BIOS 755: Introduction & Data Examples

Alexander McLain

Outline

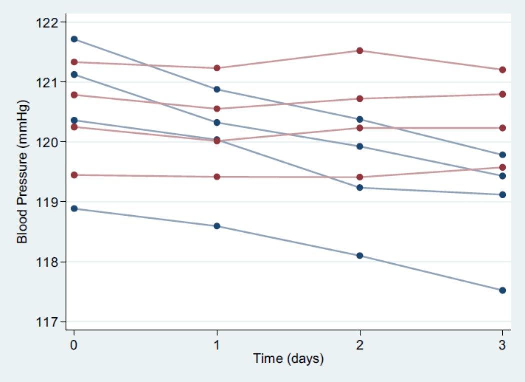
Introduction to Longitudinal Data

Examples

Features of Longitudinal Data

Introduction

- Longitudinal Studies: Studies in which individuals are measured repeatedly through time (as opposed to cross-sectional).



Longitudinal vs. Cross Sectional

Example

- Suppose investigators are interested in determining the increase in body fatness in girls after menarche.
- Two ways to do this:
  - (Cross Sectional) take a sample of 10-year-olds and a sample of 15-year-olds.
  - (Longitudinal) take a sample of 10-year-olds, measure at 10 and 15-years-old.

Longitudinal data vs. Clustered data

- Clustered data: data that exhibit correlation (typically positive correlation) within a cluster.
- Longitudinal data are actually a special case of clustered data, that have a temporal ordering.
- In this class we will mostly discuss longitudinal data:
  - Most popular clustered data
  - Good introduction to clustered data
- Later in the class we'll discuss multi-level models, which contain clustered (non-longitudinal) data.

Features of Longitudinal Data

- Repeated measures on the same variable over time
  - Height or weight in growth studies
  - Tumor size in cancer studies
  - Score on a depression scale in a mental health treatment study

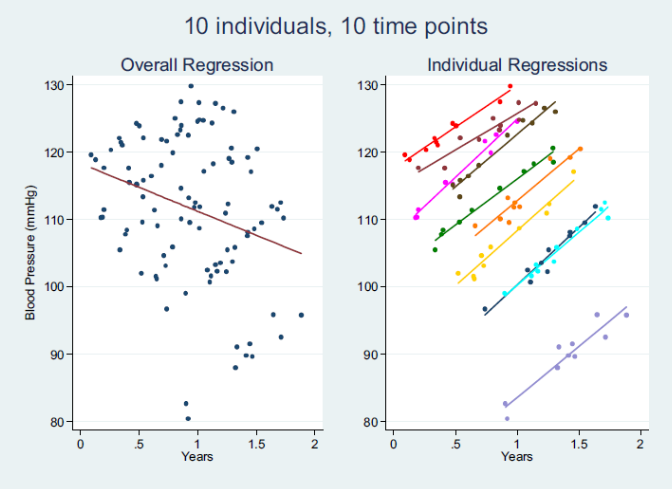
Features of Longitudinal Data

- Repeated measures on the same variable over time
  - Height or weight in growth studies
  - Tumor size in cancer studies
  - Score on a depression scale in a mental health treatment study
- Dependence among within individual measures
  - Observations within individuals tend to be more similar than observations from different individuals
  - Observations closer in time tend to be more similar than observations farther apart

Features of Longitudinal Data

- Repeated measures on the same variable over time
  - Height or weight in growth studies
  - Tumor size in cancer studies
  - Score on a depression scale in a mental health treatment study
- Dependence among within individual measures
  - Observations within individuals tend to be more similar than observations from different individuals
  - Observations closer in time tend to be more similar than observations farther apart
- Interest lies in change of some response variable over time
  - Comparing time-trends across populations
  - Individual trends
  - Within-individual variability
  - Between-individual variability

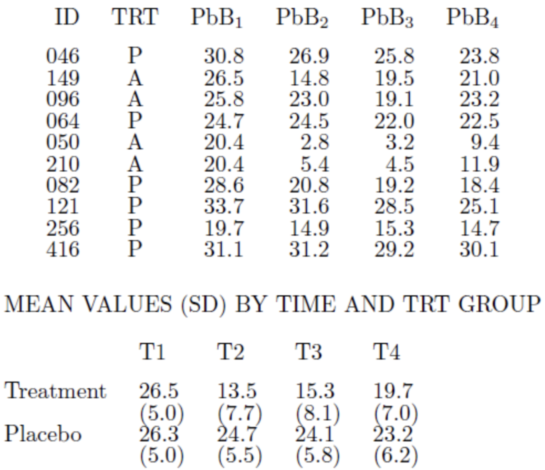
Why Do Longitudinal Data Need Special Methods?



