Statistical Methods for Life History Analysis

STAT 437

Winter 2022 (1221)[1]

Cameron Roopnarine[2] Dylan Spicker[3]

13th January 2022

[1] Online Course
[2] LaTeXer
[3] Instructor

Contents

Chapter 1

What are Longitudinal Data?

WEEK 1
5th to 7th January

1.1 Introduction

Syllabus.

1.2 What are Longitudinal Data?

NIH RESEARCH MATTERS

April 27, 2021

Lack of sleep in middle age may increase dementia risk

> At a Glance
>
> - People who slept six hours or less per night in their 50s and 60s were more likely to develop dementia later in life.
> - The findings suggest that inadequate sleep duration could increase dementia risk and emphasize the importance of good sleep habits.

What would a study **need** to look like to conclude this?

1.2.1 The Design of a Longitudinal Study

- Can we conclude this by taking a sample of elderly individuals directly?
  - **No**. How do we determine how much they slept 20 years prior?
- Can we conclude this by taking a sample of middle-aged individuals directly?
  - **No**. How do we determine who will develop dementia later on?
- Can we conclude this by taking independent samples of middle-aged individuals *and* elderly individuals?
  - **No**. How do we pair the individuals?

We would *need* to be able to follow individuals, starting when they are middle-aged, recording information like how often they sleep, and continue following them until the onset of dementia.

This is a longitudinal study.

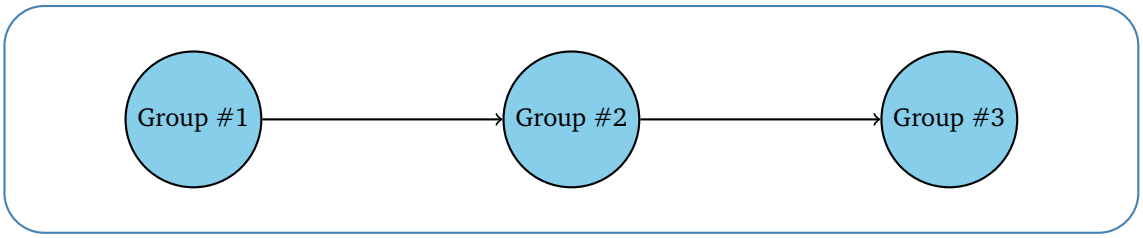

Figure 1.1: Longitudinal Study

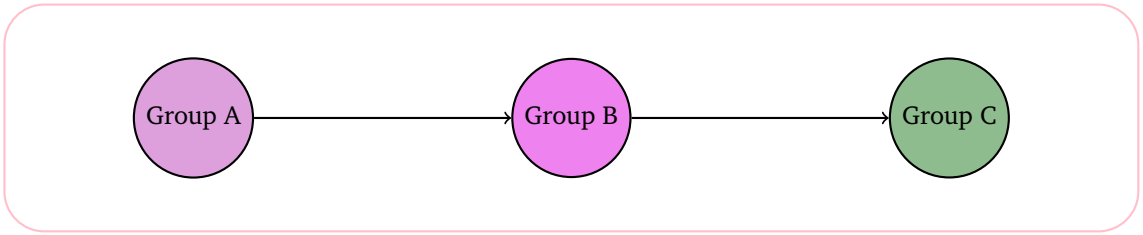

Figure 1.2: Cross-Sectional Study

A research study in which **subjects are followed over time**. Typically, this involves repeated measurements of the same variables. Longitudinal studies differ from **cross-sectional** studies and **time series** studies.

1.2.2 Uses for Longitudinal Studies

- To detect *changes* in outcomes, both at the population and individual level.
- **Longitudinal effects** as compared to **cohort effects**.
- Correctly ascertain the exposures.
- Understand different sources of variation
- **Between**- and **within**-subject variation.
- To detect **time effects**, both directly and as interactions with other relevant factors.

Bottom line: There are many questions of interest which can only be answered using longitudinal data. We should probably learn how to analyse it.

1.2.3 Why are Longitudinal Data Special?

What makes longitudinal data more difficult to analyse?

