Multilevel models

Dr Andi Fugard *(they/them)*

Intermediate
Quantitative Social Research

10 Feb 2020

 a.fugard@bbk.ac.uk

Today

We will cover the last new idea of the module: **random effects**

It may help to know that residuals are an example random effect

The plan for this eve

We started the module with **linear regression** for normally distributed data

and saw how it **generalises** to other distributions for

binary (yes/no) &
count data

This week: we will generalise the **residuals**

So far…

Linear model (regression)

```
lm(outcome ~ predictors, data = dat)
```

Generalised linear model (GLM)

```
glm(outcome ~ predictors, data = dat,
    family = binomial)
glm(outcome ~ predictors, data = dat,
    family = poisson)
glm(outcome ~ predictors, data = dat,
    family = quasipoisson)
```

Preview of this week's new command

Linear model (regression)

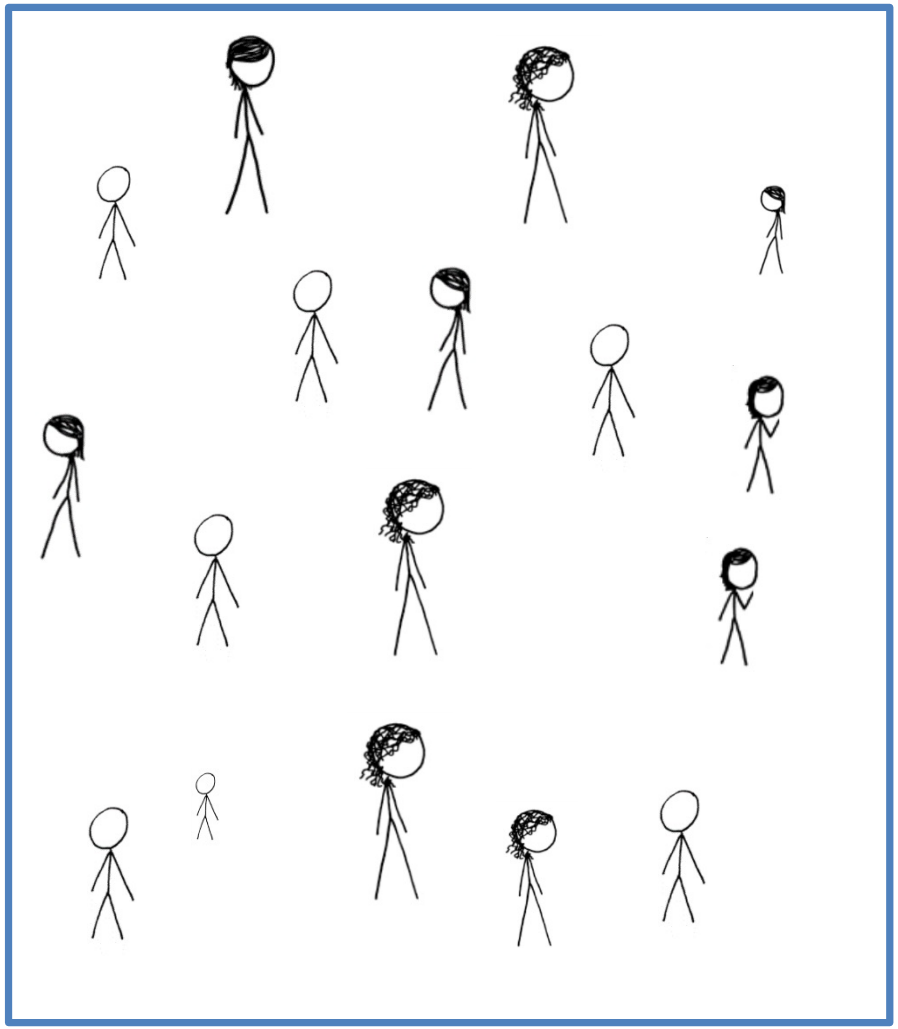
```
lm(outcome ~ predictors, data = dat)
```

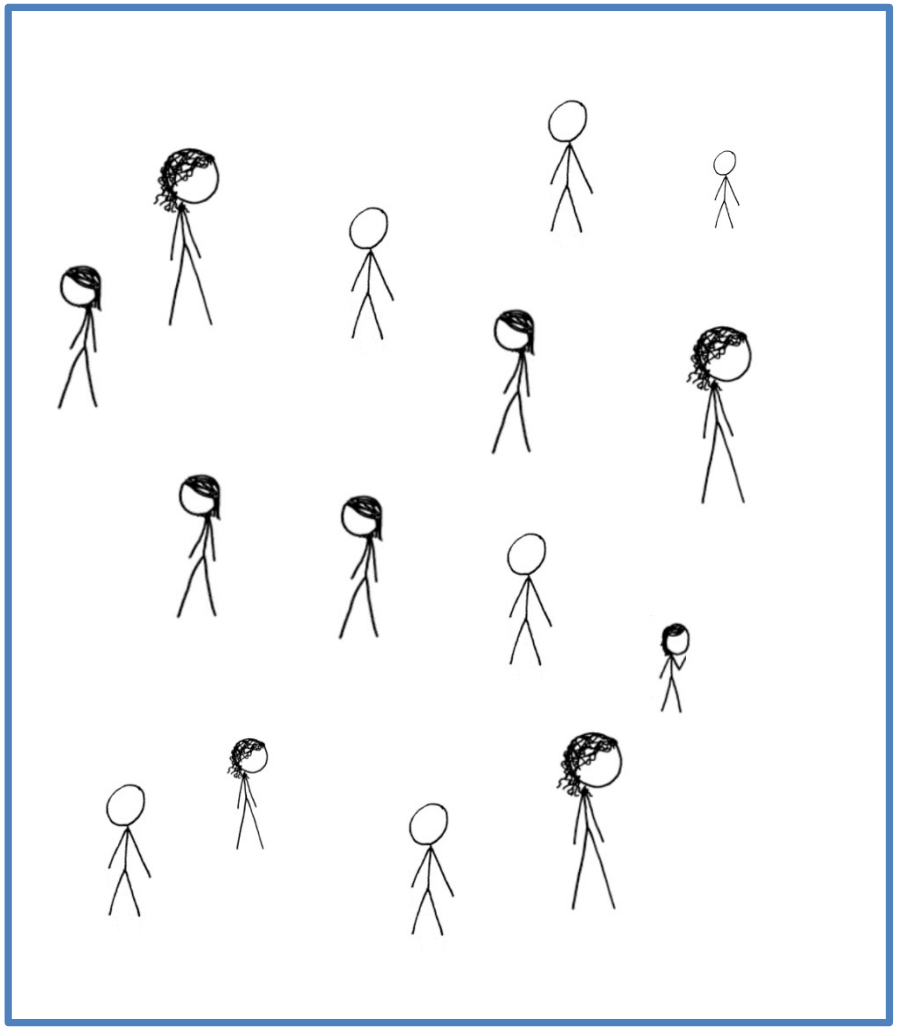
Linear mixed-effect model (LMM)

```
library(lme4)
lmer(outcome ~ predictors +
   (predictors | group), data = dat)
```

