ST4242

Jialiang Li

January 2015

INTRODUCTION

- Longitudinal Studies: Studies in which individuals are measured repeatedly through time.

- This course will cover the design, analysis and interpretation of longitudinal studies.

- The course will emphasize model development, use of statistical software, and interpretation of results.

- The theoretical basis for results will be mentioned but not developed.

- ~~No calculus or matrix algebra is assumed.~~

Text

- **Applied Longitudinal Analysis** by Garrett M. Fitzmaurice, Nan M. Laird and James H. Ware
- Supplementary materials will be mentioned in selected classes.

Assessment

- Homework: 40%
  - assigned from the textbook
  - analysis-based
  - discussed in tutorial
- Final exam: 60%

Computing

- Computing is important for statistical research.
- Standard software accepted by industry and government: SAS and S-plus (R).
- Will not be tested in the exam.
- Know which program (or function) to use, can interpret the output and seek proper help.

Review of Concepts

- Population:

  - A statistical population is the entire collection of subjects about which information is sought.

- Sample:

  - A sample is the subset of population that are actually collected in the course of an investigation.

Common Problems in Sampling

Bad sampling practice:

- (Call-in polls) Self-selected sample;
- (Animal experiments) Samples from a different population;
- Non-representative sample.

Statisticians like random (or probability) sample.

Study Designs

- Observational studies: population based case control, cross sectional, or cohort studies.
- Experiments: randomized clinical trials.

Types of Data

- Numerical (Quantitative) variable:

  - discrete variable: number of hospital visits in a year.

  - continuous variable: blood pressure, body height.
- Categorical (Qualitative) variable:

  - nominal variable: gender, race.

  - ordinal variable: degree of sunburn.

Review of Concepts

- Parameter:

  - A numerical feature of a population is called a parameter.

- Statistic:

  - A statistic is a numerical valued function of the sample observations.

Estimation

- Unbiased estimator. (Accuracy)

  - Mean of the estimator is equal to the parameter.
- Consistent estimator. (Accuracy)

  - Estimator converges to the parameter as the sample size tends to infinity.
- Efficient estimator. (Precision)

  - Estimator has the smallest variance comparing to other estimators for estimating the same parameter.

back

Hypothesis Test

- Two types of mistakes:

  - Type I: Reject the null hypothesis when it is true (significance level).

  - Type II: Fail to reject the null hypothesis when it is not true (power).

back

Confidence Interval

- Confidence level: 95%

  - Coverage rate is related to the accuracy of the estimator.

- Length of the interval

  - Length is related to the precision of the estimator.

back

Some conventions

- Capital Latin letters: variables (X, Y) or distributions (N).
- Lower Latin letters: observed values of the variables (x, y)
- Greek letters: parameters (θ, β, μ, σ)

- Violations may happen sometimes.

Convention in Regression

- X is the predictor variable (may also be called independent variable, explanatory variable, treatment variable, and covariate).
- Y is the response variable (may also be called dependent variable, outcome variable).
- Subscript i indexes the observation.

Ordinary Linear Regression

- For i=1,…,n, assume the following model for the two variables x and y

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{or}$$

$$\mu_{y|x} = \beta_0 + \beta_1 x, \quad \text{or}$$

$$E(y \mid x) = \beta_0 + \beta_1 x$$

Assumption I

- Linearity assumption:

  The conditional expectation (mean) of the response variable y is linearly dependent on the predictor variable x.

  - two parameters of interest: intercept β 0 and slope β1.

Assumption II

- Equal variance assumption:

$$\text{Var}(\varepsilon \mid x) = \sigma^2_{y|x}$$

Assumption III

- Normality assumption:

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Error distribution is normal: yield correct inference in a small sample study.
- Not a concern for a large sample study.

Assumption IV

- Independence assumption:

  - We assume that ε_i and ε_j are independent for $i \neq j$.

  - Observations from the same subjects are usually correlated.