to some defined baseline group
  - We should not display the probabilities as we have done with cohort or cross-sectional sampling

17

And Recall

- The previous distinctions refer to making estimates of the scientific magnitude of the association
  - Estimating θ within groups
  - Contrasting θ across groups
- Recall, however, that if all we want to do is test for associations, we do not have to worry so much about the sampling scheme
  - We will get fairly similar p values by analyzing D | R or R | D no matter whether sampling conditioned on disease or risk factor

18

Matching

19

Scientific Questions

- Most times: Comparing distribution of response across groups defined by predictor of interest
  - Cohort: Distribution of disease across exposure groups
  - Case-control: Distribution of exposure across disease groups
  - Cross-sectional: Either of the above
- Occasionally, other variables also need to be considered because comparison is different in strata
  - Detecting effect modification
  - Answering question within each subgroup
  - Gaining precision
- Quite often other variables are considered because we want to detect effect of POI across groups that are “otherwise similar”
  - Confounding: Groups being compared differ in other ways
  - Precision: Less variable response if we control for other variables

20

Covariate Adjustment

- Response: Distribution is summarized within groups
- Predictor of interest (POI): Primary definition of groups to be compared
  - May be modeled as several covariates
  - E.g.: Smoking modeled as binary ever smoked, current intensity of smoking, pack-years history of smoking, years since quit
  - E.g.: Cholesterol modeled as linear and quadratic term
- Covariates other than the POI are included in the model as
  - Effect modifiers
  - Confounders
  - Precision variables

21

Matching

- An alternative approach is to design a study in such a way as to make the sampled individuals more comparable
- Such matching can be expected to
  - Reduce confounding, because there will be no association between the POI and any variables used in matching
  - Increase precision, because there will be more homogeneity in each group
- It is useful to consider cases in which subjects are matched by
  - only fixed effects (e.g., sex, age, smoking), or
  - some random effects (e.g., family, neighborhood)

22

Linear Regression

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$$

  - β_1 = Diff in mean Y for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)

$$\text{Adjusted Model}: \quad E\left[Y_i \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

23

Relationships: True Slopes

- The slope of the unadjusted model will tend to be

$$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if
  - $\rho_{XW} = 0$ (X and W are truly uncorrelated), OR
  - $\gamma_2 = 0$ (no association between W and Y after adjusting for X)

24

Increased Precision

- Difference in means across groups can be estimated by mean difference
- Comparisons within a pair of positively correlated subjects leads greater precision
  - Adjusting for a highly predictive random effect
    - Correlation of matched measurements near 1

Variance of difference with matched samples :

$$Var(W - X) = Var(W) + Var(X) - 2\rho\sqrt{Var(W)\,Var(X)}$$

Variance of difference with independent samples :

$$Var(W - X) = Var(W) + Var(X)$$

25

Matched Samples: RCT

- Many studies make use of matched samples to study associations
  - E.g., cross-over studies in which each subject receives both treatments *in random order*
  - E.g., "split-plot" designs in which each subject receives both treatments in different locations *(randomized)*
    - Eye disease, skin disease
  - E.g., matched subjects in which one of each pair receives a treatment *(randomized)*
    - Twin studies, matched communities

26

Matched Sample: Case Control

- A disease registry might be used to identify all eligible cases of incident disease
- Sampling from the population might be used to identify nondiseased subjects who are similar with respect to
  - sex
  - age
  - race
  - neighborhood
  - propensity to be seen at the same hospital
  - …

27

Collapsing Data on Clusters

- So far: We have primarily considered analyses in which we would adjust for confounders and precision variables that are fixed effects
- When we take several measurements on a cluster, we often combine them in order to make comparisons only to the other subjects that are most comparable
- It is also possible to combine the approaches
  - Make direct comparisons on each cluster, and
  - Borrow information on some effects across clusters

28

Conditional Logistic Regression

29

Matching: Fixed Effects

- Frequently case-control studies will merely match on age, sex, race, and other fixed effects
- Frequency matching: Ensure same general *marginal (univariate)* distribution of covariates for the cases and controls
  - Same age distribution, sex distribution, etc.
  - But may not be same distribution of sex-age combinations
- Individual matching: Match each case to a control that matches on all covariates
  - Matches on fixed effects and their interactions