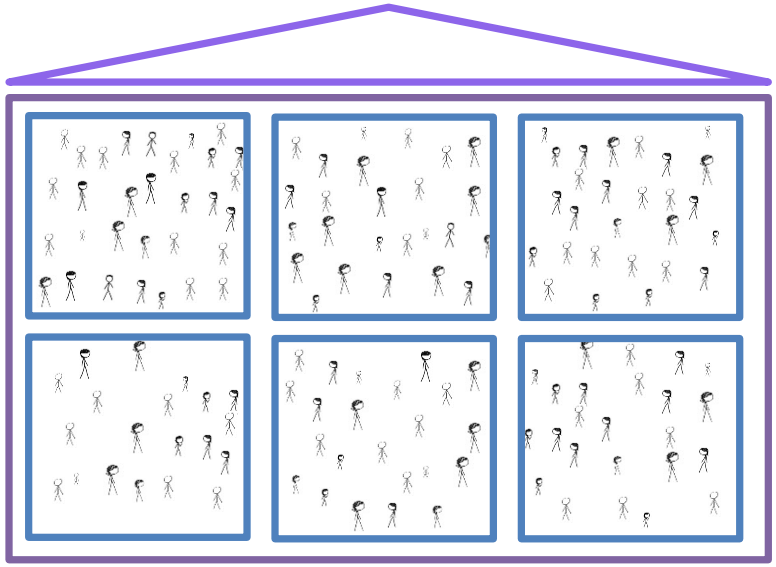
Multilevel data

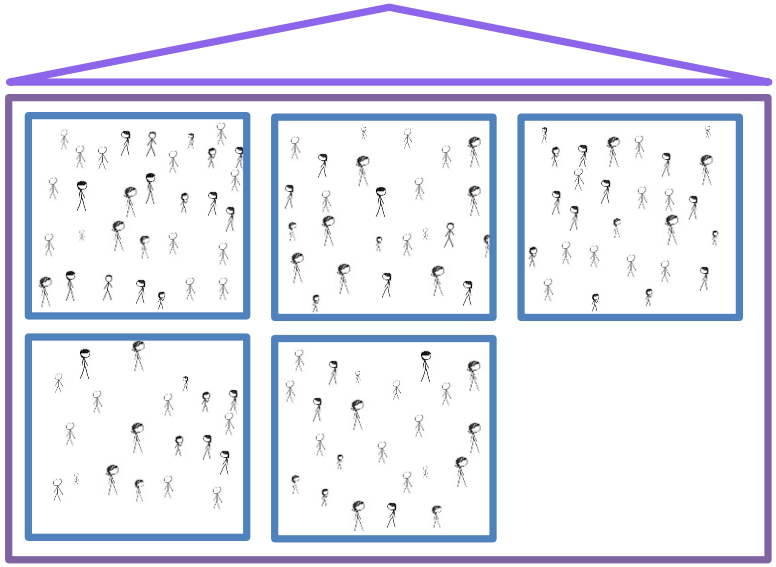
Students are in classrooms

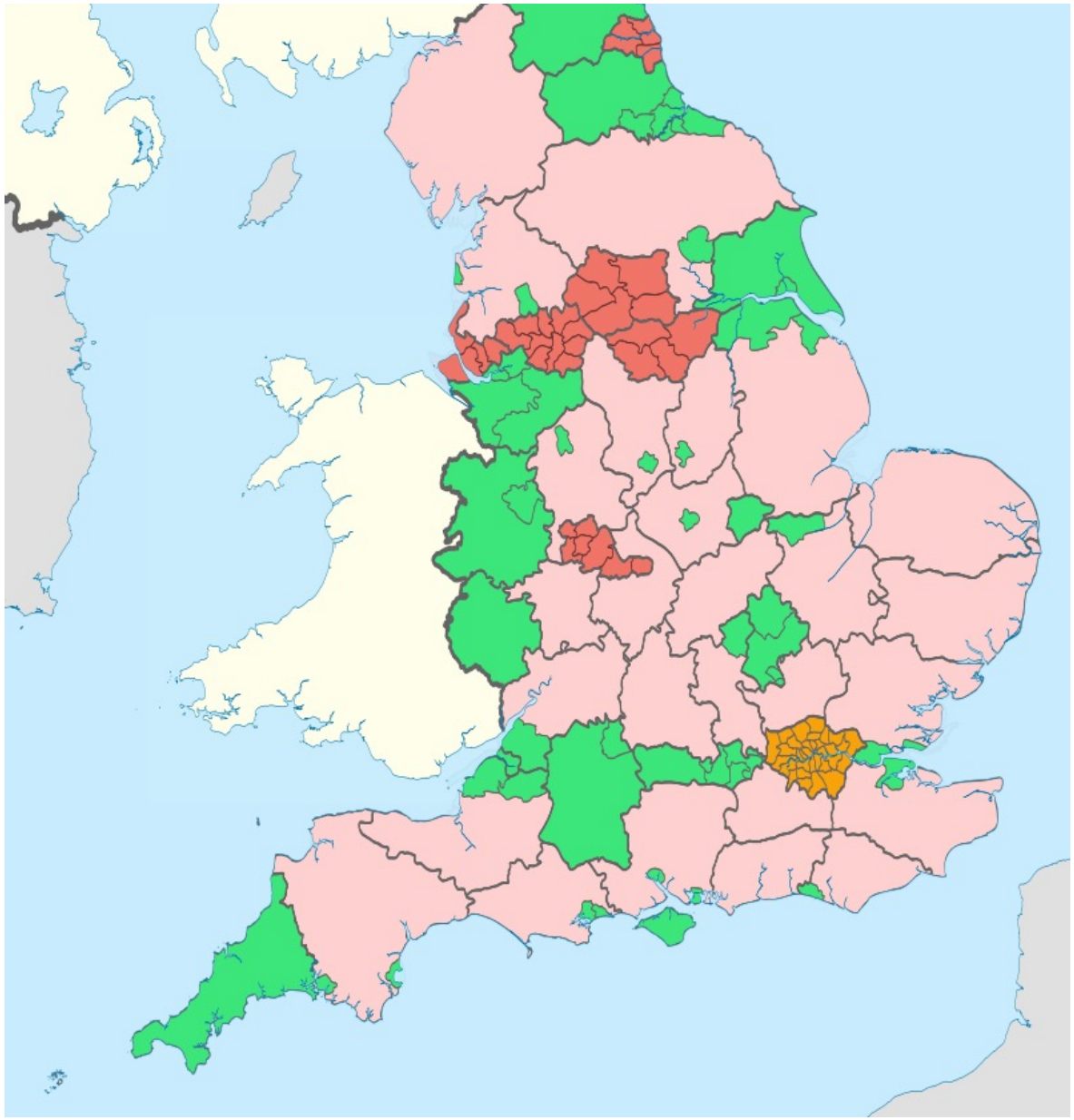


