Longitudinal Data Analysis: Introduction

Arnab Maity

NCSU Department of Statistics ~ 5240 SAS Hall ~ 919-515-1937 ~ amaity[at]ncsu.edu

Contents

Introduction

The most straightforward design for a longitudinal study consists of a random sample of subjects/items, where multiple observations correspond to a variable observed at multiple follow-up times. Longitudinal data analysis refers to statistical techniques for studying the behavior of the variable of interest *over time*. The need often arises in agriculture and the life sciences, medical and public health research, and physical science and engineering, among other fields.

Recall that longitudinal data are different from multivariate data, as the **order of the repeated measurements is essential in the analysis of longitudinal data**, whereas permuting the order of the variables in multivariate analysis yields the same results.

Consider the following examples. Pay attention to what is the response variable, what is the observational unit, how many measurements are collected per unit.

Example 1: Treatment of Lead Exposed Children (TLC) Trial

The TLC trial was a placebo-controlled, randomized study of a chelating agent (succimer) in children with blood lead levels of 20-44 micrograms/dL.[1] We only consider a subsample of size 50 from the children who received succimer. The dataset consists of four repeated measurements of blood lead levels obtained at baseline (week 0), week 1, week 4, and week 6 on each of the 50 children.

[1] The dataset and its description are available at `https://content.sph.harvard.edu/fitzmaur/ala2e/`

```
tlc <- read.table("data/lead-data.txt", header = F)
colnames(tlc) <- c("ID", "Week 0", "Week 1", "Week 4",
    "Week 6")
tlc[1:3, ]
```

```
##   ID Week 0 Week 1 Week 4 Week 6
## 1  1   26.5   14.8   19.5   21.0
## 2  2   25.8   23.0   19.1   23.2
## 3  3   20.4    2.8    3.2    9.4
```

Each row in the matrix corresponds to one child, and each column corresponds to one time point. A plot of the dataset is shown in Figure 1; each line corresponds to one child.

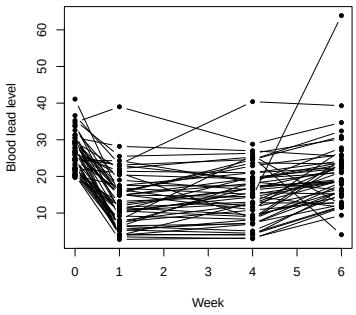

Figure 1: Blood lead levels of 50 Children over weeks in the TLC data. Each profile corresponds to one child.

Notice that we used "week" as the x-axis variable to account for the unequal spacing of the weeks (0, 1, 4, 6) to properly display the data. In this example,

- **response variable:** blood lead levels,
- **observational unit:** a child,
- **number of measurements collected per unit:** 4.

Example 2: Six Cities Study of Air Pollution and Health

The dataset contains a subset of the pulmonary function data[2] collected in the Six Cities Study. The data consist of all measurements of log-transformed Forced Expiratory Volume (FEV1)[3], height and age obtained from a randomly selected subset of size 300 of the female participants living in Topeka, Kansas.

[2] The dataset and its description are available at `https://content.sph.harvard.edu/fitzmaur/ala2e/`

[3] FEV1 is the maximal volume of air one can forcefully exhale in one second.

```
fev <- read.table("data/fev1-data.txt", header = F)
colnames(fev) = c("id", "Height", "age", "height_base",
    "age_base", "logFEV1")
head(fev)
```

```
##   id Height     age height_base age_base logFEV1
## 1  1   1.20  9.3415         1.2   9.3415 0.21511
## 2  1   1.28 10.3929         1.2   9.3415 0.37156
## 3  1   1.33 11.4524         1.2   9.3415 0.48858
## 4  1   1.42 12.4600         1.2   9.3415 0.75142
## 5  1   1.48 13.4182         1.2   9.3415 0.83291
## 6  1   1.50 15.4743         1.2   9.3415 0.89200
```

Notice the format in which this dataset is displayed is somewhat different from before. Here each row corresponds to *one girl and one time point for that girl*. Thus all the rows which have "id" value of 1 correspond to the same girl, and so on. This is because the variable "age" takes the role of the time component. Since each girl was measured at different ages, it is not reasonable to put the data in the format as in the previous example (where each column was a time point). Figure 2 shows a histogram of the ages at which the measurements were taken for the girls. For example, the ages of the first two girls (with id 1 and 2) are shown below.

