Biostatistics II 514-5509

Course Description:

Modern multivariable statistical analysis based on the concept of generalized linear models. Includes linear, logistic, and Poisson regression, survival analysis, fixed-effects analysis of variance and repeated measures analysis of variance. Course emphasizes the underlying similarity of these methods, the choice of the right method for specific problems, common aspects of model construction, the testing of model assumptions through influence and residual analyses, and the use of graphical and other methods to present results that are readily understood by clinicians. Prerequisite, Biostatistics 1 or consent of the course director. (Dupont, 4 hours)

Course Objectives: to teach the following competencies

Select and generate appropriate graphical and numerical summaries of data. Graphical and tabular methods of presenting data are emphasized throughout the course. Particular emphasis is placed on drawing confidence bands for regression curves. See in particular syllabus sections 1.8, 2.3, 2.9, 3.6, 4.10, 5.2, 5.3, and 6.4

Use principles of hypothesis testing to make inferences about populations from samples. Introduced in Biostatistics I, hypothesis testing is a major topic of this course. It is explained and illustrated in sections devoted to linear regression, logistic regression, survival analysis, Poisson regression and analysis of variance. Each section starts with the appropriate simple regression technique (syllabus sections 1, 3, and 5), and is followed by the corresponding approach to multivariable regression (syllabus sections 2, 4, and 6). Simple and multiple Poisson regression analysis are introduced in section 7 while fixed and mixed effects analysis of variance are introduced in sections 8 and 9, respectively.

Perform residual analyses and draw plots to assess how well models fit the data and to detect outliers. See syllabus sections 1.8, 2.7, 4.9, 6.7, 6.8.1, 7.10, and 10.2.4

Communicate statistical findings to others. This course provides detailed explanations as to how to convert parameter estimates into relative risks or odds ratios that have meaning to a clinical audience. The use of multivariable regression methods to calculate adjusted odds ratios and relative risks is explained in syllabus sections 2.2, 2.4, 4.2, 4.3, 4.5, 4.6, 6.2, 6.3, and 7.9.1. The assessment of effect modifiers is discussed in syllabus sections 2.4, 2.5, 4.8, 7.9.5 and 9.2.

Use computer software to conduct simple statistical analysis. The Stata statistical software package is used throughout this course.

Understand assumptions underlying multiple regression models. Know how to make inferences from these models. The assumptions underlying the multiple regression models discussed in this course are given in syllabus sections 2.1, 2.2, 4.1, 4.2, 6.1, 6.2, 7.6.1, 8.6, 8.7, 9.1, and briefly in 9.3.

Syllabus

1. Simple Linear Regression

   1.1. Distinction between a parameter and a statistic
   1.2. The normal distribution
   1.3. Inference from a known sample about an unknown target population
   1.4. Simple linear regression: Assessing simple relationships between two continuous variables
   1.5. Interpreting the output from a linear regression program
      1.5.1. Analyzing data with Stata
   1.6. Plotting linear regression lines with confidence bands
   1.7. Making inferences from simple linear regression models
   1.8. Lowess regression and residual plots.
      1.8.1. How do you know you have the right model?
   1.9. Transforming data to improve model fit
   1.10. Comparing slopes from two independent linear regressions

2. Multiple Linear Regression

   2.1. Extending simple linear regression to models with multiple covariates
   2.2. Meaning of parameters in a multiple linear regression model
   2.3. Exploratory data analysis
      2.3.1. Density distribution sunflower plots for displaying high-density bivariate data
      2.3.2. Matrix scatterplots
   2.4. Additive models and models with interaction terms
   2.5. Building and interpreting complex linear models
   2.6. Stepwise methods of building regression models
   2.7. Model validation: Evaluating residuals, leverage and influence
   2.8. Least squares estimation and maximum likelihood estimation
   2.9. Information criteria for assessing statistical models
      2.9.1. Akaike's information criteria
      2.9.2. Schwarz's Bayesian information criteria
   2.10. Restricted cubic splines: Using multiple linear regression to model non-linear relationships between continuous variables.
   2.11. Calculating 95% confidence bands for regression curves from restricted cubic spline models.