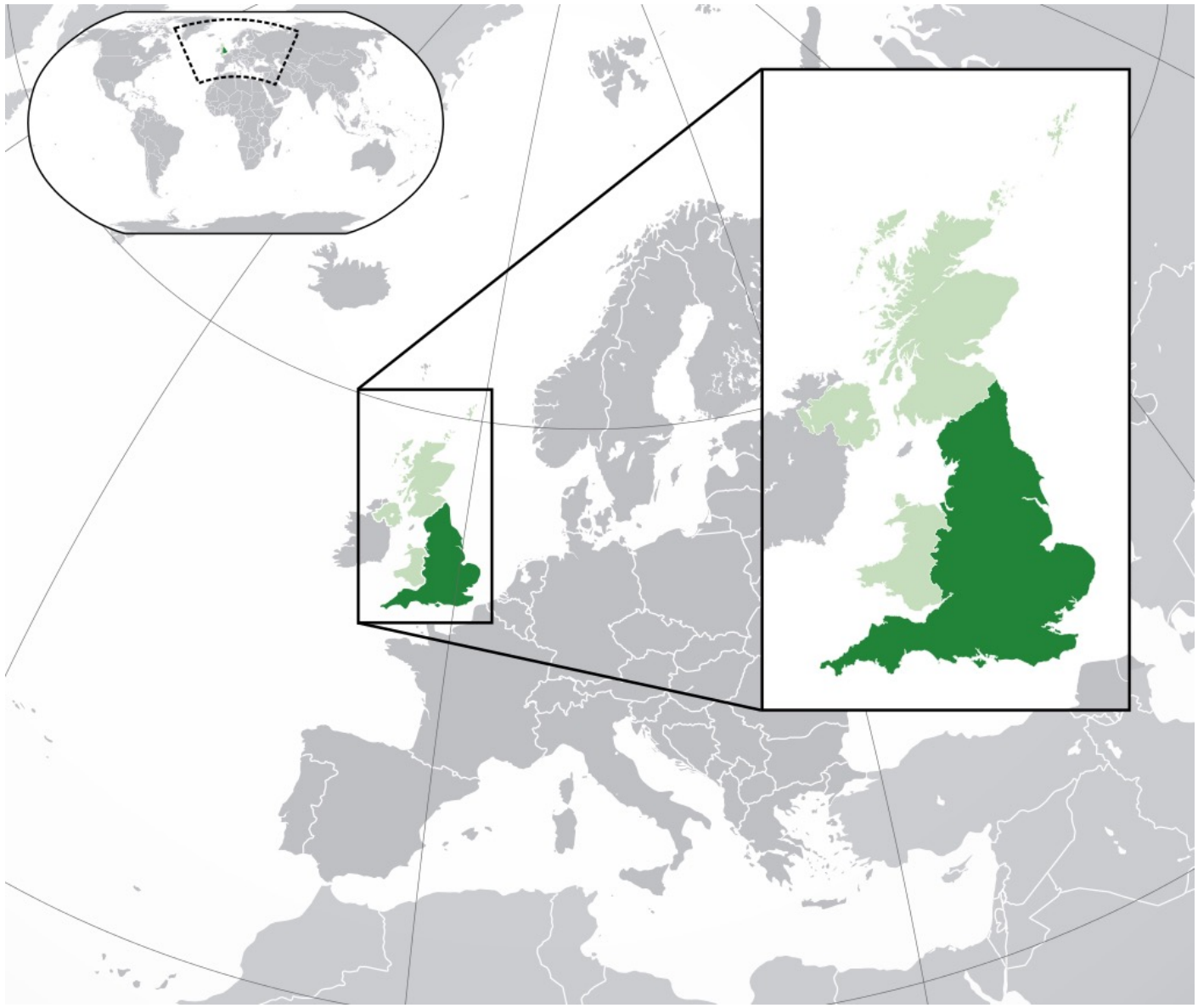


Classrooms are in schools

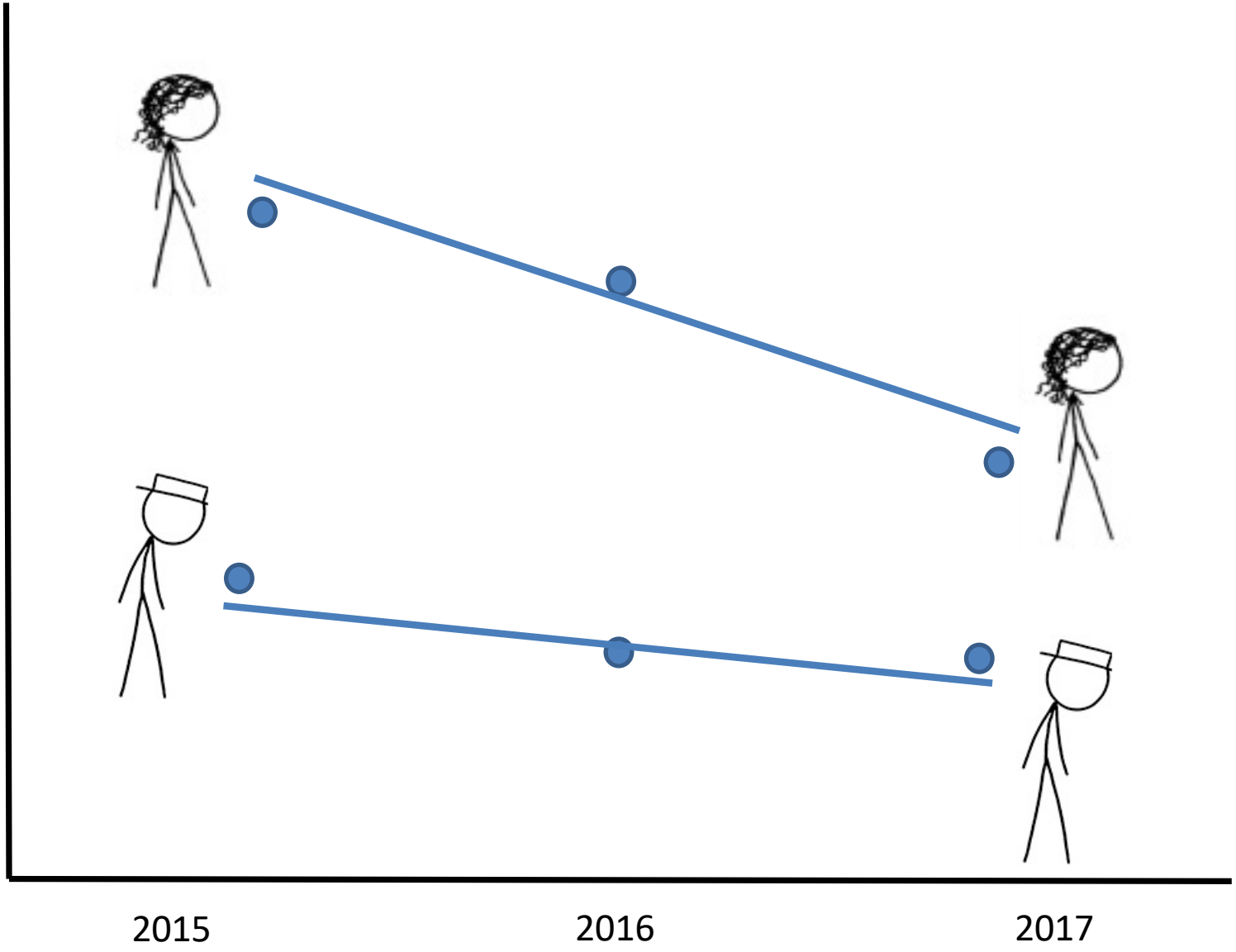








Each student can also be tracked over time (another level is “within participant”)