30

Matching: Fixed Effects - Analysis

- Having matched on variables thought to be important, we have avoided confounding
  - However, failure to adjust for important (precision) variables will attenuate the OR for the predictor of interest
  - Analyze a marginal effect rather than a within-cluster effect
- When matching on a few fixed effects *(and modeling them with relatively few parameters)*, it is most common to analyze the data using logistic regression
- Covariates that measure the matching variables are included in the model
  - Various forms of inclusion and interactions can be considered
    - Continuous terms or dummy variables as appropriate

31

Logistic Regression: Special Issues

- There is a problem, however, that arises when the data becomes too sparse to estimate parameters in logistic regression
- For instance, modeling age by dummy variables with or without the sex interaction may yield strata with very few subjects
- Contrasts across exposure within those strata may yield many noninformative strata
  - There is no information about the odds ratio in a stratum that has all subjects exposed or unexposed
  - There is no information about the odds ratio in a stratum that has the same outcome (all events or non-events) in both the exposed and unexposed

32

Options That We Can Consider

- Logistic regression with dummy variables for matching variables
  - Fixed effects
  - Conditioning on random effects
- Logistic regression that does not adjust for matching variables
  - Possible attenuation of within-cluster effects
  - Effect on power more muted
- Stratified analyses
  - Conditional logistic regression

33

Modeling With Many Dummy Variables

- Of ultimate concern is
  - The number of distinct levels J in a sample size of n
  - The number of cases, controls at each level
  - The control log event rate β_0 and the true log odds ratio β_1
  - The variance of the true parameter values $Var(\beta_{2j})$

$$\text{Nominal variable} \quad W_i \in \{c_1, c_2, c_3, \ldots, c_J\}$$

$$\text{Dummy variables} \quad W_{ij}^* = \begin{cases} 1 & W_i = c_j \\ 0 & else \end{cases}$$

$$\text{logit}(p_i \mid X_i, W_i) = \beta_0 + X_i \times \beta_1 + \sum_{j=1}^{J} W_{ij}^* \beta_{2j}$$

34

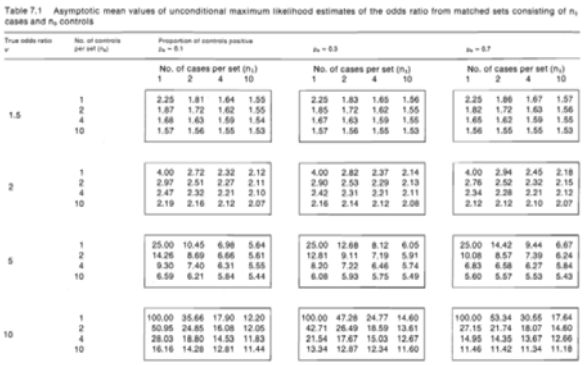
Logistic Regression: Bias

- If there are too few subjects in many strata, logistic regression is known to lead to biased results
  - Arises from noninformative strata
  - The estimated OR can tend toward the square of the true OR
- We can quantify the degree of bias according to the number of cases and controls in each stratum
  - Also a function of the control event rate and the true OR
- With the fewest data in each stratum, estimated OR tends to the square of the OR
  - Less biased as number of cases / controls increases

35

Bias From Too Many Strata

- Breslow & Day: Considering ratio of cases to controls in each stratum, control event rate, and true OR



36

Take Home Message

- When analyzing OR in the presence of many strata with sparse data, we cannot just adjust for indicators of the strata in logistic regression
- This holds true especially when matching on random effects
  - When matching on each case individually, the number of strata increases to infinity as the sample size increases
  - We thus have too many "nuisance parameters" modeling the stratum effects
- But it is also true with finely graded fixed effects in small samples

37

Solution: General Idea

- We define strata according to the matching variables
  - We will not be able to consider an association between the disease incidence and the matching variables
  - (Presumably they are associated with outcome, or why did we bother? We just do not care about estimating their effect)
- We consider the *s*-th stratum of *S* total strata
  - Assume there is 1 case and M_s controls
- In the absence of an association between a POI *X* and the disease, we expect the case to look like a randomly chosen individual from the stratum
  - We will try to estimate the amount of "biased" sampling that had to occur for us to (for instance) keep seeing that the cases had higher values of *X* among individuals in the stratum

