Why OLS Is A Bad Model For Longitudinal Data

Andy Grogan-Kaylor

2024-10-22

Table of contents

1 A Beginning Idea

> "The language we have in that world is not large enough for the territory that we've already entered." (Whyte and Tippett 2016)

2 An Empirical Example

Figure 1: Happiness as a Function of Time and Pizza

3 Introduction

We are all familiar with the idea of:

$y_i = \beta_0 + \beta_1 x + e_i$ (OLS)

get substantive example

Table 1: Data in WIDE format

id	x1	x2	x3	y1	y2	y3
1						
2						
3						

4 A First Longitudinal Model

We could imagine a longitudinal model where we regress y_i at time 2 on y_i at time 1….

$y_{i2} = \beta_0 + \beta_1 x + \beta_2 y_{i1} + e_i$

And we could even make this (*perhaps confusingly*) a multilevel model for individual i in social unit j:

$y_{i2j} = \beta_0 + \beta_1 x + \beta_2 y_{i1j} + u_{0j} + e_{ij}$

… and add all of the usual random slope terms…

5 What About Change Scores?

$y_{i2} - y_{i1} = \beta_0 + \beta_1 x + e_i$

What Happens To The Regression Coefficients in a Change Score Model?

βy_{i1}

6 What If We Have More Than Two Time Points?

$y_{i3} = \beta_0 + \beta_1 x + \beta_2 y_{i1} + \beta_3 y_{i2} + e_i$

> **Tip**
>
> What is the problem here? We have 2 terms that are likely to be collinear:
> β_2 & β_3

This issue only becomes worse the more time points we add.

As a result, we are not really modeling y_2 and y_1.

7 Two Conceptual Diagrams

7.1 OLS or MLM for 2 Timepoints

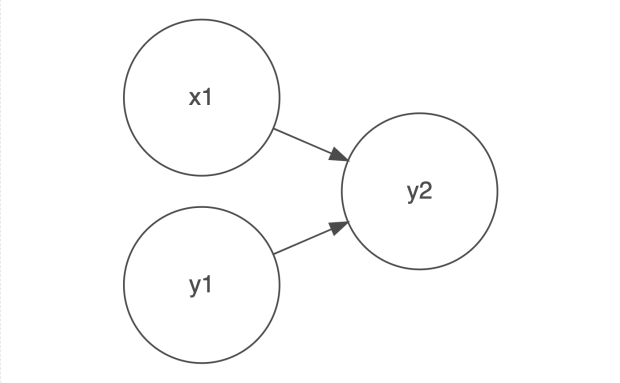

Figure 2: An OLS Or Multilevel Model For 2 Timepoints

7.2 Cross-Lagged Model

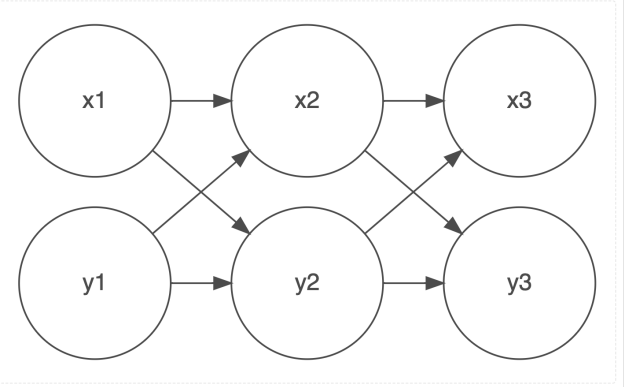

Figure 3: A Cross Lagged Model For 3 Timepoints

8 Additionally ...

No Explicit Function of Time

Additionally, we do not have an explicit function of time. We don't know really have a clear idea of whether our outcome increases with time, or decreases with time. Or whether the effect is curvilinear e.g. t^2 or $\ln(t)$.

Unbalanced Data Are A Problem

Additionally, any data that is *unbalanced* i.e. study participants enter the study late, or leave the study early are going to be difficult for this kind of model to deal with.

Missing Data Are A Problem

Similarly, data that is *missing at one time point, but present at other time points*, is going to be a problem for this kind of model. (and it is going to be difficult for many of our colleagues to see how we can get around this issue.)

9 Our Answer To the Problem

We Reshape The Data and Use the SAME Notation!!!

"Mathematics is the art of giving the same name to different things." (Poincare 1908)

9.1 Data in Long Format

Table 2: Data in LONG format

id	t	x	y
1	1		
1	2		
1	3		
2	1		
2	2		
2	3		
3	1		
3	2		
3	3		

So.... we take our standard multilevel notation.

$y_{ij} = \beta_0 + \beta_1 x + u_{0j} + e_{ij}$ (Simple MLM)

cross out j write in t.

$y_{it} = \beta_0 + \beta_1 t + u_{0i} + e_{it}$ (LONGITUDINAL MLM)

> **Tip**
>
> Every row is a *person-observation* (person i observed at time t). Every person has *multiple rows*.

10 This Has The Following Advantages:

10.1 First...

1. No multicollinearity issue.
2. *Unbalanced data is less of a problem*, the data structure and estimation are robust to these possibilities.
3. *Missing data is less of a problem* (assuming *MCAR*). When a person observation is missing, that person simply has fewer rows of data. But all rows of data are "matched" to the same person by i.

10.2 How To Address Missing Data?

> **Addressing Missing Data is Complicated!!!**
>
> It is sometimes best to (a) do nothing; (b) do something complicated.

- Ignore it.
- Fill in the mean.
- Use previous observation.
- Use next observation.
- Linearly interpolate previous and next observation.
- Regression imputation.
- Multiple imputation.

10.3 Further...

3. We now have an *explicit function of time* $\beta_1 t$ and could even add $\beta_2 t^2$ or substitute $\beta \ln(t)$.
4. *Multiple time-points are not a problem.* Same algebra for 2 time points as for 10,000 time points. (Helpful when we start to think about intensive longitudinal data *e.g.* George Holden's *recording study*).
5. We are *measuring exactly the time at which events take place* for each individual. Not simply saying *Wave 1*, *Wave 2*, *Wave 3*, etc...

6. Every individual could have a *completely different set of time points* and even a *completely different number of time points*.

And we can even add βx back into the model.

> **Caution**
>
> We do need to think carefully about what is the appropriate variable for time. Is it the variable we used to reshape the data–often `wave`–or some other more appropriate metric, like `age`?

Let's continue to explore how this model works.

References

Bryk, Anthony S, and Stephen W Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods. Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications, Inc.

Hox, Jop J, Mirjam Moerbeek, and Rens van de Schoot. 2018. *Multilevel Analysis: Techniques and Applications. Multilevel Analysis: Techniques and Applications.* Third edition. Routledge, Taylor & Francis Group,.

Poincare, Henri. 1908. *Science Et Methode.* Flammarion.

Whyte, David, and Krista Tippett. 2016. "David Whyte: Seeking Language Large Enough." The On Being Project. https://onbeing.org/programs/david-whyte-seeking-language-large-enough/.