- The data are **correlated**.
- Everyone's favourite assumption (assume that $X_1, \ldots, X_n$ are iid) will **not** hold.
- Now what? STAT 437.

1.2.4 Example Datasets

TLC Trial

ID	Treatment	W0	W1	W4	W6
1	P	30.8	26.9	25.8	23.8
2	A	26.5	14.8	19.5	21
3	A	25.8	23	19.1	23.2
⋮	⋮	⋮	⋮	⋮	⋮
98	A	29.4	22.1	25.3	4.1
99	A	21.9	7.6	10.8	13
100	A	20.7	8.1	25.7	12.3

- Is there a difference between **placebo** and **treatment**?
- How does the blood lead level **change over time** (in each group)?
- Is the **change** over time **equal** between treatment groups?

Sales Data

DATE	brand	prod	QTY	PROMO
2014-01-02	1	1	7	0
2014-01-02	1	2	3	0
2014-01-02	1	3	0	0
⋮	⋮	⋮	⋮	⋮
2018-12-31	4	8	1	1
2018-12-31	4	9	0	0
2018-12-31	4	10	3	1

- Are the **different brands comparable** in terms of overall sales?
- Are the **different products comparable**?
- Do **promotions increase** the quantity sold? If so, **by how much**?
- Do the effects of time, and promotion, **change by brand** or product?

Podcast Data

Rating	No. Reviews	Title	Date	⋯
4.9	6400	Dissect	2019-11-01	⋯
4.9	26300	The Adventure Zone	2019-11-01	⋯
4.8	3700	Song Exploder	2019-11-01	⋯
⋮	⋮	⋮	⋮	⋮
4.2	1100	Finding Fred	2019-12-01	⋯
3.9	648	Inside Frozen 2	2019-12-01	⋯
4.6	6400	Pop Culture Happy Hour	2019-12-01	⋯

- Can we **predict** the number of ratings that a podcast will receive over time?

- Can we **predict** the average rating value that a podcast will receive over time?

Stroke Data

year	Prop. $(0,0)$	Prop. $(0,1)$	Prop. $(1,0)$	Prop. $(1,1)$
1	$57/344$	$17/72$	$17/79$	$5/23$
2	$27/287$	$8/55$	$9/62$	$4/18$
3	$23/260$	$8/47$	$5/53$	$3/14$
⋮	⋮	⋮	⋮	⋮
8	$10/129$	$1/15$	$5/23$	$1/4$
9	$17/119$	$3/14$	$4/18$	$0/3$
10	$13/102$	$1/11$	$2/14$	$0/3$

- $0 =$ placebo treatment, $1 =$ active treatment; $0 =$ no previous stroke, $1 =$ previous stroke.
- This is **time to event** data.
- What is **probability of surviving** beyond some point?
- Does this **differ** if you previously had a stroke? If you **received treatment**?

1.2.5 Summary

- Longitudinal data occur when we take repeated measurements on the same individuals over time.
- Longitudinal data are required for answering questions about changes within an individual (compared to between individuals) and to capture time effects.
- Longitudinal data are challenging to work with because the data are correlated.

1.3 Exploring Longitudinal Data (Application)

R Demo.

1.4 Notation for Longitudinal Data (Theory)

General Notation

- Random variables: X, Y, Z.
  - Realizations of these random variables: x, y, z.
- Unknown parameters: θ, β, α.
  - Estimates of these parameters with "hat:" $\hat{\theta}$, $\hat{\beta}$, $\hat{\alpha}$.
- Transpose of a matrix $\boldsymbol{X}$: $\boldsymbol{X}^\top$.

Individual Notation

- Individual outcome for individual i at time j: Y_{ij}, where $i = 1, \ldots, n$ are the individuals, and $j = 1, \ldots, k_i$ are the time points. We may also use Y_{it_j} to denote the outcome for individual i at time t_j when more complex times are used.
- Individual variate: X_{ijk}, where i and j index over individuals and times, respectively, and k indexes over the different variates of interest.