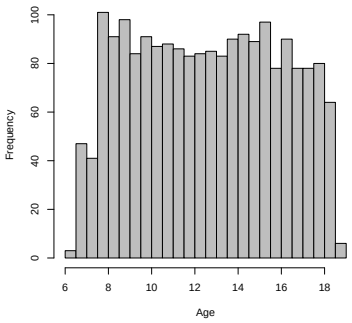

Figure 2: Ages at which measurements were taken in the air pollution data.

```
# ID=1
fev$age[fev$id == 1]
```

```
## [1]  9.3415 10.3929 11.4524 12.4600 13.4182
## [6] 15.4743 16.3723
```

```
# ID=2
fev$age[fev$id == 2]
```

```
## [1]  6.5873  7.6496 12.7392 13.7741 14.6940
## [6] 15.8220 16.6680 17.6318
```

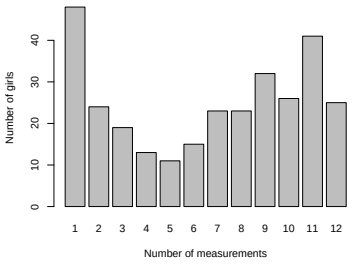

Figure 3: Number of measurements taken per girl.

Thus the number of measurements taken for each girl, and the time of measurements vary.[4] The distribution of the number of measurements is shown Figure 3.

[4] This is different from the TLC example, where every child was measured exactly four times.

Consider the variable `logFEV1` for this discussion. We plot `logFEV1` against `age` of each individual (top panel), and the same plot for a subset of the data for better visualization (bottom panel) of Figure 4. In this example,

- **response variable:** log(FEV1),
- **observational unit:** a girl,
- **number of measurements per unit:** varies between 1 to 12.

Our goal is to build models for analyzing such longitudinal data. We shall focus on the following questions:

1. How does the population mean response vary over time?
2. How does the population variance (standard deviation) behave over time?
3. How to incorporate the correlations between observations within subjects?
4. If covariates are presemt, how to model the effects of covariates on mean, variance and correlation?

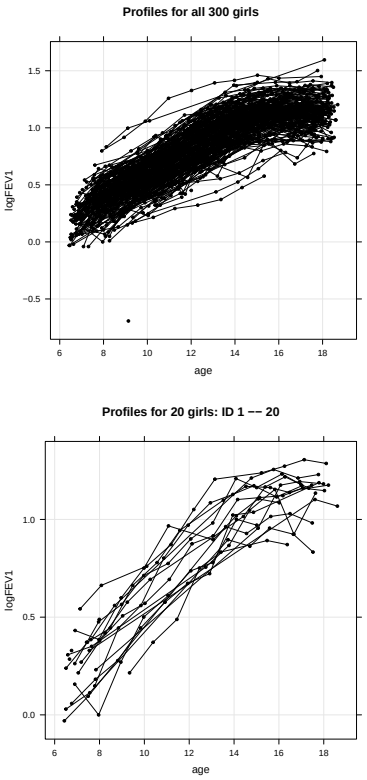

Figure 4: Plot of log(FEV1) vs. Age for all the girls (top panel) and for only girls with ID = 1 to 20 (bottom panel).

Basic concepts and notations

Let us first introduce some important concepts and notation that will be used throughout the course.

- *Response* is the outcome of interest;
- *Unit* is the object/subject on which measurements are taken;
- *Time* is the generic term for the condition of measurement.

We define

$$Y_{ij} = \text{the } j\text{th measurement taken on the } i\text{th subject or unit,}$$

$$t_{ij} = \text{time at which the } j\text{th measurement on the } i\text{th unit is taken,}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Here we use n to denote the total number of units and m_i is the number of measurements for unit i.[5]

Response vector

We can define the *response vector* of unit i as the *random vector*

$$\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^T.$$

We assume the responses to be *independent across units* ($\boldsymbol{Y}_i$ and $\boldsymbol{Y}_{i'}$, $i \neq i'$, are independent). However, the responses are *correlated within each unit*, that is, Y_{ij} and $Y_{ij'}$, $j \neq j'$, are correlated.

[5] Although not specified explicitly, it is assumed that times occur in an increasing order $t_{i1} < t_{i2} < \ldots < t_{im_i}$.

Time is an important covariate in longitudinal data. Both the mean and the covariance of the response vector, $E(\boldsymbol{Y}_i)$ and $cov(\boldsymbol{Y}_i)$, may depend on the time variable. Depending on the observation times and number of observations, we define different types of designs:

- We say the design is **balanced** when $m_i = m$ (same number of repeated measurements per unit) and $t_{ij} = t_j$ (measurements for each unit taken at the same time) for all the units. Otherwise we say the design is *unbalanced*. In our two examples, the TLC trial data is balanced, but the six cities air pollution data is unbalanced.
- We say the design is **equally spaced** if the time between consecutive observations, $t_{j+1} - t_j$, is the same for all j and subjects (e.g., week 1, 2, 3 and 4). Both of our examples are not equally spaced.
- The design is **balanced with missing data** if the design is balanced but each subject is only measured on a subset of possible times. For example, we may have planned to measure children's weight on years $0, 1, \ldots, 5$, but one child only has data on years 1, 2 and 5, and another on years 0, 1, 2, and 4.