Example 1: Treatment of Lead-Exposed Chidren (TLC) Trial

- Exposure to lead during infancy is associated with substantial deficits in tests of cognitive ability
- Chelation treatment of children with high lead levels usually requires injections and hospitalization
- A new agent, Succimer, can be given orally
- Randomized trial examining changes in blood lead level during course of treatment
- 100 children randomized to placebo or succimer
- Measures of blood lead level at baseline, 1, 4 and 6 weeks

Example 1

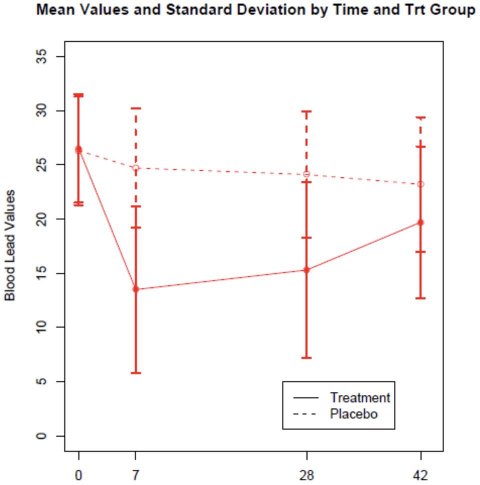


ID	TRT	PbB_1	PbB_2	PbB_3	PbB_4
046	P	30.8	26.9	25.8	23.8
149	A	26.5	14.8	19.5	21.0
096	A	25.8	23.0	19.1	23.2
064	P	24.7	24.5	22.0	22.5
050	A	20.4	2.8	3.2	9.4
210	A	20.4	5.4	4.5	11.9
082	P	28.6	20.8	19.2	18.4
121	P	33.7	31.6	28.5	25.1
256	P	19.7	14.9	15.3	14.7
416	P	31.1	31.2	29.2	30.1

MEAN VALUES (SD) BY TIME AND TRT GROUP

	T1	T2	T3	T4
Treatment	26.5 (5.0)	13.5 (7.7)	15.3 (8.1)	19.7 (7.0)
Placebo	26.3 (5.0)	24.7 (5.5)	24.1 (5.8)	23.2 (6.2)

Example 1

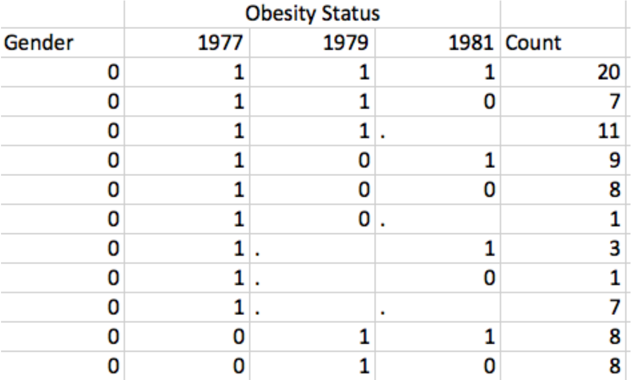

Mean Values and Standard Deviation by Time and Trt Group
Blood Lead Values
0
5
10
15
20
25
30
35
0
7
28
42
Treatment
Placebo

Example 2: Muscatine Coronary Risk Factor Study

- The Muscatine Coronary Risk Factor Study was a study designed to look at the change in obesity in children.
- Five age cohorts were measured in 1977, 1979 and 1981.
- The study had 4856 boys and girls.
- Children were classified as obese or not obese.

Example 2

There is a contingency table of some of the data:



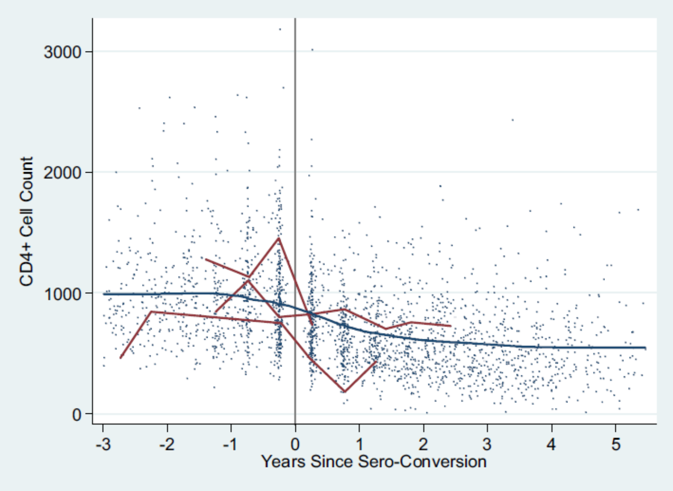
	Obesity Status			
Gender	1977	1979	1981	Count
0	1	1	1	20
0	1	1	0	7
0	1	1	.	11
0	1	0	1	9
0	1	0	0	8
0	1	0	.	1
0	1	.	1	3
0	1	.	0	1
0	1	.	.	7
0	0	1	1	8
0	0	1	0	8

