Biostatistics 140.655, 2017-18
Lab 0 Solution

Topics:

- Data management for longitudinal data
- Computation of descriptive statistics for the unit and over time
- Multivariate normal distribution

Learning Objectives:

Students who successfully complete this lab will be able to:

- Inspect longitudinal data
- Transform a dataset from "wide" to "long" format and vice versa
- Compute descriptive statistics for the unit and over time
- Write out the structure of data generated from a multivariate normal distribution
- Generate data from a multivariate normal distribution

Scientific Background:

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 that consented to participate in your study. The SF-36 will be administered at hospital discharge (time 0) and then monthly for 4 months. You are specifically interested in the mental health score of the SF-36.

A priori you believe that the mental health symptoms of the ICU survivors will improve over the course of the follow-up and you state that you will estimate the improvement in mental health symptoms comparing 1 to 4 months post hospital discharge to hospital discharge (time 0 or baseline).

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

Datasets, Stata do-file and R-code:

You should download the appropriate datasets ("lab0 wide" and "lab0 long" Stata or csv files) and code (Stata or R) from the Courseplus Lab0 Online Library Folder.

Lab Exercise:

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 is considered to be "multivariate" and "longitudinal".

   ANSWER: the outcome of interest is a list/vector of 5 measures of mental health symptoms for each patient. Since the outcome represents multiple values measured of the same outcome for each patient, the outcome is considered multivariate.

2. There are two ways in which longitudinal data will be formatted: wide and long. Use the commands below and then discuss with your peers what the differences are between these two formats.

   ANSWER: In the wide format, each patient is represented by a single row of data. The longitudinal mental health symptoms at hospital discharge and four months post discharge are represented by 5 columns in the wide dataset. Patient age, gender and illness severity are represented by 3 columns.

   In the long format, each row of data represents data for a particular patient at a given time point. In our example, each patient is represented by 5 rows of data. Mental health symptoms for a given patient are identified by the subject identifier and time identifier (a column in the dataset). Patient age, gender and illness severity are represented by 3 columns. For each row of data represented for a given patient, the values of age, gender and illness severity are repeated.

   a. Wide format

   STATA:

```
use "lab0 wide", clear
list in 1/2
```

   R:

```
wide = read.table("lab0 wide.csv",sep=",",header=T)
wide[1:2,]
```

   b. Long format

   STATA:

```
use "lab0 long", clear
list in 1/10
```

   R:

```
long = read.table("lab0 long.csv",sep=",",header=T)
long[1:10,]
```

3. In longitudinal studies, time is either discrete (i.e. at hospital discharge and then monthly for 4 months) or continuous (i.e. time from hospital discharge when patients see their primary care provider through the first year post hospitalization). When time is continuous often patients have varying numbers of longitudinal measurements; e.g. patient 1: 0, 15, 45, 90, 150 and 300 days and patient 2: 0, 49, 200 and 250 days.

   Discuss with your peers whether there is a preference in data formatting (wide vs. long) when time is discrete vs. continuous.

   ANSWER: When time is discrete and each patient is measured at the same number of measurements, there is no clear preference for the data format. However, when time is continuous and the number of measurements for each patient may vary, the long format is often preferred.

4. Using the data in the wide format, compute appropriate summary statistics for the SF36 mental health scores separately at each time and the baseline patient characteristics.

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

   Mean = 36.42, standard deviation = 10.22

   At 4-months post hospital discharge?

   Mean = 49.95, standard deviation = 9.87

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

   AGE: Mean = 49.83, minimum = 32.11, maximum = 68.49
   SEVERITY: Mean = 27.09, minimum = 16.21, maximum = 36.08

   What proportion of the patients are female?

   26% of the patients are female

STATA:

```
use "lab0 wide", clear
summ
```

```
   Variable |       Obs        Mean    Std. Dev.       Min        Max
------------+--------------------------------------------------------
         y0 |       100    36.42319    10.22092   3.664389    68.4766
         y1 |       100    39.23928    9.116138   17.71607   57.18258
         y2 |       100    45.00149    10.44998   20.25983   70.99014
         y3 |       100    50.45048    9.518102   26.98596    71.0041
         y4 |       100     49.9509    9.869278   26.86858   74.89728
------------+--------------------------------------------------------
         id |       100        50.5    29.01149          1        100
        age |       100    49.82898    8.362132   32.11149   68.48689
   severity |       100    27.08723    3.847566   16.20534   36.08126
     gender |       100         .26     .440844          0          1
```

R:

```
wide = read.table("lab0 wide.csv",sep=",",header=T)
summary(wide)
apply(wide,2,FUN=function(x) sqrt(var(x)))
```


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

   When you "reshape" the data from "wide" to "long", how many rows of data should be produced?

   Use the commands below to "reshape" the data from "wide" to "long".