General data structure for a balanced design ($m_i = m$ and $t_{ij} = t_j$) is:[6]

[6] This is the format used in Example 1: TLC data.

	t_1	t_2	t_3	$\ldots$	t_m
Units					
1	Y_{11}	Y_{12}	Y_{13}	$\ldots$	Y_{1m}
2	Y_{21}	Y_{22}	Y_{23}	$\ldots$	Y_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
n	Y_{n1}	Y_{n2}	Y_{n3}	$\ldots$	Y_{nm}

For an unbalanced (or balanced with missing data) design, we may use the "long format":[7]

[7] This is the format used in Example 2: air pollution data.

Unit	Time	Response
1	t_{11}	Y_{11}
$\vdots$	$\vdots$	$\vdots$
1	t_{1m_1}	Y_{1m_1}
	$\vdots$	
n	t_{n1}	Y_{n1}
$\vdots$	$\vdots$	$\vdots$
n	t_{nm_n}	Y_{nm_n}

The long format is also applicable for the balanced and equally spaced case as well. If a covariate is present, we simply add extra columns. For example, in the six cities data, "baseline height" and "baseline weight" are covariates that do not change over time. On the other hand, "height" is a covariate that changes over time.

Inferences about longitudinal data

Consider the observed data: $\{(Y_{ij}, t_{ij}) : j = 1, \ldots, m_i\}, i = 1, \ldots, n$, where Y_{ij} *is assumed to be continuous*. For simplicity we assume $t_{ij} = t_j$ and $m_i = m$ (*balanced* design).[8] We are interested in studying the typical behavior of the outcome over time.[9]

[8] Think of the TLC data in example 1.

[9] The most general inference we can make is on the *distribution* of $\mathbf{Y}_i$ as a function of time. This is generally very difficult. Therefore we restrict our attention to simpler quantities such as mean and variance etc.

Estimating mean and variance functions

There are two types of profiles: population and subject level profiles. The *population profile* is the plot of the mean response against time. This plot indicates how the mean response changes over time and what type of models we should use to model the mean function mathematically. The *subject profile* is a plot of the response of one individual against time. Often each individual may show trends different from the overall average. We need to model these patterns carefully for proper inference.

Since data are collected over multiple time points, the population mean is not a single number. Instead, the population mean can be thought of as a *function of time*: at time t_j, the mean response is denoted as $\mu_j = \mu(t_j)$. We can estimate $\mu(t_j)$ by the sample mean of the data observed at t_j,

$$\widehat{\mu}(t_j) = \bar{Y}_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}.$$

In general, $\mu(t)$ can be constant over time, or have a trend over time. It can also depend on other covariates. We will discuss such scenarios later.

Simularly, the population variance (standard deviation) is also a function of time. At time t_j, the population variance is $\sigma^2(t_j)$ (standard deviation is $\sigma(t_j)$), and can be estimated by the sample variance (standard deviation) of the data observed at t_j:

$$\widehat{\sigma}^2(t_j) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\bullet j})^2, \quad \widehat{\sigma}(t_j) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\bullet j})^2\right]^{1/2}.$$

In our TLC data example, the estimated mean (top panel) and variance (bottom panel) profiles are shown in Figure 5. We also plot the subject level profiles in gray in the top panel.

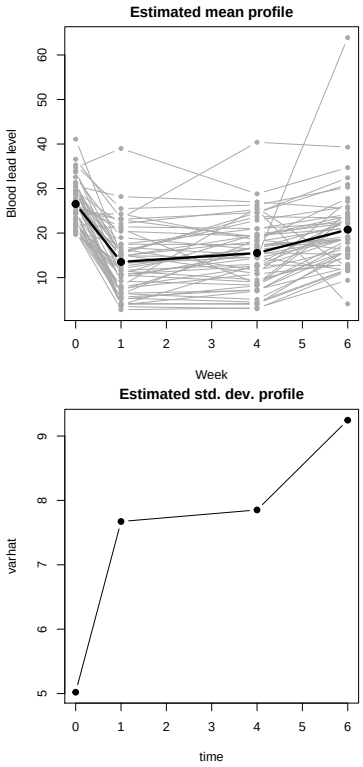

Figure 5: Mean and standard deviation profiles for TLC data.