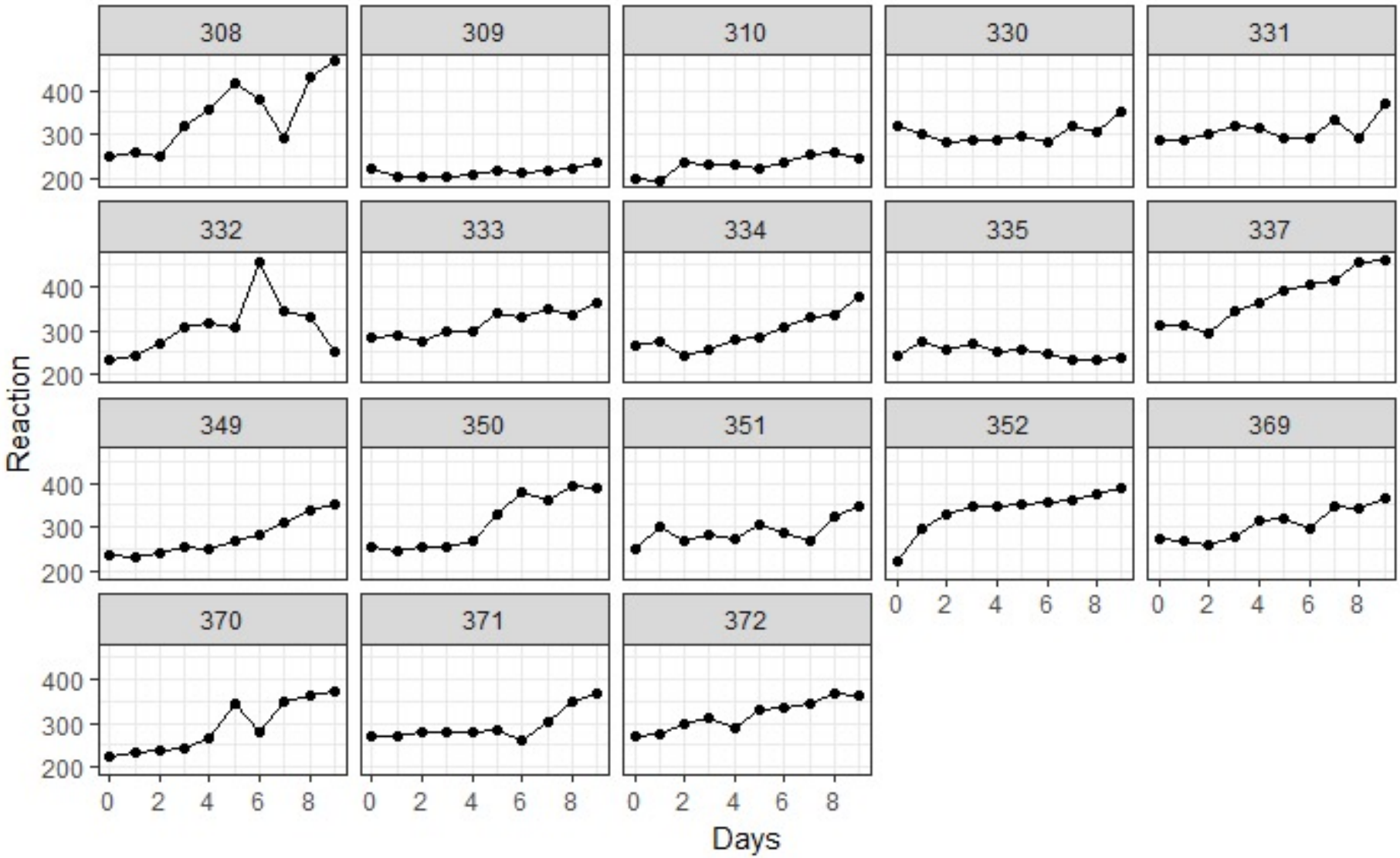


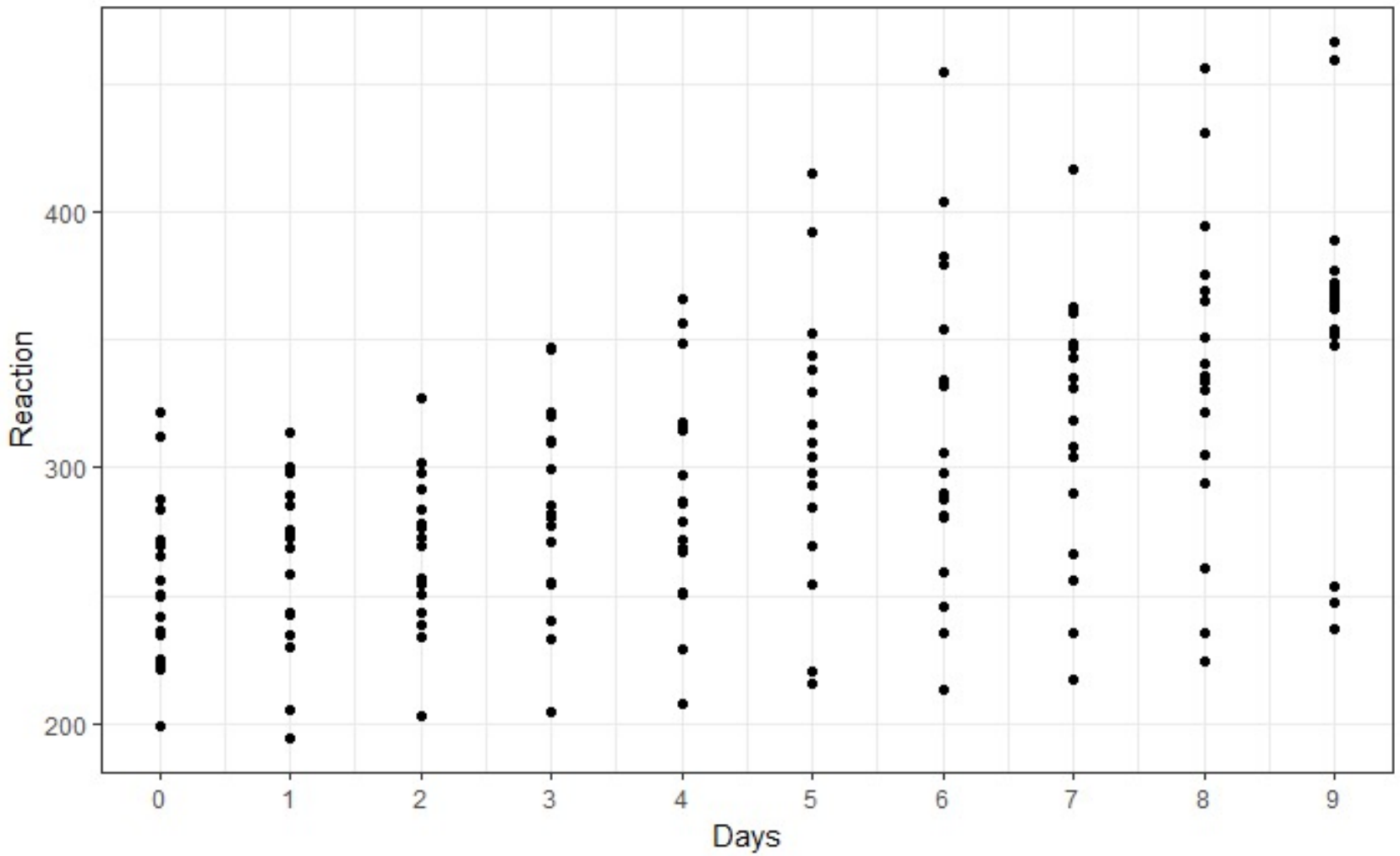
J. Sleep Res. (2003) **12,** 1–12

Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study

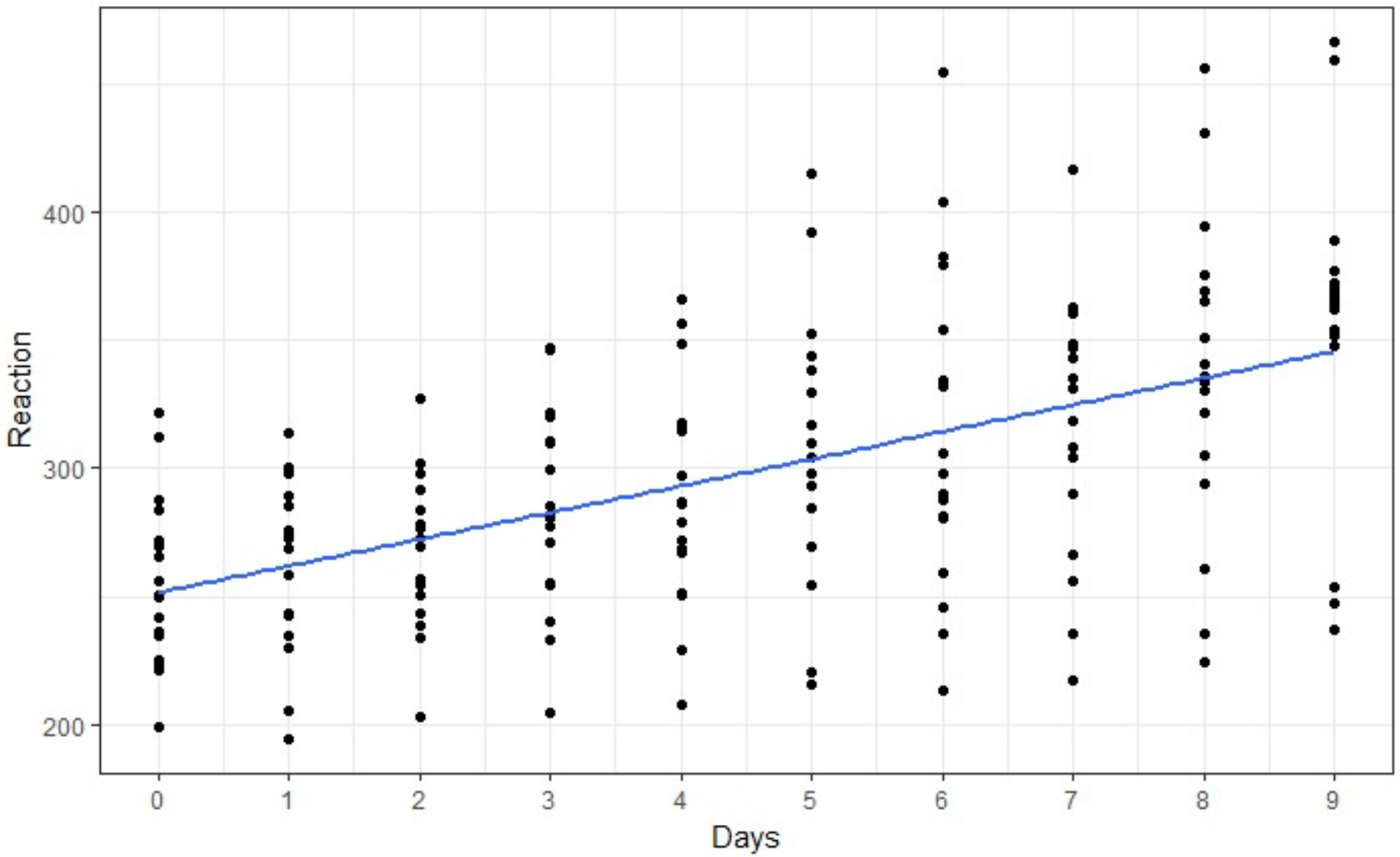
GREGORY BELENKY, NANCY J. WESENSTEN, DAVID R. THORNE, MARIA L. THOMAS, HELEN C. SING, DANIEL P. REDMOND, MICHAEL B. RUSSO and THOMAS J. BALKIN
