Statistical Methods for Life History Analysis

STAT 437

Winter 2022 (1221)[1]

Cameron Roopnarine[2] Dylan Spicker[3]

9th January 2022

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

Contents

1 What are Longitudinal Data? 2

1.1 What are Longitudinal Data? . . . 2

1.1.1 The Design of a Longitudinal Study . . . 2

1.1.2 Uses for Longitudinal Studies . . . 3

1.1.3 Why are Longitudinal Data Special? . . . 3

1.1.4 Example Datasets . . . 4

1.1.5 Summary . . . 5

1.2 Exploring Longitudinal Data (Application) . . . 5

Chapter 1

What are Longitudinal Data?

WEEK 1
5th to 7th January

1.1 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?

1.1.1 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.

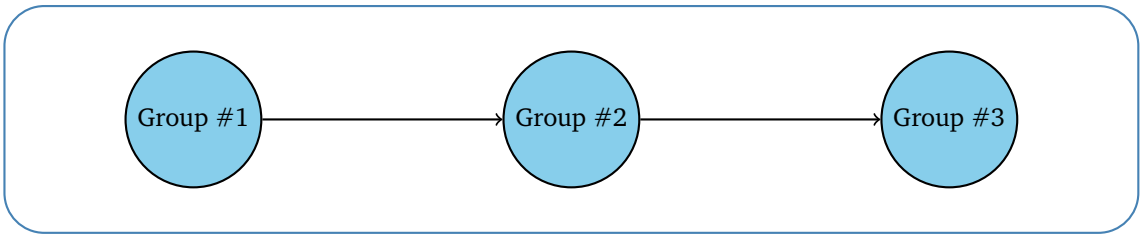


Figure 1.1: Longitudinal Study

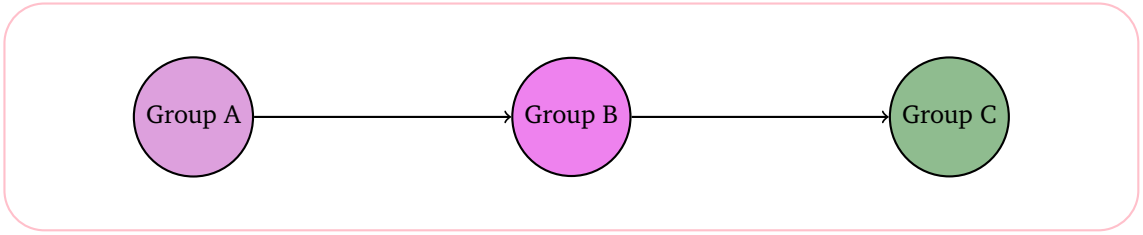


Figure 1.2: Cross-Sectional Study

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.

1.1.2 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.

1.1.3 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.

1.1.4 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**?

1.1.5 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.

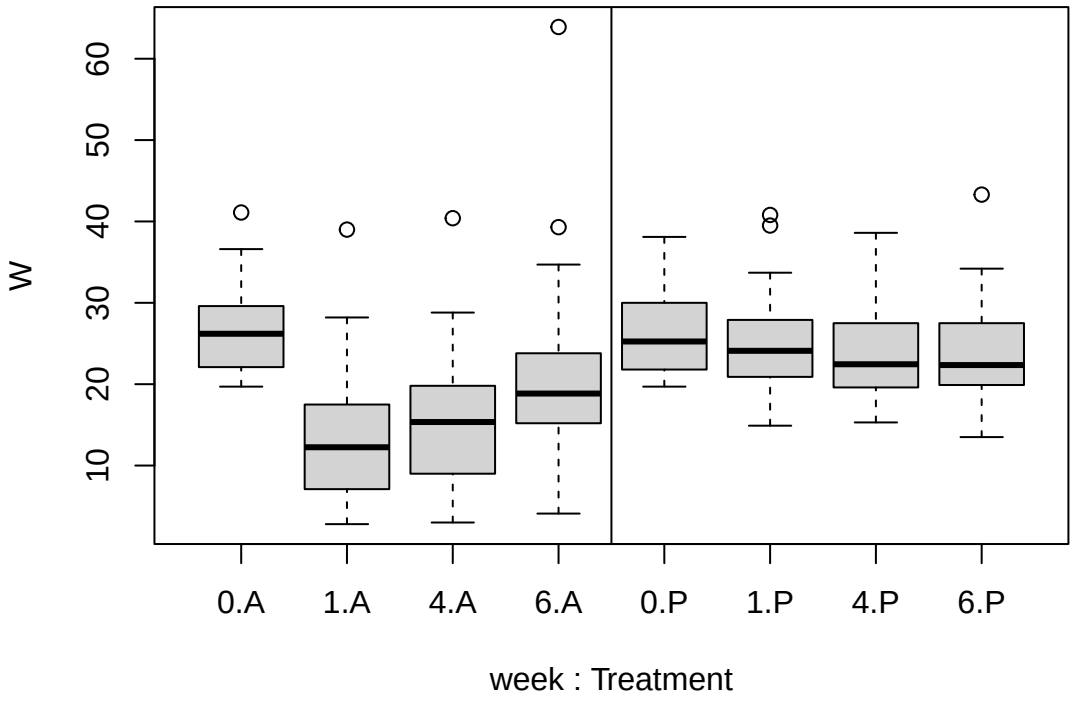
1.2 Exploring Longitudinal Data (Application)