38

Conditional Logistic Regression: EE

- Find regression parameter estimates that make the case look like it was randomly chosen among the total weights

For stratum $s, i = 0, \ldots, M_s$ ($i = 0$ is case, others controls)

$$\log\left(\frac{\Pr(D_i = 1 \mid R_i, \vec{W}_i)}{\Pr(D_i = 0 \mid R_i, \vec{W}_i)}\right) = \beta_{0s} + \beta_1 \times R_i + \beta_2 \times w_{1i} + \cdots$$

Consider relative odds of case:

$$\sum_{s=1}^{S} \frac{\exp(\beta_1 \times R_0 + \beta_2 \times w_{10} + \cdots)}{\sum_{i=0}^{M_s} \exp(\beta_1 \times R_i + \beta_2 \times w_{1i} + \cdots)}$$

39 a2

CLR: Comments

- Comparisons are first made within a matched cluster (stratum) and then averaged across strata
- Hence: For every variable for which we want to estimate associations with the disease, we must have some strata that have discordant pairs on that variable
  - We do not estimate an "intercept" for each stratum
    - It cancels out of the estimating equation
  - We cannot estimate an effect of the matching variable(s)
    - They are constant within each stratum
- We can consider interactions between the matching variables and other covariates that do differ within strata
  - Those interactions would differ within strata

40

a2 anytwo, 11/26/2013

CLR: Analogy with Proportional Hazards

- The estimating equation for CLR looks just like the estimating equation for PH regression
- In PH regression, our strata are defined by the time at which some subject has an event
  - As our sample size goes to infinity, so do the number of strata
- In PH regression, we do not estimate the underlying hazard
  - We only estimate the hazard ratios

41

CLR: Stata

- `clogit` *`casevar predvar1 …`*`, group(`*`gvar`*`) [or]`
  - *gvar* is cluster identifier
  - option `or` will give exponentiated results

42

Example

- Case-control study of endometrial cancer from Leisure World (as described in Breslow & Day)
- 63 cases of incident endometrial cancer
- 4 matched controls
  - Living in Leisure World
  - Moved into Leisure World at same time
  - Same birth year (+/- 1 year)
  - Same marital status
- Additional covariates
  - Estrogen use
  - Hypertension
  - Gall bladder disease
  - Obesity

43

Example: Logistic Regression

```
. logistic case est hyp i.ob

Logistic regression                     Number of obs   =        315
                                        LR chi2(4)      =      38.37
                                        Prob > chi2     =     0.0000
Log likelihood = -138.44328             Pseudo R2       =     0.1217

------------------------------------------------------------------------------
        case | Odds Rat   StdErr      z    P>|z|     [95% Conf Intrvl]
-------------+----------------------------------------------------------------
         est |   8.846    4.046     4.77   0.000     3.609    21.68
         hyp |    .955    .2998    -0.15   0.883     .5159    1.767
             |
          ob |
           1 |   1.78     .6087     1.68   0.092     .9100    3.479
           9 |   1.96    1.155      1.15   0.251     .6205    6.217
------------------------------------------------------------------------------
```

44

Example: Logistic Regression

```
. logistic case est hyp i.ob i.set
Logistic regression                 Number of obs   =        315
                                    LR chi2(66)     =      48.41
                                    Prob > chi2     =     0.9489
Log likelihood = -133.42345         Pseudo R2       =     0.1535

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   15.4     8.20     5.14   0.000     5.44      43.7
         hyp |   .989     .390    -0.03   0.978     .457      2.14
             |
          ob |
           1 |   2.35     1.03     1.95   0.051     .997      5.537
           9 |   2.31     1.73     1.12   0.264     .532      10.0
             |
         set |
           2 |   .973     1.70    -0.02   0.987    .0314      30.1
           3 |   .445     .772    -0.47   0.641    .0148      13.3
------------------------------------------------------------------------------
(lots deleted)
```