Division of Neuropsychiatry, Walter Reed Army Institute of Research, Silver Spring, MD, USA

Accepted in revised form 11 December 2002; received 28 June 2002

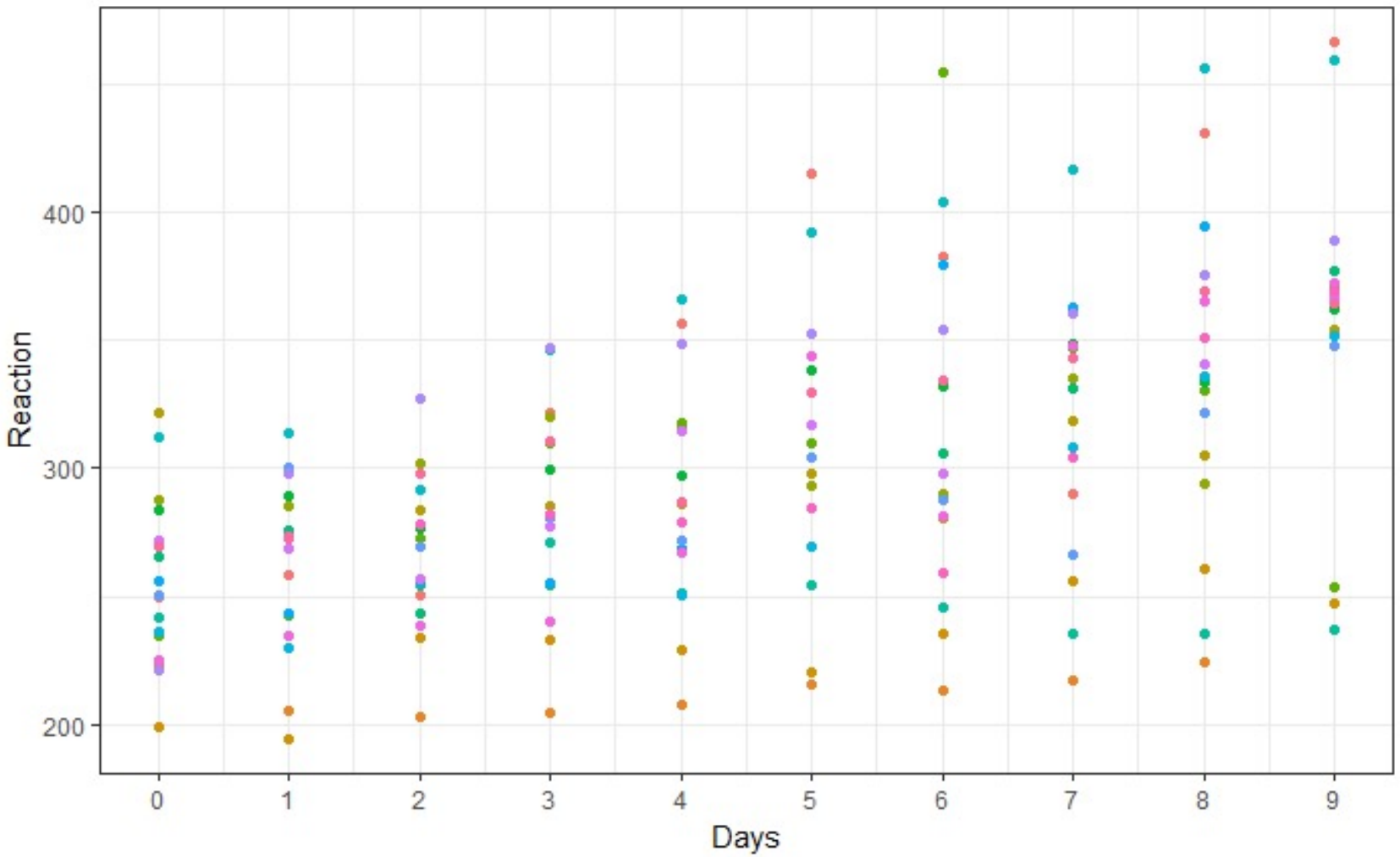

308
309
310
330
331
332
333
334
335
337
349
350
351
352
369
370
371
372
Reaction
Days
200
300
400
0
2
4
6
8