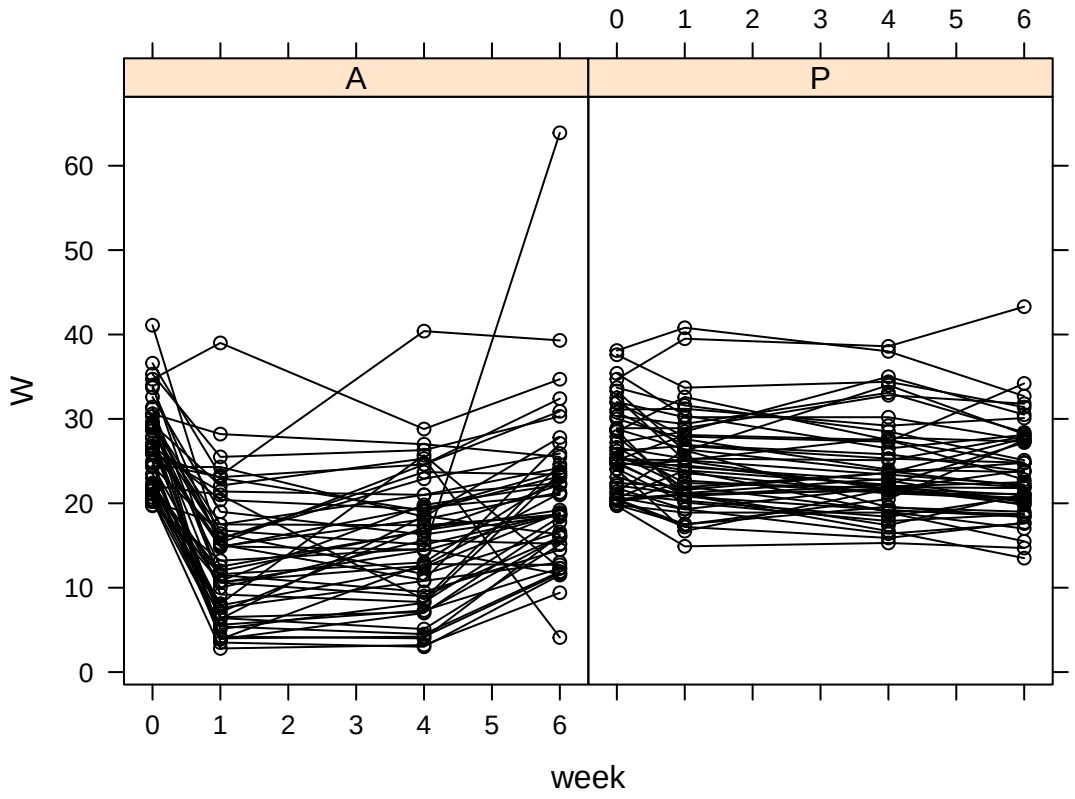
```r
# Read in the TLC Data Note: This is stored for me at
# data/TLC/TLC.csv, you should update for yourself
TLC <- read.csv("data/TLC/TLC.csv")
head(TLC)  # Outputs the first few rows of the data to take a look
```

```
  ID Treatment   W0   W1   W4   W6
1  1         P 30.8 26.9 25.8 23.8
2  2         A 26.5 14.8 19.5 21.0
3  3         A 25.8 23.0 19.1 23.2
4  4         P 24.7 24.5 22.0 22.5
5  5         A 20.4  2.8  3.2  9.4
6  6         A 20.4  5.4  4.5 11.9
```