Less than 40% of the children provided complete data.

Example 3: Multicenter AIDS Cohort Study of HIV

- A cohort of 369 men was followed before and after HIV sero-conversion (which is the development of detectable specific antibodies to microorganisms in the blood serum as a result of infection or immunization)
- Important indicator of immune function is the CD4+ cell count
- CD4+ count was taken on each subject approximately every six months
- Over all 2376 observations

Example 3: Multicenter AIDS Cohort Study of HIV



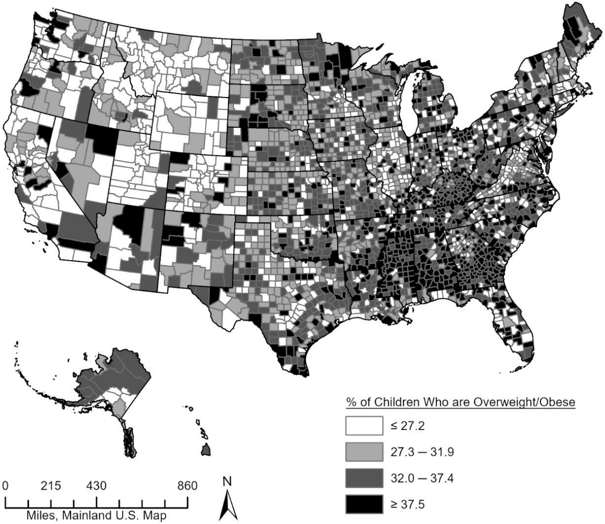
Example 3: Multicenter AIDS Cohort Study of HIV

Scientific goals of the study are to:

- Characterize the typical time course of CD4+ cell depletion after HIV infection (natural history)
- Characterize heterogeneity within and across men in CD4+ count and in progression of CD4+ depletion
- Estimate time course (trajectory) of CD4+ count for individual men, accounting for substantial measurement error in CD4+ count
- Study factors predicting levels and changes in CD4+ cell count

Example 4: County level disease rates/summaries.

- The National Study of Children's Health (NSCH) gathers data from roughly 50K 2–17 year-old children at each survey.
- The NSCH variables include indicators of ADHD, ASD, and many other conditions.
- It also includes BMI percentile, and other continuous variables.
- The public version of the data includes state, but the restricted version includes county and zip code.
- It is of interest to predict the rate of obesity for each county.


% of Children Who are Overweight/Obese
□ ≤ 27.2
▨ 27.3 – 31.9
▨ 32.0 – 37.4
■ ≥ 37.5
0
215
430
860
Miles, Mainland U.S. Map
N

Features of the examples

- Types of outcomes.
- Equally and unequally spaced measurements
- Goals.
- Missing data.
- **Types of studies**
  - Observational
  - Clinical trials
- Magnitude and type of correlation.

Covariates Collected

- **Time-invariant covariates**
  - Sex
  - Race
  - Age at baseline
  - Treatment assignment
- **Time-varying covariates**
  - Age or time since baseline
    - age/time is the most challenging and common time-varying covariate.
  - Physical activity.
  - Difficult life events in past months in a depression study.
  - Screen time.

Regression Model

- In this class, we will use statistical models to approximate reality to help us determine if there are patterns in the data and if those patterns are related to public health factors.
    - We want to make an accurate approximation of reality, such that claimed associations are true (with a certain degree of confidence) and causal (up to the limitations of the data).

Regression Model

- The statistical models we use are **“regression models”**, which is a broad term that contains models for:
  - Continuous outcomes
  - Binary outcomes
  - Count outcomes
  - Correlated outcomes (mostly by subject).
- Interpretations will vary from model to model and will be the main focus of this course.

Challenging topics

- Probability
- Math/Linear Algebra
- Interpreting and using interaction terms
- Control for continuous variables (continuous/count/binary)