45

Comments

- The matching variable is believed to be a precision variable
  - Its distribution is similar across cases and controls
- We would expect some attenuation of OR if we fail to adjust for the matching
  - It will depend upon the range of the precision variable and strength of association
  - (I did not see too much here: cf the clogit model later)
- Adjusting for the matching variable as a dummy variable leads to marked anti-conservative bias in the OR

46

Example: Conditional Logistic Regression

```
. clogit case est hyp i.ob, group(set) or

Conditional (fixed-effects) logistic regression
                                    Number of obs   =        315
                                    LR chi2(4)      =      38.41
                                    Prob > chi2     =     0.0000
Log likelihood = -82.188646         Pseudo R2       =     0.1894

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   8.652    4.033     4.63   0.000     3.47      21.6
         hyp |   .9929    .3352    -0.02   0.983     .512      1.92
             |
          ob |
           1 |   1.861     .696     1.66   0.097     .894      3.87
           9 |   1.903    1.250     0.98   0.327     .525      6.89
------------------------------------------------------------------------------
```

47

Comments

- We do not have any estimate of the effect of age, time in community, or marital status
- Age was only imprecisely matched (+/- 1 year), so we could fit a model with age
  - We would not expect much precision with only one year's worth of variation
- We could fit interactions with age, marital status, etc. by estrogens, because those interactions will vary within strata even if the main effects of the matching variables do not

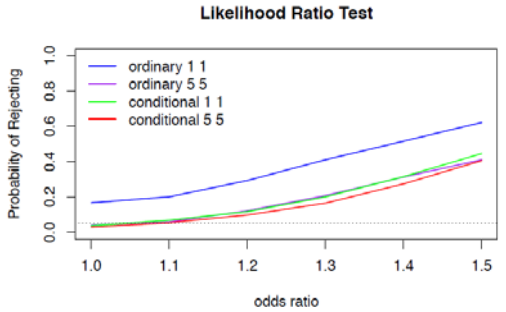
48

Comments: Recap

- We are always concerned when fitting dummy variables in which there are too few observations to estimate the parameters
  - Bias and inflation of type 1 error
  - "Too few" is not enough variation in the events: We need to be able to estimate proportions/odds and the ratios
- To get around that, we can fit
  - Logistic regression without adjusting for the matching variables
    - There will be attenuation of the OR if the matching variables are prognostic
    - There will be slight loss of power if the matching variables are strongly prognostic
  - Logistic regression with collapsed categories or continuous modeling of fixed effects
  - Conditional logistic regression

49

Comparisons of Behavior: Ratio

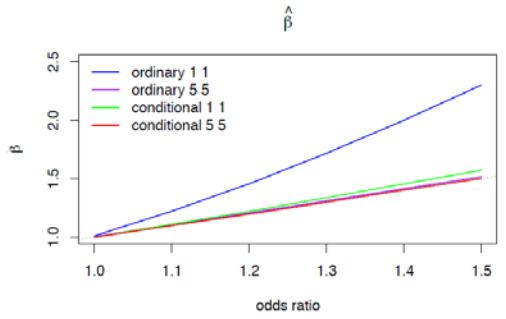


No confounding. 400 subjects. 1000 replications.

Biostatistics/Epidemiology 536 Autumn 2012 B. McKnight

50

Comparisons of Behavior: Ratio

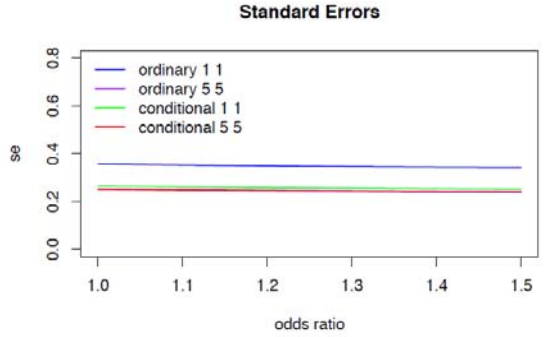


No confounding. 400 subjects. 1000 replications.