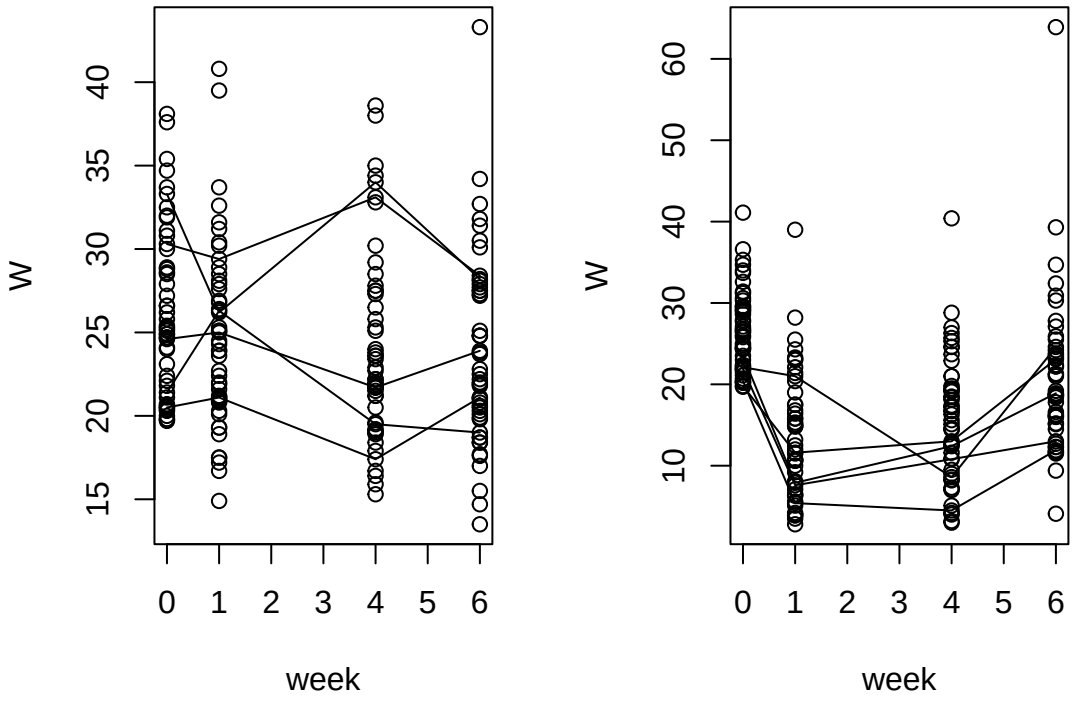
```r
# Convert from 'wide' to 'long' and back again, using
# reshape.  If you're interested, you can also use
# `pivot_wider` and `pivot_longer` from the tidyverse (If
# that doesn't mean anything to you, feel free to ignore
# it!)
TLC_long <- reshape(data = TLC, varying = c("W0", "W1", "W4",
  "W6"), timevar = "week", idvar = "ID", times = c(0, 1, 4,
  6), direction = "long", sep = "")
```

```
TLC_wide <- reshape(data = TLC_long, timevar = "week", v.names = "W",
  idvar = "ID", times = c(0, 1, 4, 6), direction = "wide",
  sep = "")
# Create a Basic Boxplot to get a Sense of the Data
boxplot(W ~ week + Treatment, data = TLC_long)
abline(v = 4.5)  # Abline v=... draws a vertical line at 4.5
# Start with an xyplot This requires the package 'lattice'
# You can install using: install.packages('lattice')
lattice::xyplot(W ~ week | Treatment, data = TLC_long, groups = ID,
  col = "black", type = c("l", "p"))
```