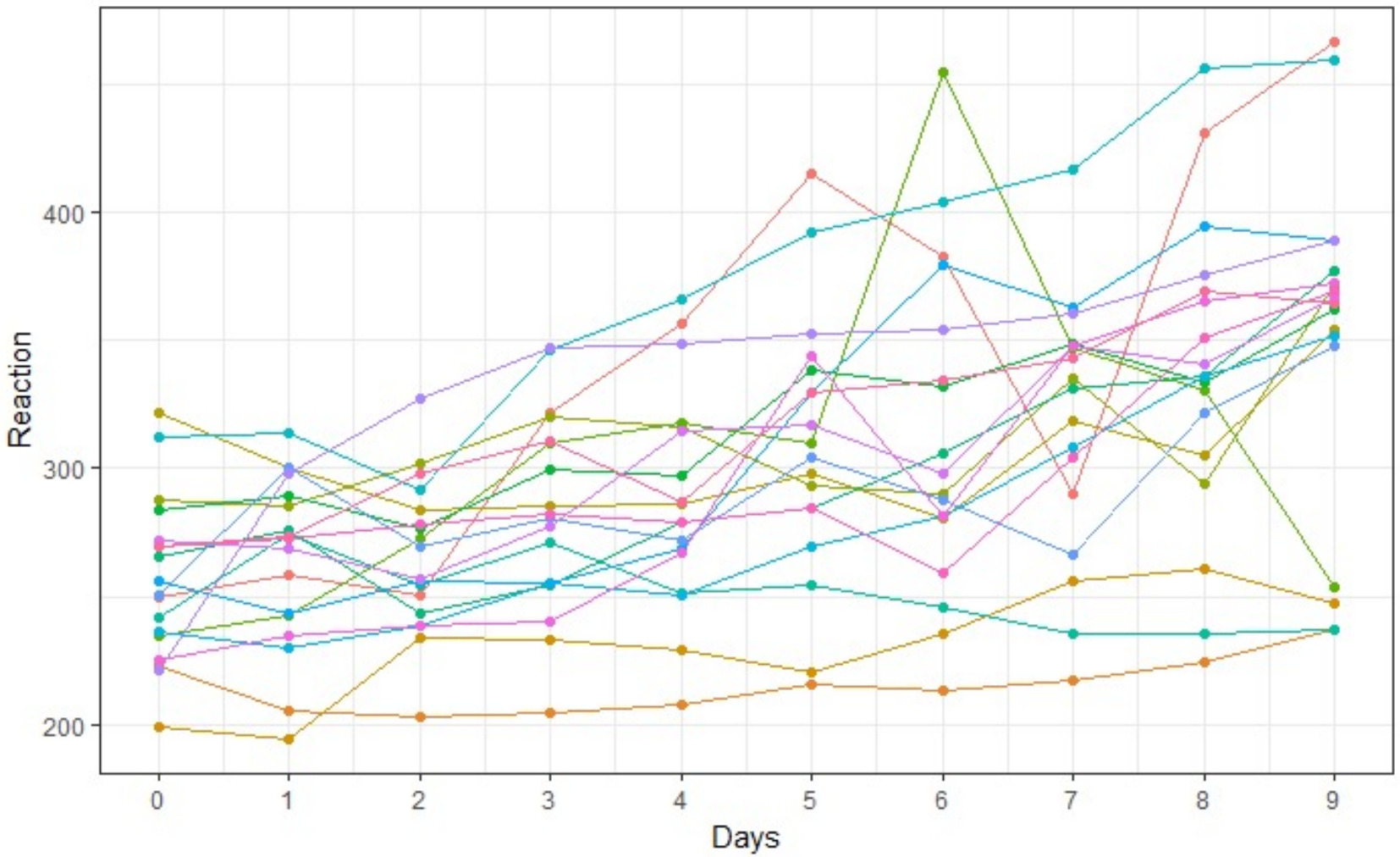

Reaction
400
300
200
0
1
2
3
4
5
6
7
8
9
Days

You could just use linear regression…



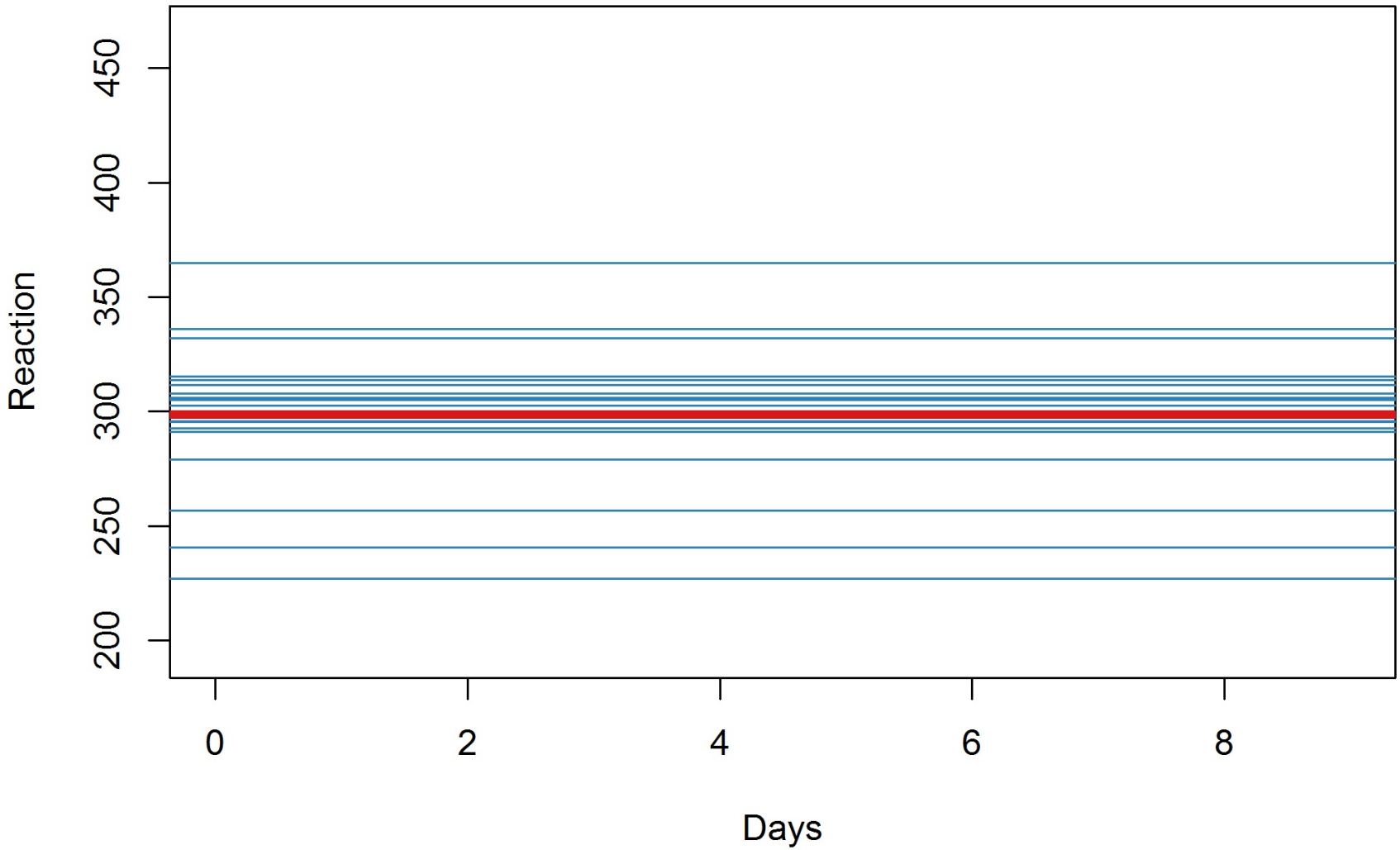
But that doesn’t “know” about the structure of the data