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

# Reshape from long to wide
wide <- reshape(long,v.names="y", idvar= "id",timevar="time",direction="wide")
```

6. When time is discrete with many longitudinal measurements or time is continuous, data in the long format is preferred. Therefore, you should feel comfortable computing descriptive statistics for the data in the long format. Within your group, discuss the commands below to ensure that you understand the key features and compare your results to those obtained in question 4.

   STATA:

```
use "lab0 long", clear
bys time: summ y

sort id time
by id: gen counter = _n
summ age gender severity if counter==1
```

   R:

```
long = read.table("lab0 long.csv",sep=",",header=T)
tapply(long$y,long$time,summary)
tapply(long$y,long$time,FUN=function(x) sqrt(var(x)))

long$counter =
unlist(tapply(long$id,long$id,FUN=function(x) seq(1,length(x))))
summary(long[long$counter==1,c("age","gender","severity")])
apply(long[long$counter==1,c("age","gender","severity")],2,FUN=funct
ion(x) sqrt(var(x)))
```


7. Throughout the course, we will be describing statistical methods where we focus our attention on a multivariate normal distribution. In this question, we will define and explore the special case of the bivariate normal distribution.

   The bivariate normal distribution is a representation of the joint distribution for two normally distributed random variables; most interestingly when these two random variables are correlated with one another.

   We would typically express the bivariate normal distribution using the following notation: Consider random variables X and Y. Assume X and Y follow a bivariate normal distribution; associated with this distribution are 5 parameters: two means (μ_x and μ_y) and two standard deviations (σ_x and σ_y) and a correlation (ρ_{xy}). We express this distribution as:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} {\sigma_x}^2 & \sigma_x\sigma_y\rho_{xy} \\ \sigma_x\sigma_y\rho_{xy} & {\sigma_y}^2 \end{pmatrix}\right)$$

   $\begin{pmatrix} {\sigma_x}^2 & \sigma_x\sigma_y\rho_{xy} \\ \sigma_x\sigma_y\rho_{xy} & {\sigma_y}^2 \end{pmatrix}$ is the variance for $\begin{pmatrix} X \\ Y \end{pmatrix}$, which is a matrix. The diagonal elements are the variances for X and Y, respectively and the off-diagonal elements are the covariance of X and Y.

   Convince yourself that Cov(X,Y) = $\sigma_x\sigma_y\rho_{xy}$

   Go to http://socr.ucla.edu/htmls/HTML5/BivariateNormal/ where you can explore the bivariate normal distribution.

   Enter the following values for the parameters of the bivariate normal distribution: $\mu_x = 35$, $\mu_y = 50$, $\sigma_x = 10$, $\sigma_y = 10$. Start with setting $\rho_{xy} = 0$ and then allow the correlation to be 0.25, 0.50 and 0.75.

   Focus on the "look" of the bivariate normal distribution in the 3-D graph produced. You can rotate the 3-D graphic to assess the marginal distribution of X and Y and notice the impact of changing the correlation from 0 to 0.75.

8. The multivariate normal distribution or k-variate normal distribution is the natural extension of the bivariate normal distribution where you can have more than 2 random variables. In our example, we will have a k-variate normal distribution with k = 5, 5 outcome values each representing the SF-36 mental health score at one of the 5 assessment times. The parameters associated with this k-variate (k = 5) normal distribution is 5 means, 5 standard deviations and $\frac{k\times(k-1)}{2} = 10$ pairwise correlations.

   Let Y_{ij} be the SF-36 mental health score for subject i , i = 1, …, 100, and j = 0,1,2,3,4 corresponding to the number of months post hospital discharge.

   Assume the data for subject i is generated from a multivariate normal distribution with means 35, 38, 43, 49, 48 for j = 0, 1, 2, 3, 4, respectively, and equal standard deviation over time of 10 and correlation ρ_{jk} for times j and k of $\rho_{01} = 0.85$, $\rho_{02} = 0.80$, $\rho_{03} = 0.72$, $\rho_{04} = 0.69$, $\rho_{12} = 0.85$, $\rho_{13} = 0.80$, $\rho_{14} = 0.72$, $\rho_{23} = 0.85$, $\rho_{24} = 0.80$, $\rho_{34} = 0.85$.

In the space below, write out the k-variate (k=5) normal distribution.

ANSWER:

$$\begin{pmatrix} Y_{i0} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 35 \\ 38 \\ 43 \\ 49 \\ 48 \end{pmatrix}, \begin{pmatrix} 100 & 85 & 80 & 72 & 69 \\ 85 & 100 & 85 & 80 & 72 \\ 80 & 85 & 100 & 85 & 80 \\ 72 & 80 & 85 & 100 & 85 \\ 69 & 72 & 80 & 85 & 100 \end{pmatrix} \right)$$

Discuss with your peers whether you notice any pattern to the correlation structure over time. See if you can come up with a rule that describes the correlation pattern.

ANSWER: The correlation for SF-36 mental health scores measured 1 month apart is 0.85, the correlation for scores measured 2 months apart is 0.80, the correlation for scores measured 3 months apart is 0.72, and the correlation for scores measures 4 months apart is 0.69. So the correlation depends on the time between the measurements (the lag) and this correlation pattern is known as the Toeplitz model where a correlation is defined for each lag.