Covariance and correlation

Since there are multiple measurements for each subject, there might be correlation between these measurements. For any subject i, the measurements Y_{ij} and $Y_{i\ell}$, taken at t_j and t_ℓ, can be correlated. The

covariance, and the correlation, are denoted as

$$\sigma(t_j, t_\ell) = cov(Y_{ij}, Y_{i\ell}), \qquad \rho(t_j, t_\ell) = \sigma(t_j, t_\ell)/\{\sigma(t_j)\sigma(t_\ell)\},$$

respectively. These quantities can be estimated by the sample covariance and correlation, respectively:

$$\widehat{\sigma}(t_j, t_\ell) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\bullet j})(Y_{i\ell} - \bar{Y}_{\bullet \ell}), \qquad \widehat{\rho}(t_j, t_\ell) = \frac{\widehat{\sigma}(t_j, t_\ell)}{\widehat{\sigma}(t_j)\widehat{\sigma}(t_\ell)}.$$

In general, an unbiased estimator for the population covariance is the sample covariance[10]

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T;$$

the (j, ℓ)-th element of $\widehat{\mathbf{\Sigma}}$ is $\widehat{\sigma}(t_j, t_\ell)$.

[10] Recall the definition of sample covariance from multivariate data.

For the TLC data, the estimated correlation coefficients are shown below.

```
##         Week 0 Week 1 Week 4 Week 6
## Week 0  1.000  0.401  0.384  0.495
## Week 1  0.401  1.000  0.731  0.507
## Week 4  0.384  0.731  1.000  0.455
## Week 6  0.495  0.507  0.455  1.000
```

Such plots and summary statistics help us to decide how to model the polulation mean and variances for a dataset. In the TLC dataset, the mean profile plot (top panel of Figure 5) indicates that the mean function is not simply linear in time – we need to use other formulations for $\mu(t)$. Perhaps a "piecewise liner function" would be appropriate here.

From the standard deviation profile (bottom panel of Figure 5), we see that we have non-constant standard deviation across time – standard deviation at week 6 is almost twice that standard deviation at week 0. Thus equal variance assumption across time is not reasonable here.

From the estimated correlation matrix, we see that obervations from weeks 1 and 4 are highly correlated compared to other weeks. Thus simple structures like "equal correlation among weeks" might not be appropriate here.

Unbalanced data

When we have unbalanced design (where time of measurements are random) or balanaced with many missing data, we may not have enough data at a given time to calculate the mean or standard deviation. For example, take the six cities pollution data in Example

2. At time point $t = 7$, we have no data points at all. One way to bypass this issue is to *pool responses taken from observations with similar times* to calculate summaries. Specifically, for a given time t, we will take all the Y_{ij} whose measurement times are within a window $[t - h/2, t + h/2]$, where h is a pre-specified *window width*. The mean and sd at t will be computed on these observations.

For the six cities data, suppose we want to estimate $\mu(t)$ at $t = 7$. Let us take the window width $h = 1$. Thus we will take all the observation whose measurement times fall in the window $[7 - 1/2, 7 + 1/2] = [6.5, 7.5]$.

```
# subset the data
sub <- fev$logFEV1[fev$age>=6.5 & fev$age<=7.5]
# number of observations to pool
length(sub)
```

```
## [1] 88
```

```
# mean
mean(sub)
```

```
## [1] 0.2434999
```

Thus we have 88 observations within the specified window, and averaging those gives us $\hat{\mu}(7) \approx 0.24$. Overall, the mean and standard deviation profiles are shown in Figure 6. We see that the mean profile is roughly linear for the most part execpt at the end where it becomes constant. The standard deviation profile shows us that the standard deviations are almost constant over age.

We typically pick h so that we have enough data to reasonably estimate mean or sd – do not choose h so large that the estimates become meaningless.

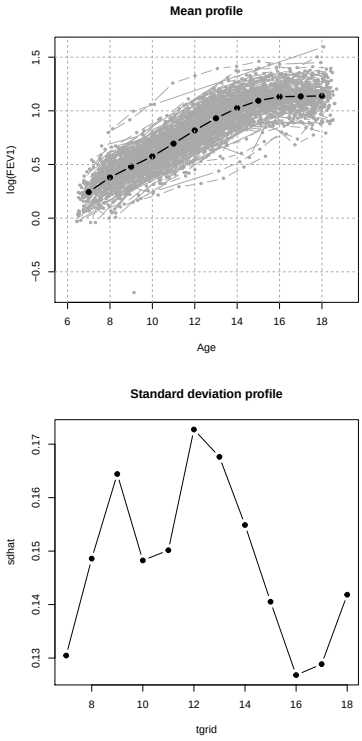

Figure 6: Mean and standard deviation profiles of the six cities airpollution data.