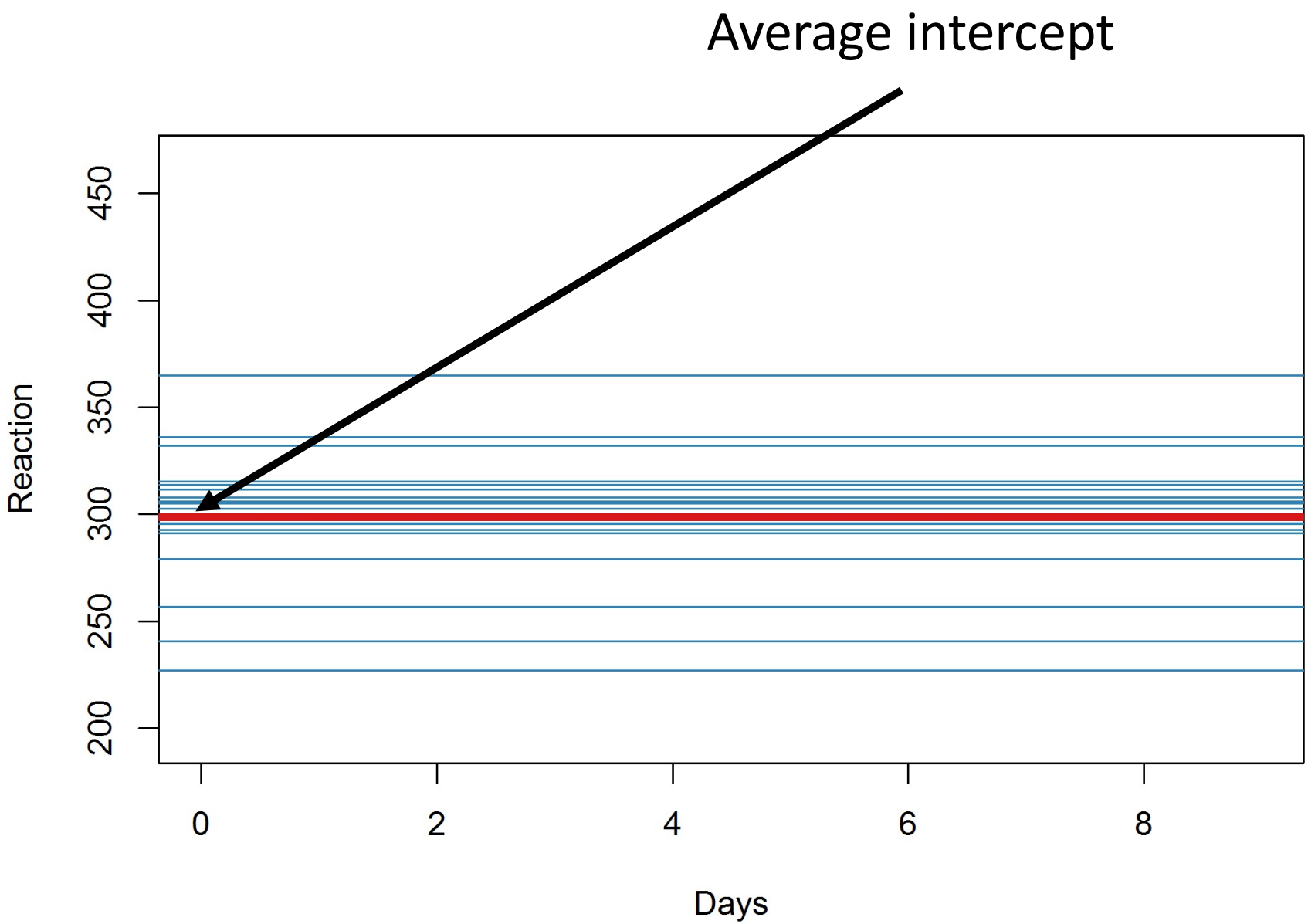


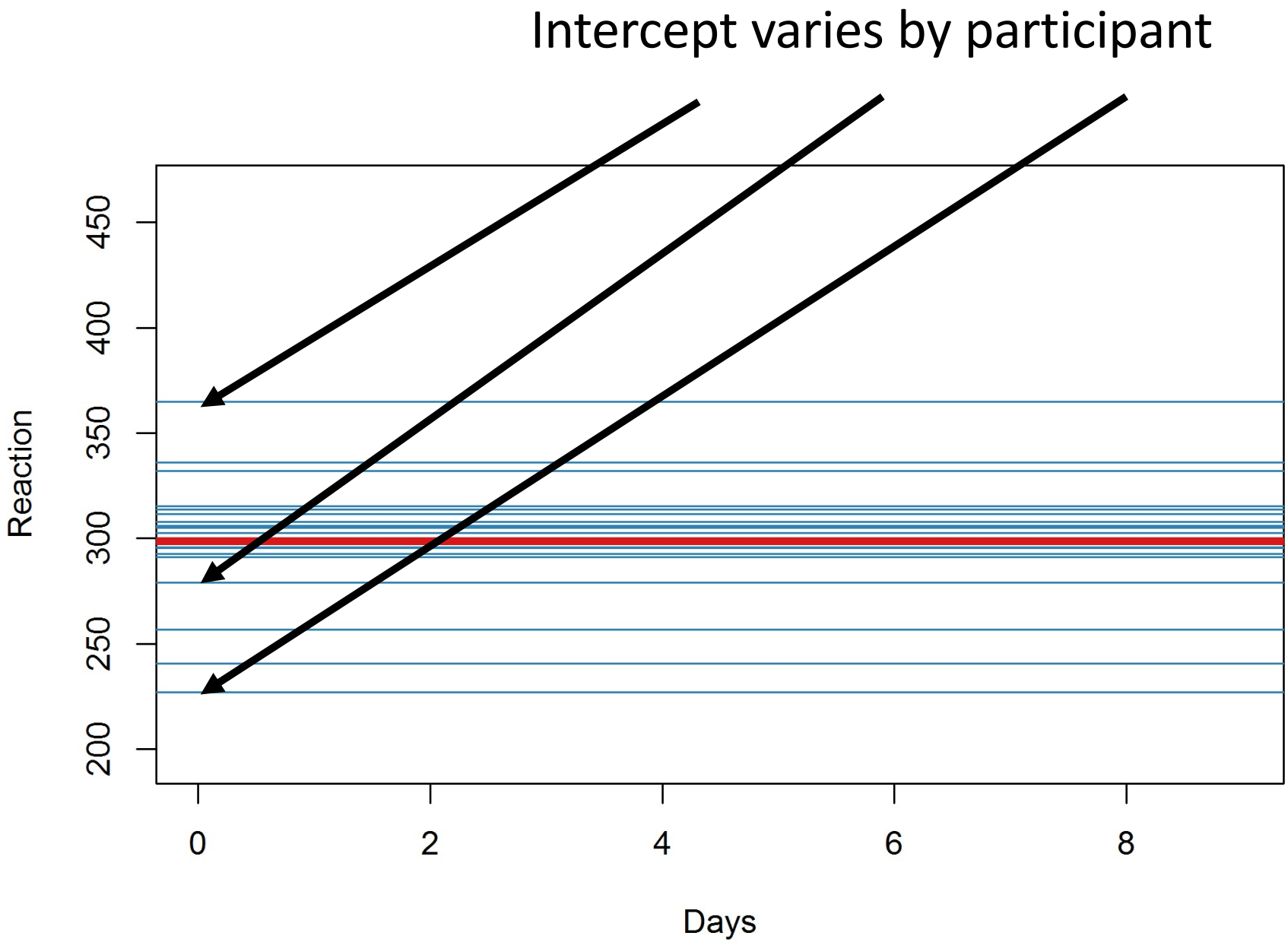

Reaction
400
300
200
0
1
2
3
4
5
6
7
8
9
Days

Why not just fit a separate regression model for each participant?

