Biostatistics 140.655

Lab 0

**Topics:**

* Managing multilevel and longitudinal data (MLLD)
* Visualizing MLLD
* Computing descriptive statistics from MLLD

**Learning Objectives:**

Students who successfully complete this lab will be able to:

* Transform a dataset from “wider” to “longer” format and vice versa
* Make graphical displays of MLLD with a macro-micro design that make apparent the observed data and patterns of potential in the means, variances and covariances
* Compute descriptive statistics to quantify patterns of interest

**Scientific Background - Longitudinal Data (2-level: time within person)**:

Assume you are a researcher interested in mental health symptoms among critically ill ICU survivors. You administered the Short Form (36) Health Survey (SF-36) to 100 patients who consented to participate in your study. The SF-36 was administered at hospital discharge (time 0) and then monthly for 4 months. As your study outcome variable, you are specifically interested in the mental health score from the SF-36.

*A priori* (before looking at the data)*,* you believe that the mental health symptoms will improve over the course of the follow-up and your goal is quantify the average change from discharge (time 0) to one and 4 months later.

NOTE: We are going to assume we have no deaths in our study patients or drop-out/missing data. We will address these issues later in the course.

You also collected data on age, sex (female = 1, male = 0) and baseline illness severity (higher scores indicate more severe illness) of the critically ill ICU patients.

**Scientific Background - Clustered Data (2-level: student within school):**

This data set if from a cross-sectional sub-study of high school mathematics achievement from the High School and Beyond (HS&B), a longitudinal study conducted within the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES). The NELS was established to study the educational, vocational, and personal development of young people beginning with their elementary or high school years, and following them over time as they begin to take on adult roles and responsibilities. The HS&B survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992.

We have available data from one of the assessments for 7042 students within 156 schools.

The study variables include:

Level 1: student

* mathach: a measure of mathematics achievement
* minority: dummy variable for student being non-white
* female: dummy variable for student being female
* ses: socioeconomic status (SES) based on parental education, occupation and income (z-score)

Level 2: school

* schoolid: school identifier
* sector: dummy variable for a school being Catholic
* pracad: proportion of students in the academic track
* disclim: scale measuring disciplinary climate
* himinty: dummy variable for more than 40% minority enrollment
* s: number of students enrolled at the school
* newid: rescaled school identifier, counts 1 to 156 (we created this for you)

**Datasets, Stata do-file and R-code:**

You should download the appropriate datasets (e.g. “Lab0\_wide” and “Lab0\_long” Stata or csv files) and code (Stata or R) from the CoursePlus Lab0 Online Library Folder. For example, for the multilevel data, here is Stata and R code for downloading.

**Stata:**

use “hsb\_data”, clear

bys newid: gen N = \_N

bys newid: gen n = \_n

list newid N n ses sector in 1/47

**R:**

library(tidyverse)

data = read\_csv("hsb\_data.csv")

data = data %>%

group\_by(newid) %>% # Number of elements in each group

mutate(N = n(), n = row\_number()) %>% # Index within group

ungroup()

data %>%

select(newid, N, n, ses, sector) %>%

head()

**Lab Exercise: Longitudinal SR-36 data**

1. The outcome in your study is the SF-36 mental health score measured at hospital discharge and every monthly for 4 months after hospital discharge. *Discuss with your peers why the outcome in your study can be called “multivariate”, "multi-level", and “longitudinal”. Be prepared to define each of these descriptors of a study outcome.*

2. There are two ways in which longitudinal data will be formatted: wider and longer. Wider is used for making displays of autocorrelation; longer is used for most regression analyses. *Use the commands below and then discuss with your peers what the differences are between these two formats.*

* 1. Wider format

**STATA:**

use "Lab0\_wide", clear

list in 1/2

**R:**

wide = read\_csv ("Lab0\_wide.csv")

wide[1:2,]

* 1. Longer format

**STATA:**

use "Lab0\_long", clear

list in 1/10

**R:**

long = read\_csv ("Lab0\_long.csv")

long[1:10,]

3. In longitudinal studies, the times of measurements can either be *discrete*, that is, take values from a list of fixed set of times (i.e. at hospital discharge and then monthly for 4 months) or continuous, taking values at any times. *Discuss with your peers whether there is a preference in data formatting (wider vs. longer) when time is discrete vs. continuous.*

4. In longitudinal studies, some variables are said to be *static* or *time invariant* as opposed to *dynamic* or *time varying*. Static do not change over time during the study. Examples include sex, patient condition at discharge, or SES. An example of a dynamic variable is SF36. *Using the data in the wide format, compute appropriate summary statistics for the time-varying SF36 mental health scores separately at baseline and at each follow-up time.*

*What is the mean (standard deviation) of the SF36 mental health scores at hospital discharge?*

*At 4-months post hospital discharge?*

*What is the mean (minimum and maximum) age and illness severity of the patients?*

*What proportion of the patients are female*?

**STATA:**

use "Lab0\_wide", clear

summ

**R:**

wide = read\_csv ("Lab0\_wide.csv")

summary(wide)

apply(wide,2,FUN=function(x) sqrt(var(x)))

5. In question 4, you used the longitudinal data in the wider format. This format is nice for computing summary statistics associated with the outcome over time and time invariant patient level variables. However, the data is often required to be in the longer format for plotting or fitting regression models.

*When you “reshape” the data from “wider” to “longer”, how many rows of data should be produced?*

*Use the commands below to “reshape” the data from “wider” to “longer”.*

**STATA:**

\* Reshape from wide to long

reshape long y, i(id) j(time)

\* Reshape from long to wide

reshape wide y, i(id) j(time)

**R:**

### Reshape from wider to longer

# using base R "reshape"

reshape(wide, varying=1:5, ids=seq(1,100), direction="long", v.names="y") %>%

head

# using tidyverse "pivot\_longer"

wide %>%

pivot\_longer(-c(id, age, severity, gender), names\_to = “time”, values\_to = “y”) %>%

arrange(id) %>%

head()

### Reshape from longer to wider

# using base R "reshape"

reshape(long,v.names="y", idvar= "id",timevar="time", direction="wide") %>%

head

# using tidyverse "pivot\_wider"

long %>%

pivot\_wider(names\_from = time, values\_from = y) %>%

head()