Biostatisics

Lecture 13

Design and Analysis Techniques for Epidemiologic Studies

Dummy Variables

- Use ‘dummy’ (indicator) variables to represent the levels of a variable measured on the nominal scale
  - e.g., education, income, race, gender
  - levels have no intrinsic numeric relationship
- Procedure: choose one level as the ‘reference’ level and then form n-1 dummy variables to represent the n levels of the original variable
- Use set of dummy variables in regression model

Example of Constructing Dummy Variables

- Income has three levels: low, middle, and high
- Choose low income as the reference group
- Form two indicator variables:
  - Income1: 0=low, 1=middle
  - Income2: 0=low, 1=high
- Note that both income1 and income2 have to be in the model to properly represent the ‘low’ income group
- A person from the middle income category would have income1=1 and income2 = 0

Interactions

- Interaction terms represent the joint action of two+ variables
- The interaction term of race and income in a model for TV watching is an indicator for whether the relationship of income with TV watching varies significantly by race
- Most effective when constructed using a binary variable (such as race in NGHS) along with another classification variable or a continuous variable
- Mathematically, can construct interactions from any type of variable, but becomes more difficult to interpret

Example of Constructing an Interaction Term

- Determine what variables might represent different relationships
  - Example of race and income from before
- Multiply the two terms together
  - Race*income where race is coded 0=white, 1=black and income is low, middle, high
  - Thus, interaction term applies only to black girls

Interpretation of Interaction Terms

- Test of significance is a test for a different relationship between income and TV watching for black compared to white girls
  - For white girls, relationship of income and TV watching represented by the income 'main effects' term
- If test of significance for an interaction term is significant, may be necessary to fit separate models based on the levels of the interaction terms
  - If the race*income interaction is significant, fit separate models for white girls and for black girls

Interpretation of Interaction Terms

- Principal of hierarchy
  - The appropriate main effects must remain in the model if the interaction term remains in the model
  - Otherwise, impossible to interpret with no reference group

Variable Selection

- Start with the full model
- Eliminate most non-significant term and refit the model
- Continue until only significant terms remain
- Keep sets of dummy variables together
- Keep main effects in when interaction terms are in the model even if main effects are not significant
- Step-up and step-wise regression may miss important variables

Logistic Regression Models

- Logistic regression used for binary or multinomial outcomes to predict probability (odds) of the 'event' happening
  - Ex., what are odds of developing diabetes by age 50 given body mass index and dietary factors
  - Logistic regression models the effects of different predictors on the odds of the event in a particular person or group while staying within the probability limits of zero to one
  - Multinomial logistic regression predicts the odds of a particular level of an outcome
- Predictors act as in multivariate regression, except that prediction uses the logit

Logistic Regression Model – II

- Logistic regression model is written as
  - Ln (p / 1-p) = $\alpha + \beta x$
  - Or
  - $$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$
  - where $\alpha + \beta x$ is called the logit
- To predict the odds of an event for an individual, use
  - Exp($\alpha + \beta x$)

Odds Ratios

- Odds ratios used extensively in epidemiology to express the odds of developing a condition due to the presence of a factor compared to the absence of the factor
  - Odds for success are the ratio of the probability of success to the probability of failure
  - Odds ratio is the ratio of the odds of a disease with a specific factor to the odds of a disease without the factor
- Thus, an odds ratio of 2.0 indicates that a person with the specific factor has twice the odds of contracting the disease than a person without that factor

Interpreting the Logistic Model

- Parameters are the logits which are used to calculated the odds ratios
  - Odds ratio for high income is exp(parameter for high income) and is the increase in odds compared to the reference group
- Odds ratio of 1 is equivalent to a regression parameter of 0
  - Odds ratios greater than 1 indicate an increase in the odds of the event
  - Odds ratios less than 1 indicate a decrease in the odds