Suppose for an individual that we measure age, treatment, and symptom status. We have $k = 3$ since we have three variables.

- Usually, X_{ijk} will not change over time, so we may write $X_{ijk} = X_{ij'k}$ for all j and j'. Usually $X_{ij1} = 1$ to include the intercept in our models. However, if a variate is time-changing, then we need to be more careful about $X_{ij1} = 1$.
- For an individual, define $\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik_i} \end{bmatrix}_{k_i \times 1} \equiv (Y_{i1}, Y_{i2}, \ldots, Y_{ik_i})^\top$ to be a vector of outcomes.
- For variates, take $\boldsymbol{X}_{ij} = \begin{bmatrix} X_{ij1} & X_{ij2} & \cdots & X_{ijp} \end{bmatrix}_{1 \times p} \equiv (X_{ij1}, X_{ij2}, \ldots, X_{ijp})$, where p different variates are measured.
- Define $\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \\ \vdots \\ \boldsymbol{X}_{ik_i} \end{bmatrix}_{p \times k_i}$ to be a matrix containing of all the variates.
- In certain contexts, we may write $\boldsymbol{Y}_i$ as a row vector or to take the transpose of $\boldsymbol{X}_i$.

1.4.1 Notation and Considerations for Time

- Time for the i^{th} individual at the j^{th} measurement: t_{ij}.
  - Sometimes, we take $t_{ij} = j$, where j is an index of visits.
  - If the scale of time is related to calendar time, we may have $t_{i1} = 0$ and $t_{i2} = 14$ to indicate the first visit and second visit are two weeks apart, where time is measured in days.
- The design is **balanced** if $t_{ij} = t_{i'j}$ for all i and i'. In this case, we drop subscript i from the times and write $t_1, \ldots, t_k$. We will often consider balanced designs, but this is not necessary.

1.5 What is Linear Regression (Review/Theory)

The Ordinary Least Squares Estimators

- Suppose $\boldsymbol{Y}_i$ are continuous, and we want to model $\mathbb{E}[\boldsymbol{Y}_i \mid \boldsymbol{X}_i]$.
- A **linear regression model** takes
$$\mathbb{E}[\boldsymbol{Y}_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i\boldsymbol{\beta}.$$
- We take
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y},$$
and we call these **ordinary least squares** (OLS) estimators.

OLS Estimators (Two Ways)

- If $\boldsymbol{Y}_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$, then the OLS estimators are the **maximum likelihood estimators**.
- If we take $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ is non-normal, then the OLS estimators are simply the best (in terms of *mean squared error*) predictor of $\boldsymbol{\beta}$.

Assumptions for OLS

1. The conditional mean is **linear** (in parameters).
2. All values of $\boldsymbol{Y}_i$ have constant variance, denoted σ^2 (conditionally).
3. The $\boldsymbol{Y}_i$ are **independent**.

Asymptotic Analysis

- As $n \to \infty$, $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\boldsymbol{\beta}, \mathsf{Var}(\hat{\boldsymbol{\beta}})\big)$, where

$$\mathsf{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

  We can use this result for **confidence intervals** and **hypothesis tests**.

Summary

- Linear Regression allows us to estimate a functional form for the conditional mean of a continuous outcome.
- The OLS estimators are valid MLE-type estimators when normality is assumed, and are LS estimators otherwise.
- The asymptotic analysis is valid in large samples, regardless of distributional assumptions, and can be used for Wald-type analysis.

1.6 Why Can't We Just Use Regression? (Linear Marginal Models)

Stated Mathematically

We want to **fit a model** that gives $\mathbb{E}[Y_{ij} \mid \boldsymbol{X}_{ij}, t_{ij}]$ in terms of **interpretable parameters**.

Let's use an example!