3. Introduction to Logistic Regression

   3.1. Simple logistic regression: Assessing the effect of a continuous variable on a dichotomous outcome

- 3.2. How logistic regression parameters affect the probability of an event
- 3.3. Probability, odds and odds ratios
- 3.4. Generalized linear models: The relationship between linear and logistic regression
- 3.5. 95% confidence intervals for proportion.
  - 3.5.1. Wilson confidence interval for a proportion
- 3.6. Plotting probability of death with 95% confidence bands as a function of a continuous risk factor
- 3.7. Review of classic 2x2 case-control studies
- 3.8. Analyzing case-control studies with logistic regression

4. Multiple Logistic Regression

- 4.1. Extend simple logistic regression to models with multiple covariates
- 4.2. Similarity between multiple linear and multiple logistic regression
- 4.3. Multiple 2x2 tables and the Mantel-Haenszel test
  - 4.3.1. Estimating an odds ratio that is adjusted for a confounding variable
- 4.4. Using logistic regression as an alternative to the Mantel-Haenszel test
- 4.5. Using indicator covariates to model categorical variables
- 4.6. Making inferences about odds ratios derived from multiple parameters
- 4.7. Analyzing complex data with logistic regression
  - 4.7.1. Multiplicative models
- 4.8. Models with interaction
- 4.9. Assessing model fit
  - 4.9.1. Testing the change in model deviance in nested models
  - 4.9.2. Evaluating residuals and influence
- 4.10. Using restricted cubic splines in logistic regression models
  - 4.10.1. Plotting the probability of an outcome with confidence bands
  - 4.10.2. Plotting odds ratios and confidence bands

5. Introduction to Survival Analysis

- 5.1. Survival data: follow-up time and fate at exit
  - 5.1.1. Right censored data
- 5.2. Kaplan-Meier survival curves
- 5.3. Kaplan-Meier cumulative mortality curves
  - 5.3.1. Greenwood confidence bands for survival and mortality curves
  - 5.3.2. Displaying censoring times and numbers of patients at risk
- 5.4. Estimating survival probabilities
- 5.5. Censoring and biased Kaplan-Meier survival curves
- 5.6. Log-rank test for comparing survival curves
- 5.7. Hazard functions and cumulative mortality
- 5.8. Simple proportional hazards regression model
  - 5.8.1. Hazard rate ratios and relative risk
  - 5.8.2. Estimating relative risks from proportional hazards models
- 5.9. Tied failure times and biased relative risk estimates

6. Hazard Regression Analysis of Survival Data

- 6.1. Extending simple proportional hazards regression to models with multiple covariates
- 6.2. Model parameters, hazard ratios and relative risks
- 6.3. Similarities between hazard regression and linear regression
  - 6.3.1. Categorical variables, multiplicative models, models with interaction
  - 6.3.2. Estimating the effects of two risk factors on a relative risk
  - 6.3.3. Calculating 95% confidence intervals for relative risks derived from multiple parameter estimates.
  - 6.3.4. Adjusting for confounding variables
- 6.4. Restricted cubic splines and survival analysis
- 6.5. Stratified proportional hazards regression models
- 6.6. Using age as the time variable in survival analysis
- 6.7. Checking the proportional hazards assumption
  - 6.7.1. Comparing Kaplan-Meier plots to analogous plots drawn under the proportional hazards assumption
  - 6.7.2. Log-log plots
- 6.8. Hazards regression models with time-dependent covariates
  - 6.8.1. Testing the proportional hazards assumption

