Why OLS Is A Bad Model For Longitudinal Data

Andy Grogan-Kaylor

2025-01-02

Table of contents

# 1. Some Beginning Ideas

“Despite the incredible diversity existing among and within human cultures, there are many phenomena that occur regularly in all known societies. These commonalities, or universals, while deriving in part from human nature, may also have specific social, cultural, and systemic sources. We need to develop a working understanding of these universals so that we might advance legitimate, empirically based human science set on creating knowledge that is politically relevant to fostering real solutions to the problems that complicate human co-existence in the Age of the Anthropocene.” (Antweiler 2016)

“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



Happiness as a Function of Time and Pizza

# 3. Introduction

We are all familiar with the idea of:

$y\_{i}=β\_{0}+β\_{1}x+e\_{i}$ (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….

$y\_{i2}=β\_{0}+β\_{1}x+β\_{2}y\_{i1}+e\_{i}$

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

$y\_{i2j}=β\_{0}+β\_{1}x+β\_{2}y\_{i1j}+u\_{0j}+e\_{ij}$

… and add all of the usual random slope terms…

|  |
| --- |
|  Tip |
| Any problems yet? |

# 5. What About Change Scores?

$y\_{i2}−y\_{i1}=β\_{0}+β\_{1}x+e\_{i}$

|  |
| --- |
|  What Happens To The Regression Coefficients in a Change Score Model? |
| $βy\_{i1}$ |

# 6. What If We Have More Than Two Time Points?

$y\_{i3}=β\_{0}+β\_{1}x+β\_{2}y\_{i1}+β\_{3}y\_{i2}+e\_{i}$

|  |
| --- |
|  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



An OLS Or Multilevel Model For 2 Timepoints

## 7.2 Cross-Lagged Model



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 $ln\left(t\right)$. |

|  |
| --- |
|  *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 |  |  |

 |

## 9.2 Equation

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

### 9.2.1 Simple MLM

$$y\_{ij}=β\_{0}+β\_{1}x+u\_{0j}+e\_{ij}  \left(1\right)$$

cross out *j* write in *t*.

### 9.2.2 Longitudinal MLM

$$y\_{it}=β\_{0}+β\_{1}t+u\_{0i}+e\_{it}  \left(2\right)$$

|  |
| --- |
|  Person-Observations |
| 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. By inspection of [Equation 2](#eq-longitudinal-mlm), we see that there is only a single $β$ coefficient for each variable, $∴$ no multicollinearity problem.
2. *Unbalanced data is less of a problem*, the data structure and estimation are robust to these possibilities (Singer and Willett 2003; Raudenbush and Bryk 2002).
3. *Missing data is less of a problem* (assuming *MCAR*). When a person observation is missing, that person simply has fewer rows of data (J. Hox 2010; Luke 2004; Raudenbush and Bryk 2002; Rabe-Hesketh and Skrondal 2012). 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…

1. We now have an *explicit function of time* $β\_{1}t$, and could treat time more flexibly, by creating a polynomial function of time e.g. by adding $β\_{2}t^{2}$, etc. (Raudenbush and Bryk 2002; Singer and Willett 2003). (We could even substitute $βln\left(t\right)$.)
2. Again, by inspection of [Equation 2](#eq-longitudinal-mlm), we see that *multiple or many 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*).
3. We are *measuring exactly the time at which events take place* for each individual (Singer and Willett 2003; Luke 2004). Not simply saying *Wave 1*, *Wave 2*, *Wave 3*, etc…
4. Every individual could have a *completely different set of time points* and even a *completely different number of time points* (J. Hox 2010; J. J. Hox, Moerbeek, and van de Schoot 2018; Singer and Willett 2003; Luke 2004).

And we can even add $βx$ back into the model.

## 10.4 Appropriate Metric For Time

|  |
| --- |
|  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 or time in study (Singer and Willett 2003)? |

## 10.5 Visually

|  |
| --- |
| Figure 1: A Multilevel Model For Longitudinal Data |

## 10.6 Lastly

|  |
| --- |
|  Caution |
| Generating appropriate descriptive statistics can be a problem. |

## 10.7 Let’s continue to explore how this model works.

# 11. References

Antweiler, Christoph. 2016. *Our Common Denominator: Human Universals Revisited*. Berghahn.

Hox, Joop. 2010. *Multilevel Analysis: Techniques and Applications*. 2nd ed. Routledge.

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,.

Luke, Douglas. 2004. *Multilevel Modeling*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412985147>.

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

Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata - Volume i: Continuous Responses*. *Stata Press*. 3rd ed. Stata Press.

Raudenbush, Stephen W, and Anthony S Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications.

Singer, Judith D, and John B Willett. 2003. *Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence*. *Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence*. Oxford University Press.

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/>.