ID	Trt	W0	W1	W4	W6	ID	Trt	time	W
1	P	30.8	26.9	25.8	23.8	1	P	1	30.8
2	A	26.5	14.8	19.5	21	2	A	1	26.5
3	A	25.8	23	19.1	23.2	3	A	1	25.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	A	29.4	22.1	25.3	4.1	98	A	4	4.1
99	A	21.9	7.6	10.8	13	99	A	4	13
100	A	20.7	8.1	25.7	12.3	100	A	4	12.3

- Consider the TLC trial data, in **wide format** (left-hand side) and then in **long format** (right-hand side).
- In the right-hand side we have an outcome (W), with two explanatory factors $\big(\{\texttt{Trt}, \texttt{time}\}\big)$.
  - We want $\mathbb{E}[W \mid \texttt{Trt}, \texttt{time}]$. **Is this familiar**?

Using Linear Regression

We can fit the model in R, using `lm`. **Is this valid**?

	Estimate	Std. Error	$\mathbb{P}(>\|t\|)$
`(Intercept)`	26.540	0.937	0.000
`time2`	−13.018	1.325	0.000
`time3`	−11.026	1.325	0.000
`time4`	−5.778	1.325	0.000
`TreatmentP`	−0.268	1.325	0.840
`time2:TreatmentP`	11.406	1.874	0.000
`time3:TreatmentP`	8.824	1.874	0.000
`time4:TreatmentP`	3.152	1.874	0.093

What does this `lm` imply about our data?

- There is a **linear conditional mean structure**:

$$\begin{aligned}\mathbb{E}[W_{ij} \mid \texttt{Trt}_i, t_j] = \beta_0 + \beta_1 \texttt{Trt}_i + \beta_2 \mathbb{I}\{t_j = 2\} + \beta_3 \mathbb{I}\{t_j = 3\} + \beta_4 \mathbb{I}\{t_j = 4\} \\ + \beta_5 \texttt{Trt}_i \mathbb{I}\{t_j = 2\} + \beta_6 \texttt{Trt}_i \mathbb{I}\{t_j = 3\} + \beta_7 \texttt{Trt}_i \mathbb{I}\{t_j = 4\}.\end{aligned}$$

- There is **constant variance** such as $\mathsf{Var}(W_{ij}) = \sigma^2$ for all i and j.
- The values of W_{ij} are **independent**. However, this assumption is clearly violated.

What makes longitudinal data special?

Longitudinal data are characterized by **correlation** *within* individuals.

TODO figure Therefore, the previous `lm` will work **only if** we are willing to assume that the observations are **independent**.

Longitudinal Data as Multivariate Data

How can we adapt linear regression to allow for this association?

- When the data are in **long format**, it appears that the outcomes are univariate.
- When the data are in **wide format**, we can view the outcome as a vector of outcomes, (e.g., $\boldsymbol{W} = (W_0, W_1, W_4, W_6)$).
- The analysis of longitudinal data is **multivariate analysis**.
  - This accounts for the **lack of independence** in the outcomes!

Multivariate Normal

Instead of assuming that $Y_{ij} \sim \mathcal{N}(\boldsymbol{X}_{ij}\beta, \sigma^2)$, what if we took

$$\boldsymbol{Y}_i \sim \text{MVN}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)?$$

Recall: The multivariate normal (MVN) has a density given by

$$f(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})^\top \right\}.$$

Linear Marginal Models

- In this proposal, we specify a **linear form** for the conditional mean.
  - That is, $\mathbb{E}[\boldsymbol{Y}_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i\beta$, where $\boldsymbol{X}_i$ is a matrix and $\boldsymbol{Y}_i$ is a vector.
- We allow for **correlation** through the individual covariance matrix, $\boldsymbol{\Sigma}_i$.
- We could (theoretically) find the **MLE** under the assumption of multivariate normality.