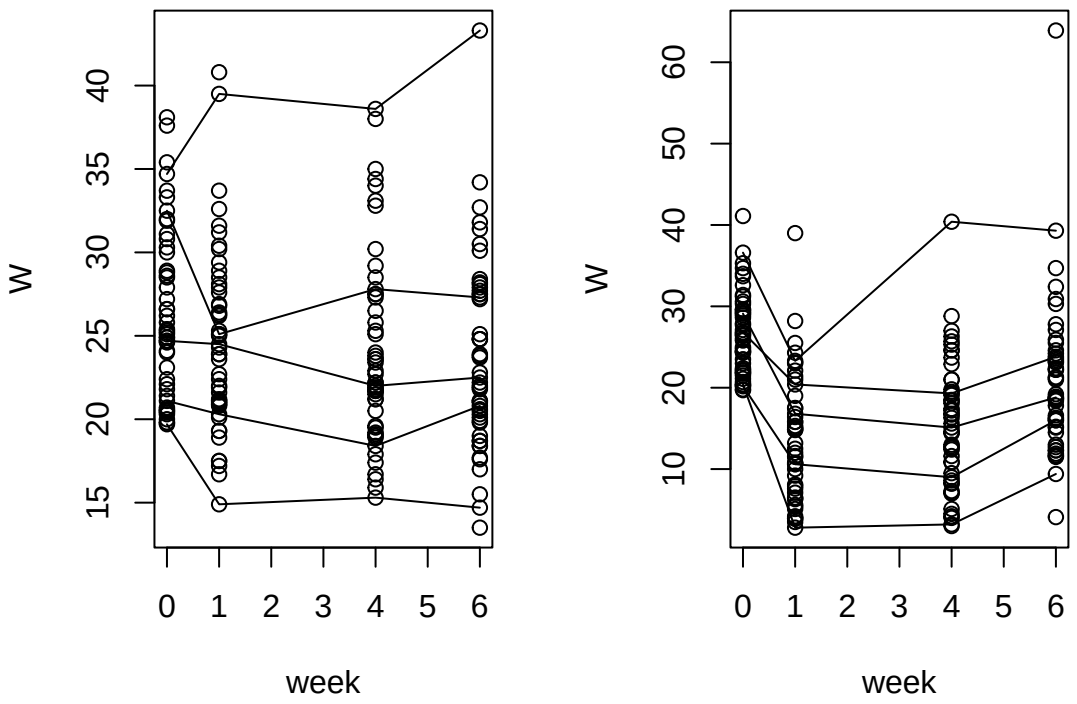



```
# The plot is a mess, as-is, so instead we can subset!
plot_num <- 5  # Select a fixed number
# This is Just Randomly Sampling from Each Group
random_samples_P <- sample(unique(TLC_long$ID[which(TLC_long$Treatment ==
  "P")]), size = plot_num, replace = FALSE)
random_samples_A <- sample(unique(TLC_long$ID[which(TLC_long$Treatment ==
  "A")]), size = plot_num, replace = FALSE)
## Actually Draw the Plots
par(mfrow = c(1, 2))
plot(W ~ week, data = TLC_long, subset = (Treatment == "P"))
for (rid in random_samples_P) {
  # Loop through the Random Points and Draw the
  # Corresponding Lines
  lines(W ~ week, data = TLC_long, subset = (ID == rid), type = "l")
}
# Repeat it for Active Treatment
plot(W ~ week, data = TLC_long, subset = (Treatment == "A"))
for (rid in random_samples_A) {
  lines(W ~ week, data = TLC_long, subset = (ID == rid), type = "l")
}
```

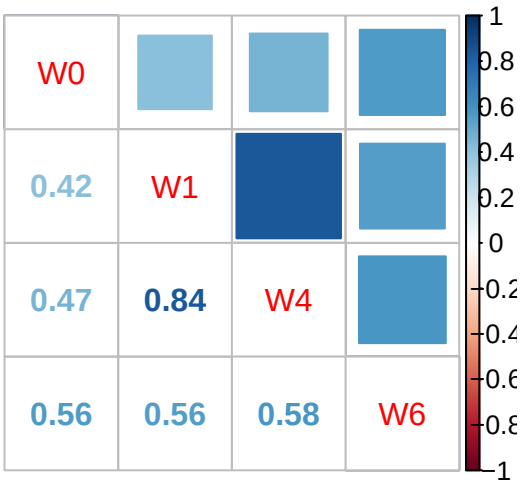



```
### Is there Smarter way of plotting?  What if we ordered
### by the median observation?
TLC_wide$median <- apply(TLC_wide[c("W0", "W1", "W4", "W6")],
  MARGIN = 1, FUN = median)  # Generate the Medians
TLC_long <- reshape(data = TLC_wide, varying = c("W0", "W1",
  "W4", "W6"), timevar = "week", idvar = "ID", times = c(0,
  1, 4, 6), direction = "long", sep = "")  # Reshape to long again, with the Median
# Sort the Data By The Medians
sorted_medians_P <- sort(TLC_wide$median[which(TLC_wide$Treatment ==
  "P")])
sorted_medians_A <- sort(TLC_wide$median[which(TLC_wide$Treatment ==
  "A")])
plot(W ~ week, data = TLC_long, subset = (Treatment == "P"))
for (row_num in floor(seq(1, 50, by = 12.25))) {
  # Here we are looping over a sequence of (1,50) by 12.5
  # which selects out every 12.5-th individual from the
  # dataset There are 50 in each group so this is
  # essentially grabing the quantiles
  rid <- TLC_wide$ID[which(TLC_wide$median == sorted_medians_P[row_num])][1]
  lines(W ~ week, data = TLC_long, subset = (ID == rid), type = "l")
}
plot(W ~ week, data = TLC_long, subset = (Treatment == "A"))
for (row_num in floor(seq(1, 50, by = 12.25))) {
  rid <- TLC_wide$ID[which(TLC_wide$median == sorted_medians_A[row_num])][1]
  lines(W ~ week, data = TLC_long, subset = (ID == rid), type = "l")
```

```
}
```




```
# This is a basic correlation plot It requires the
# 'corrplot' library, which can be installed with
# install.packages('corrplot')
corrplot::corrplot.mixed(cor(TLC_wide[c("W0", "W1", "W4", "W6")]),
  lower = "number", upper = "square")
```


W0
0.42
W1
0.47
0.84
W4
0.56
0.56
0.58
W6
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1