University of Pittsburgh

Using Medical Research Data to Motivate Methodology Development among Undergraduates in SIBS Pittsburgh

Megan Marron and Abdus Wahed

Graduate School of Public Health

Department of Biostatistics

Outline

- My Experience
- Motivation for Furthering Statistical Knowledge
- Advanced Statistical Topic: Missing Data
- Why is Missing Data a Problem?
- Missing Data Mechanisms
- Methods of Handling Missing Data
- Our Process
- Summary and Conclusion

My Experience

- **3rd year PhD student in Biostatistics**

- **BS in Applied Statistics at Rochester Institute of Technology, 2011**

- **SIBS 2010 cohort**
  - Center for Oral Health Research in Appalachia (COHRA) project
    - Used logistic regression to examine demographic variables that were associated with whether or not a subject had dental caries
  - Gave insight and advice on graduate school

- **SIBS Teaching Assistant**

My Experience

- **Graduate Student Researcher** for NIMH sponsored Center of Excellence in the Prevention and Treatment of Late Life Mood Disorders
  - Clinical Trials and Observational Studies in older adults

  What I do:

  - Attend scientific oversight meetings with PIs and collaborators
  - Consult with clinicians about their hypotheses
  - Develop analytic plans to answer their hypotheses
  - Analyze data from a variety of independently funded research projects
  - Assist clinicians in presenting their results
  - Prepare statistical methods and results for manuscripts

Motivation for Furthering Statistical Knowledge

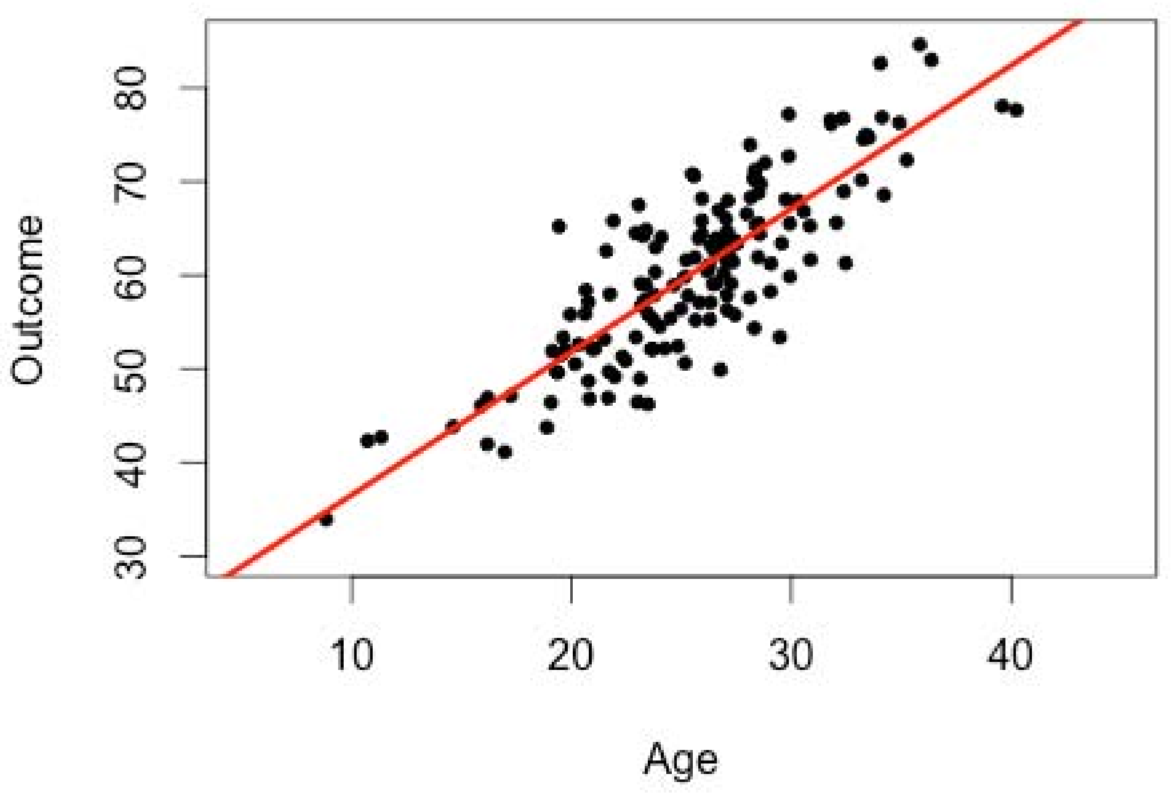
- **Statisticians are in high demand**
- **Researchers need to be aware of potential statistical issues**
  - Ensure valid findings
- **Limited background in statistics does <u>not</u> inhibit learning advanced topics**
  - SIBS Pittsburgh has successfully demonstrated this over the past 4 years