Covariance Matrix

Recall that $\mathsf{Cor}(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}$, and so, re-arranging,

$$\mathsf{Cov}(X,Y) = \mathsf{Cor}(X,Y)\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}.$$

Moreover, recall that a variance/covariance matrix is

$$\mathsf{Cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i = \begin{bmatrix} \mathsf{Var}(Y_{i1}) & \mathsf{Cov}(Y_{i1},Y_{i2}) & \cdots & \mathsf{Cov}(Y_{i1},Y_{ip}) \\ \mathsf{Cov}(Y_{i2},Y_{i1}) & \mathsf{Var}(Y_{i2}) & \cdots & \mathsf{Cov}(Y_{i2},Y_{ip}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(Y_{ip},Y_{i1}) & \mathsf{Cov}(Y_{ip},Y_{i1}) & \cdots & \mathsf{Var}(Y_{ip}) \end{bmatrix}.$$

Covariance Matrix Simplification

If **we assume** that $\mathsf{Var}(Y_{ij}) = \sigma^2$ for all i, j, *and* we denote $\mathsf{Cor}(Y_{ij}, Y_{i\ell}) = \rho_{j\ell}$ for all i, then note that

$$\mathsf{Cov}(Y_{ij}, Y_{i\ell}) = \mathsf{Cor}(Y_{ij}, Y_{i\ell})\sqrt{\mathsf{Var}(Y_{ij}\,\mathsf{Var}(Y_{i\ell}))} = \sigma^2\rho_{i\ell}.$$

We write

$$\boldsymbol{R}(\boldsymbol{\rho}) = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}.$$

With this notation,

$$\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{R}(\boldsymbol{\rho}).$$

Linear Marginal Models

Under the previous specification we can find the MLE to be

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{Y}_i.$$

For the variance parameter, we get

$$\sigma^2 = \frac{1}{nk}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{R}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

then we can solve numerically for $\boldsymbol{R}_i$. We want to model $\mathbb{E}[Y_{ij} \mid \boldsymbol{X}_i]$ (for some purpose) and so we specify a **multivariate linear model**. By assuming that the variance **is constant across different times**, and we can accommodate the correlation expected within each individual.

The multivariate normality assumption gives us a process for computing the MLE, which can produce estimates for the parameters of interest, denoted $\hat{\boldsymbol{\beta}}$.

Next Steps

- How can we conduct **inference** on the estimated parameters? (Why do we want to?)
- How can we specify **time trends** in the model for the mean?
- How can we use this model to answer **scientific questions of interest**?
- What can we do about the **correlation matrix**? (Are there any shortcomings with our assumptions?)

Asymptotic Normality

It can be shown that, asymptotically

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}\big(\boldsymbol{\beta}, \mathsf{Var}(\hat{\boldsymbol{\beta}})\big),$$

where

$$\mathsf{Var}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{\sigma^2}\sum_{i=1}^{n} \boldsymbol{X}_i^{-1}\boldsymbol{R}_i^{-1}\boldsymbol{X}_i\right),$$

which can be estimated by plugging in $\hat{\sigma}^2$ and $\hat{\boldsymbol{\rho}}$. We get

$$\mathsf{se}(\hat{\beta}_j) = \left[\widehat{\mathsf{Var}}(\hat{\boldsymbol{\beta}})\right]^{1/2}_{(j,j)}.$$

Inference based on Wald Statistics

As a result,

$$\frac{\hat{\beta}_j - \beta_j}{\mathsf{se}(\hat{\beta}_j)} \sim \mathcal{N}(0,1).$$

This can be used to test H_0: $\beta_j = \beta^\star$, or for confidence intervals, **just like with linear regression**! An equivalent expression is

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\mathsf{Var}(\hat{\beta}_j)} \sim \chi^2_1.$$

Time as a Covariate

- Generally speaking, we can simply include **time** as a **covariate** in the model.
- If the data are **balanced** and there are *relatively few* time points, we can include it as a factor.
- If the data are **not balanced** or there are *too many* time points, we can include it as a continuous variable.
  - We can also include **quadratic** time trends, or **logarithmic** time trends, or any other functional form.
- We can include time as **calendar time**, **time since baseline**, **index of time point**, **age**, etc.
  - This will depend on what we have **measured** and what we are **interested** in.

The **choice** of how we include time will be dictated **both** by the *available* data, and by the **scientific questions of inquiry**. This goes for the **form** it takes in the model, and the **timescale** that we choose to use.