7. Introduction to Poisson Regression: Inferences on Morbidity and Mortality Rates

- 7.1. Elementary statistics involving rates
  - 7.1.1. Incidence and relative risk
- 7.2. Classical methods for deriving 95% confidence intervals for relative risks
- 7.3. Relationship between the binomial and Poisson distributions
- 7.4. Poisson regression and 2x2 contingency tables
- 7.5. Estimating relative risks from Poisson regression models
  - 7.5.1. Offsets in Poisson regression models
- 7.6. Poisson regression is an example of a generalized linear model
  - 7.6.1. Assumptions of the Poisson regression model
  - 7.6.2. Contrast between logistic and Poisson regression
  - 7.6.3. 95% confidence intervals for relative risk estimates
- 7.7. Poisson Regression and survival analysis
  - 7.7.1. Converting survival records to person-year records with Stata
- 7.8. Poisson Regression with Multiple Explanatory Variables.
- 7.9. Generalization of Poisson regression model to include multiple covariates
  - 7.9.1. Deriving relative risk estimates from Poisson regression models
  - 7.9.2. Analyzing a complex survival data set with Poisson regression
  - 7.9.3. The Framingham data set
  - 7.9.4. Adjusting for confounding variables
  - 7.9.5. Adding interaction terms
- 7.10. Residual analysis

8. Fixed Effects Analysis of Variance

8.1. Regression analysis with categorical variables and one response measure per subject
8.2. One-way analysis of variance
8.2.1. 95% confidence intervals for group means
8.2.2. 95% confidence intervals for the difference between group means
8.2.3. Testing for homogeneity of standard deviations across groups
8.3. Multiple comparisons issues
8.3.1. Fisher’s protected least significant difference approach
8.3.2. Bonferroni’s multiple comparison adjustment
8.4. Reformulating analysis of variance as a linear regression model
8.5. Non-parametric one-way analysis of variance
8.5.1. Kruskal-Wallis test
8.5.2. Wilcoxon rank-sum test
8.6. Two-Way Analysis of Variance
8.6.1. Simultaneously evaluating two categorical risk factors
8.7. Analysis of Covariance
8.7.1. Analyzing models with both categorical and continuous covariates

9. Mixed Effects Analysis of Variance

9.1. Analysis of variance with multiple observations per patient
9.1.1. These analyses are complicated by the fact that multiple observations on the same patient are correlated with each other
9.2. Response-feature approach to mixed effects analysis of variance
9.2.1. Reduce multiple response measures on each patient to a single statistic that captures the most biologically important aspect of the response
9.2.2. Perform a fixed effects analysis on this response feature
9.2.3. Using a regression slope as a response feature
9.2.4. Using an area under the curve as a response feature
9.3. Generalized estimating equations (GEE) approach to mixed effects analysis of variance
9.3.1. GEE analysis with logistic or Poisson models

10. Other Topics

10.1. Discriminatory analysis
10.1.1. Logistic regression
10.1.2. Classification and regression trees
10.1.3. Neural networks
10.2. Meta-analyses
10.2.1. Fixed effects model for meta-analysis
10.2.2. Random effects model for meta-analysis
10.2.3. Publication bias
10.2.4. Funnel graphs

- 10.3. Approaches to extreme multiple comparisons problems
  - 10.3.1. Permutation tests
  - 10.3.2. False discovery rates
  - 10.3.3. Learning set – test set analyses
- 10.4. Complicated statistics with nasty properties
  - 10.4.1. Bootstrap Analysis
- 10.5. Assessing diagnostic tests
  - 10.5.1. Receiver operating curves
- 10.6. Adjusting observational trials for selection bias
  - 10.6.1. Propensity score analysis

Course Structure and Methods of Teaching

Lectures introducing new material are alternated with review sessions where the material from the previous day is reinforced by practical examples. Homework assignments are due at the beginning of each new lecture. The solution to each exercise is presented by a different student at the beginning of each new lecture. Much of the learning occurs while solving exercises. Students often work in small groups. Ning Chen, the teaching assistant plays an important role in helping students with questions that arise when they are working on the assigned exercises.

Emphasis is given to understanding the assumptions required by each method, how to determine whether these assumptions are adequately met for individual data sets, how to select the best model, and how to interpret and present analyses in ways that clinicians can understand. Much of the mathematical detail underlying multivariable statistics is avoided by focusing on how to use Stata to analyze multivariable data.

Evaluation and Grading

Homework	90%
Class participation	10%

Text:

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd Ed.* Cambridge University Press, 2009.