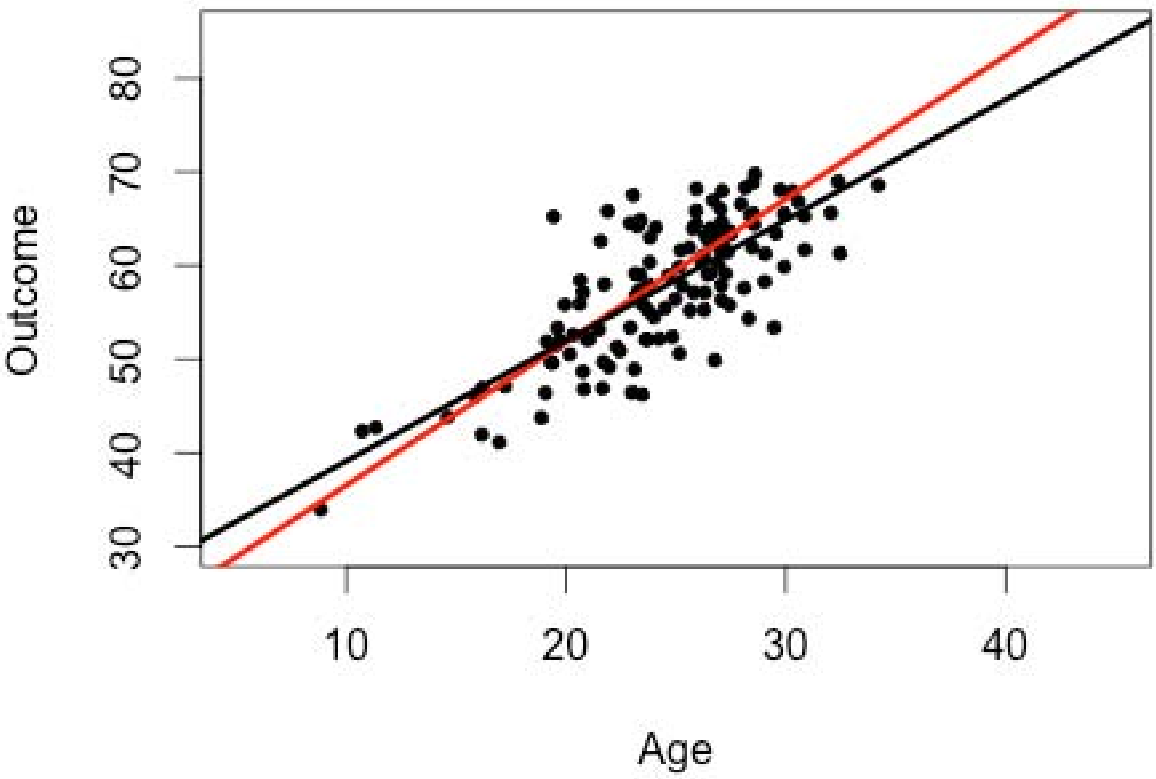
Advanced Statistical Topic: Missing Data

- **Not taught in introductory courses**
  - Given perfect datasets or perform complete-case analysis
- **Researchers should be familiar with:**
  - Different types of missing data
  - Ways of handling missing data
  - How missing data can effect results
- **Why they should be familiar with these concepts:**
  - Save time and money
  - Accurate results that are unbiased with small standard errors
  - Nobody knows your study better than you
  - Fundamental to research

Why is Missing Data a Problem?