Biostatistics/Epidemiology 536 Autumn 2012 B. McKnight

51

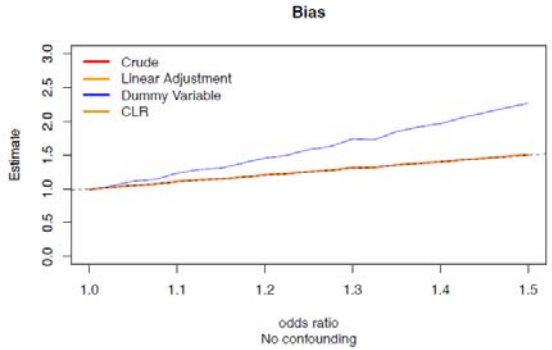
Comparisons of Behavior: Ratio



No confounding. 400 subjects. 1000 replications.

52

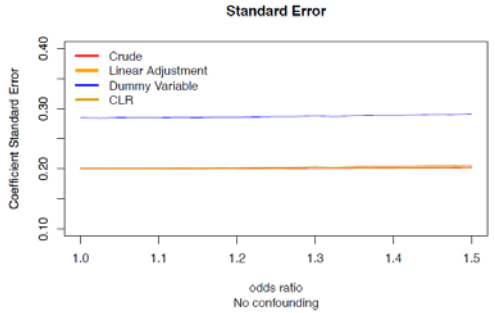
Comparisons of Behavior: Adjustment



No confounding. 400 subjects. 1000 replications.

53

Comparisons of Behavior: Adjustment

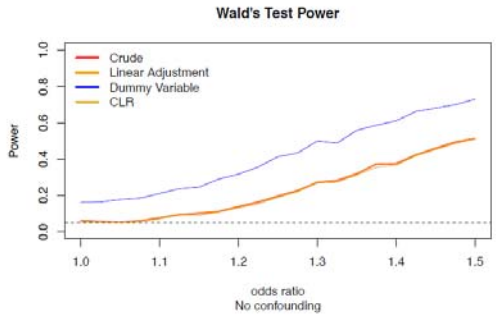


No confounding. 400 subjects. 1000 replications.

Biostatistics/Epidemiology 536 Autumn 2012 B. McKnight

54

Comparisons of Behavior: Adjustment



No confounding. 400 subjects. 1000 replications.

Biostatistics/Epidemiology 536 Autumn 2012 B. McKnight

55

Ordered Logistic Regression

56

So far…

- We have considered data analysis methods for binary data
  - RD regression using linear regression
  - RR regression using Poisson regression
  - OR regression using logistic regression
    - Conditional logistic regression with sparse data in strata
- What if we have ordered categorical data with more than two categories?
  - Stage of disease (cancer, primary biliary cirrhosis, heart failure)
  - Grade of cancer
  - Atrophy, white matter changes on MRI
  - Self-perceived general health

57

Reduce to a Problem Already Solved

- Logistic regression:
  - Dichotomize the ordered categorical variable at some arbitrary threshold
- Linear regression:
  - Treat the levels as "scores" and consider the mean score
  - Qualitatively valid: Gives direction of "effect"
  - Quantitatively hard to interpret: What does mean stage indicate?

58

Example

- What variables are predictive of a subjects self perception of general health in the CHS MRI data set
- General health is measured on an ordered scale of 1 – 5
- Probably not linear with respect to any particular physical quantity
  - So difference between 1 and 2 may have no calibration with a difference between 4 and 5

59

Separate Logistic Regressions

```
. recode genhlth 1=0 2/5=1, gen(dg1)
. recode genhlth 1/2=0 3/5=1, gen(dg2)
. recode genhlth 1/3=0 4/5=1, gen(dg3)
. recode genhlth 1/4=0 5=1, gen(dg4)

. logistic dg1 age
------------------------------------------------------------------------------
         | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
     age |   1.011702   .0113778     1.03   0.301     .9896461     1.03425

. logistic dg2 age
------------------------------------------------------------------------------
     age |   1.027365   .0067786     4.09   0.000     1.014165    1.040737

. logistic dg3 age
------------------------------------------------------------------------------
     age |   1.046692   .0082778     5.77   0.000     1.030593    1.063043

. logistic dg4 age
------------------------------------------------------------------------------
     age |   1.056785   .0175759     3.32   0.001     1.022892      1.0918
```