- Biased estimates
- Larger standard errors
- Loss of information


Outcome
30
40
50
60
70
80
10
20
30
40
Age


Outcome
30
40
50
60
70
80
10
20
30
40
Age

Missing Data Mechanisms

- **An assumption about the nature of the missing values**
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing not at Random (MNAR)

MCAR

- **Probability of missing is independent of both observed and unobserved values**
- **Example: Weight Loss Study**
  - Missing a record of weight due to the scale breaking that day
  - What we know: Subject had nothing to do with the scale breaking
  - Assumption: **MCAR**
    - ➢ Missingness has nothing to do with observed or unobserved measurements

MAR

- **Probability of missing can be explained by observed data**
- **Example: Weight Loss Study**
  - Participant drops out after a month
  - What we know: Their weight has been steadily increasing
  - Assumption: **MAR**
    - ➢ Missingness has to do with observed measurements

MNAR

- **Probability of missing depends on the unobserved**
- **Example: Weight Loss Study**
  - Participant drops out after a month
  - What we know: Past weight measurements give no clue to why they would drop out
  - What we do **not** know: Subject didn't come in because they weighed themselves at home and realized they gained weight (unobserved)
  - Assumption: **MNAR**
    - ➢ Missingness has to do with unobserved measurements

Methods of handling Missing Data

- **Complete Case Analysis**
  - Delete all records that have missing
  - Assumes MCAR
  - Loss of precision
- **Inverse Probability Weighting**
- **Last Observation Carried Forward**
- **Multiple Regression Imputation**

Our Process:

1 **Introduce advanced statistical concepts in a small-group setting**

- Actively involve trainees in collaborative research projects

2 **Data analysis**

- Apply statistical techniques to a Virahep-C data

3 **Simulation**

- Show trainees what happens when changing certain conditions

4 **Presentation**

- One of the best ways to learn something is to have to teach it to others

Our Process:

1 **Introduce advanced statistical concepts in a small-group setting**

- Missing data:
  - Different types
  - Why is it a problem?
  - Methods of handling each type
  - Potential impact on study results
  - Importance of justifying the type
  - Examples to differentiate between types

Our Process:

2 **Data analysis:** Virahep-C Study

- NIH/NIDDK-funded Study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C)
- Background:
  - African Americans (AA) with chronic Hepatitis C are less likely to respond to interferon-based antiviral treatment than Caucasian Americans (CA)
- Multicenter treatment trial with 196 AA and 205 CA
  - Treatment: peginterferon and ribavirin

Our Process:

2. **Data analysis:** Virahep-C Study in SIBS
   - Outcome: Change in log viral levels between week 12 and baseline
     - Contains missing values
   - Objectives:
     1. Estimate mean change in viral levels between week 12 and baseline and mean differences between race
     2. Assess associations of baseline demographic and clinical variables on the change in viral level
   - Address objectives using each technique for handling missing data
   - Compare results obtained from each technique

Our Process:

3 **Simulation:** How to Create Missing Data

- **MCAR:**
  - Generate a random Binomial distribution
  - If subject got a 0, then the value for their outcome was deleted
- **MAR:**
  - Generate probabilities using a logistic model based off of observed values (age, sex, and treatment)
  - Generate a Bernoulli random variable for each subject using their generated probability
  - If subject got a 0, then the value for their outcome was deleted
- **MNAR:**
  - If a subject's outcome is greater than 65 then it was deleted

Our Process:

3 Simulation

- Modify sample code to examine how different methods of analysis can result in different conclusions

- Calculate relative bias and standard error to see when each type of missing data is a problem

- Benefit of a simulation:
  - True values are known
  - Type of missing data is known

Our Process:

4 Presentation

- Teach other SIBS trainees and faculty:
  - Why missing data is a problem
  - Different types of missing data
  - Methods of handling missing data
  - How results differed under each method of analysis applied to the Virahep C study
  - How results differed under each method of analysis using a simulation to create each type of missing data

Summary and Conclusion

- **Our Process:**

  1. Introduce advanced statistical concepts in a small-group setting
  2. Data analysis
  3. Simulation
  4. Presentation

- **Using our project-based training program:**

  – Advanced statistical topics **can** be taught to those with limited statistical preparation

  – Trainees were able to effectively explain techniques with useful examples that were easy to understand

  – They are better prepared for dealing with common problems in medical research

  – Gain an appreciation for statistical methods

Thank you for listening!

Contact Information:

Megan Marron

mmm133@pitt.edu

Acknowledgment:

Department of Biostatistics at the University of Pittsburgh

National Heart, Lung, and Blood Institute

National Heart Lung and Blood Institute

PITT PUBLIC HEALTH

29 March 2014