60

Extend Previously Described Methods

- "Ordinal logistic regression" using the "Proportional Odds Model"
- Simultaneously consider every possible dichotomization
- Conceptual idea:
  - Estimate an OR from each such logistic model
  - Use some sort of weighted average of the estimated OR from each possible dichotomization

61

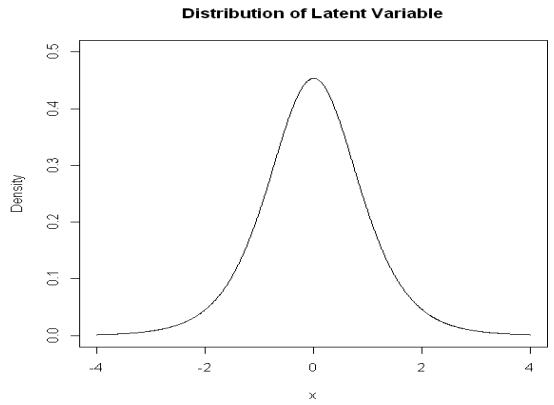
Proportional Odds Model

- Presume that the ordinal categorical variable is some arbitrary categorization of a "latent" continuous variable
  - Probably a reasonable assumption, but multifactorial contributions may only be "partially ordered"
- Presume that comparisons between two groups based on dichotomization of the latent variable will always lead to the same odds ratio
  - "Proportional odds"

62

Distribution of Latent Variable

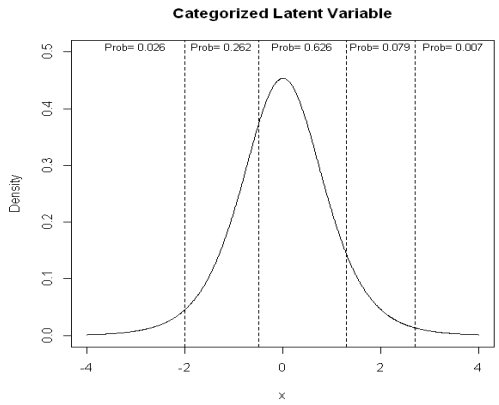
- Some continuous latent variable measuring general health



63

Categorization of Latent Variable

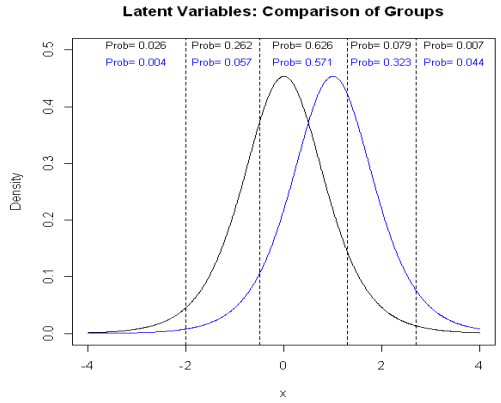
- Some continuous latent variable measuring general health



64

Distributions Across Groups: Probability

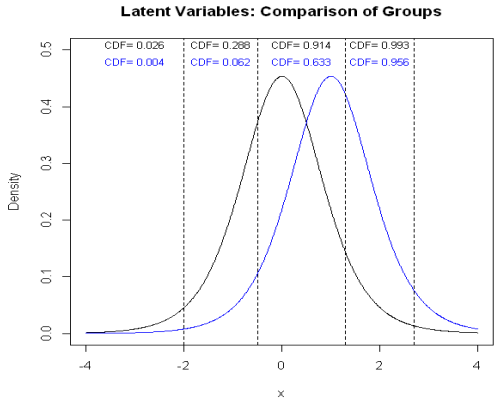
- Some continuous latent variable measuring general health



65

Distributions Across Groups: CDF

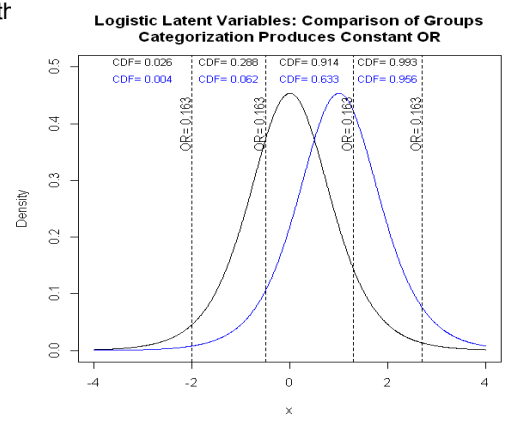
- Some continuous latent variable measuring general health



66

Odds Ratios Across Groups

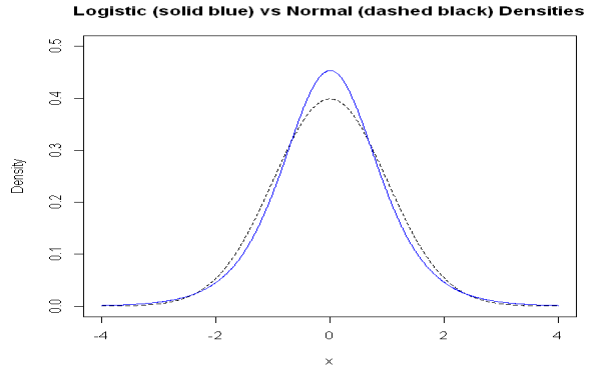
- Some **logistic distributed** latent variable measuring general healtr



67

Logistic Distribution

- Compared to standard normal distribution
  - Same mean, variance



68

Stata: ologit

- Proportional odds model is implemented in Stata with the ologit command
  - Output on the log odds scale by default, but can use "or" option
- Just like any other regression command
  - `Ologit yvar xvar1 xvar2 …`
- Provides estimates of multiple "intercepts"

69

Example: Ordered Logistic Regression

```
. ologit genhlth age

Ordered logistic regression       Number of obs   =       3775
                                  LR chi2(1)      =      30.87
                                  Prob > chi2     =     0.0000
Log likelihood = -5024.6316       Pseudo R2       =     0.0031

------------------------------------------------------------------------------
    genhlth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        age |       .033    .00596     5.55   0.000      .0214       .0447
------------+-----------------------------------------------------------------
      /cut1 |       .166    .448                         -.712        1.04
      /cut2 |      2.06     .447                          1.18        2.93
      /cut3 |      4.01     .452                          3.12        4.89
      /cut4 |      5.94     .461                          5.03        6.84
------------------------------------------------------------------------------
```

70

Example: Ordered Logistic Regression

```
. ologit genhlth age, or

Ordered logistic regression       Number of obs   =       3775
                                  LR chi2(1)      =      30.87
                                  Prob > chi2     =     0.0000
Log likelihood = -5024.6316       Pseudo R2       =     0.0031

------------------------------------------------------------------------------
    genhlth | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        age |   1.034     .00616     5.55   0.000     1.0216      1.0458
------------+-----------------------------------------------------------------
      /cut1 |    .166     .448                         -.712        1.04
      /cut2 |   2.06      .447                          1.18        2.93
      /cut3 |   4.01      .452                          3.12        4.89
      /cut4 |   5.94      .461                          5.03        6.84
------------------------------------------------------------------------------
```

71

Example: Interpretation

- When comparing two groups differing in age the odds of having self-perceived general health one category worse is 3.4% higher for every year difference in age between the groups, with the older group tending toward the worse perception of health

72

Example:Linear Regression

```
. regress genhlth age, robust

Linear regression                              Number of obs =    3775
                                               F(  1,  3773) =   26.60
                                               Prob > F      =  0.0000
                                               R-squared     =  0.0080
                                               Root MSE      =  .92668
------------------------------------------------------------------------------

        |          Robust
genhlth |  Coef.  StdErr       t     P>|t|     [95% Conf Intrvl]
    age |  .0162  .00314     5.16    0.000      .0100     .0224
  _cons |   1.51   .235      6.40    0.000      1.045     1.97
```

73

Example: Interpretation

- When comparing two groups differing in age the average level of self-perceived health is 0.016 higher for every year difference in age between the groups, with the older group tending toward the worse perception of health.

74

Comments

- I believe the assumptions of proportional odds fairly strong
- That does not mean that the weighted average of the OR across the dichotomizations might not be as reasonable a summary of an association as any other
- My own tendency, however is to just talk about the mean of ordered categorical variables
  - The difficulty in interpreting the difference in mean stage, for instance, is not inherently worst than the difficulty in interpreting what the average OR might be when proportional odds does not hold

75

Current Status Data

76

Current Status Data

- Sometimes we are interested in time to some event, but the exact timing of the event cannot be known
- Examples:
  - Time of seroconversion for HIV
  - Time of polyp recurrence
  - Time of "silent" myocardial infarctions
- Study designs:
  - Longitudinal measurements at prescribed times:
    - Interval censored time to event data (see Biost 537)
  - Cross-sectional study measuring status at a single point
    - Current status data possibly analyzed as binary outcomes

77

Data

- Measurements at a **random** time that is independent of known status

Unobserved time of event T_i

Time of measurement W_i

Indicator of event $Y_i = \begin{cases} 1 & i^{th} \text{ individual positive at } W_i \\ 0 & else \end{cases}$

Regression model $g(p_i \mid X_i, W_i) = \beta_0 + X_i \times \beta_1 + h(W_i) \times \beta_2$

78

Parameterization of Models

- Link function g( )
  - Logit (so logistic regression)
  - Complementary log log: `log (- log (1 - p_i))`
- Modeling of time h ( )
  - Need to consider interpretation of the intercept
    - Probability of an event at h (T_i) = 0
  - If that probability is near zero, then we want to avoid choices that would have the intercept at negative infinity
    - Avoid modeling *T* untransformed
  - Perhaps consider log transformed time
    - Intercept is now probability of an event prior to time 1

79

Logit Link: Logistic Regression on log(W)

- Measurements at a **random** time that is independent of known status
- Logit link modeled on log time of measurement generates the "log logistic" parametric distribution
  - log (T) has a logistic distribution

$$\Pr(Y_i = 1 | X_i = 0, W_i = w_i) = \Pr(T_i \le W_i | X_i = 0, W_i = w_i)$$
$$= \frac{\exp(\beta_0 + \log(W_i) \times \beta_2)}{1 + \exp(\beta_0 + \log(W_i) \times \beta_2)} = \frac{W_i^{\beta_2} e^{\beta_0}}{1 + W_i^{\beta_2} e^{\beta_0}}$$

- (If we had full observations, fitting a parametric model using log logistic model would be more efficient)

80

Complementary log log Link on log (W)

- Measurements at a **random** time that is independent of known status
- Complementary log log link modeled on log time of measurement generates the Weibull parametric distribution
  - Log hazard is linear in log time
  - Only distribution that is both accelerated failure time and PH

$$\begin{aligned}\Pr(Y_i = 1|X_i = 0, W_i = w_i) &= 1 - \Pr(T_i > W_i|X_i = 0, W_i = w_i) \\ &= \exp\left[-\exp(\beta_0 + \log(W_i) \times \beta_2)\right] = \\ &= \exp\left[-W_i^{\beta_2} e^{\beta_0}\right]\end{aligned}$$

- (If we had full observations, fitting a parametric model using log logistic model would be more efficient)

81

Stata

- Logistic regression to produce log logistic inference on current status data is trivial
  - Include the log time of observation as a covariate
- Using Stata command glm with the complementary log log link will correspond to the parametric Weibull model
  - `glm yvar logW xvar1 …, family(binomial) link(cloglog)`

82

Comments

- As a general rule we would prefer
  - To know the exact time that an event occurred, and
  - To use that exact information in the model
- Second choice is
  - To know the exact time for some individuals, and
  - To have right censored observations for others
- But when only current status data is available, we can still do something
- **However:** Note that current status does not make use of the exact time that an event occurred, only whether the event occurred prior to a time that was chosen independently of knowledge of the time of the event

83

Comments

- In light of the above, the methods covered for censored survival times will be of greatest use with time to event
- It is worth commenting, however, on the robustness of the Weibull parametric model
  - It is a special case of proportional hazards, so should agree in principle with the most common analyses of biomedical time to event data
    - The parametric Weibull model is used extensively in engineering
  - To the extent that the log hazard for a time to event variable is linear in log time, the Weibull model will tend to fit the data reasonably well
    - It will often agree well with other parametric AFT models depending on the "support of the censoring" distribution
    - (That is, what was the time that individuals would definitely be censored)

84