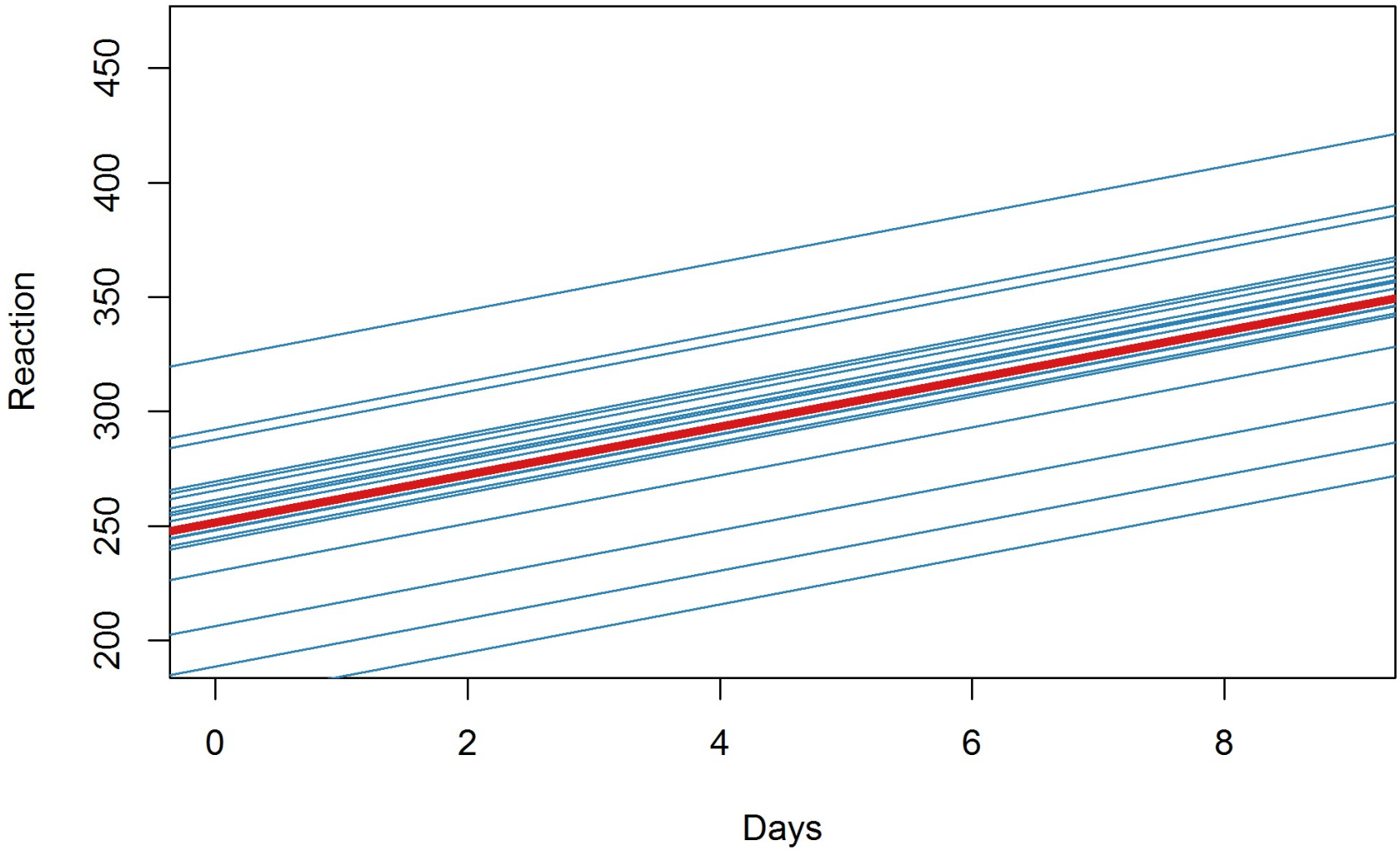
Our favourite baseline model



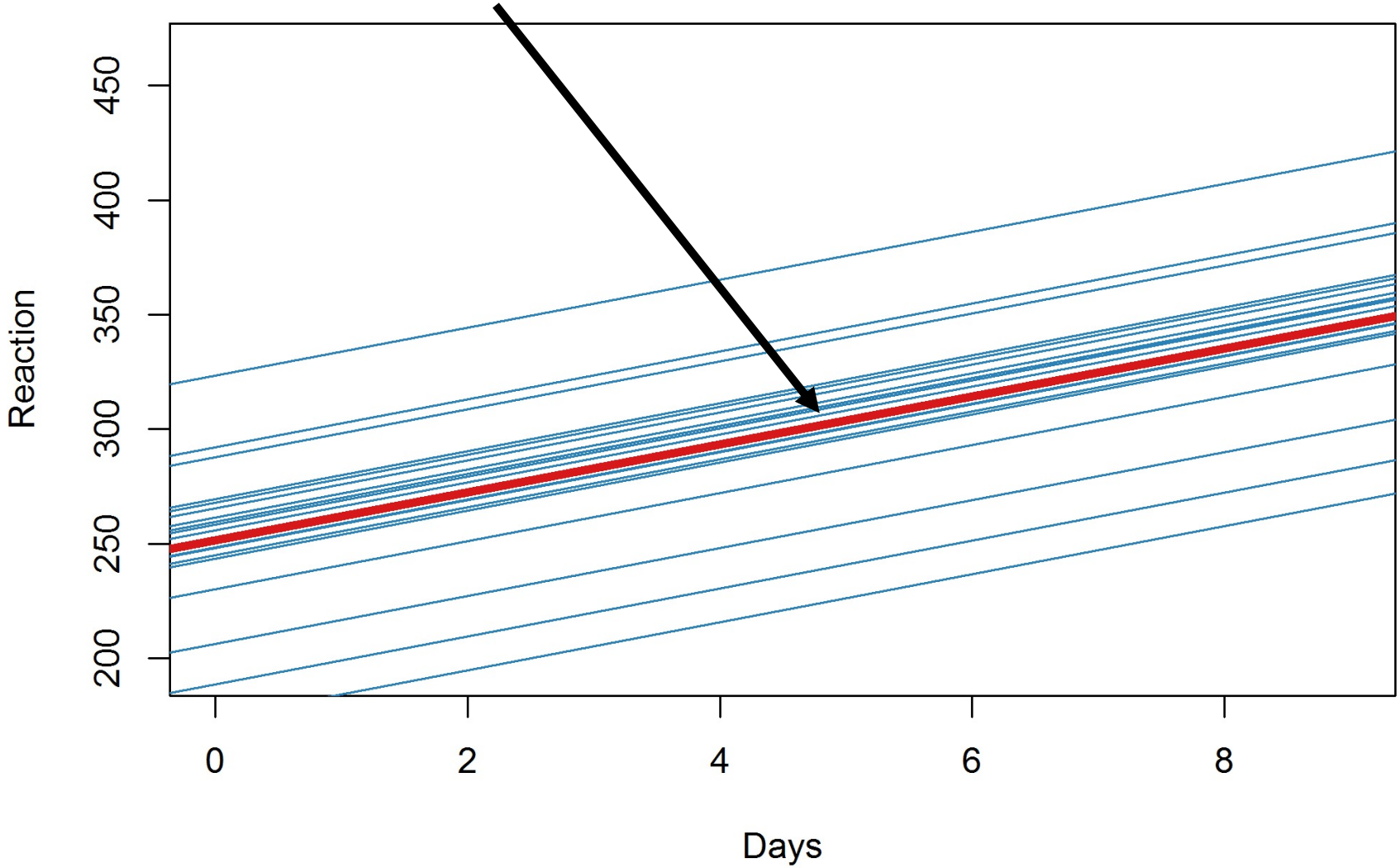

Average intercept
Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

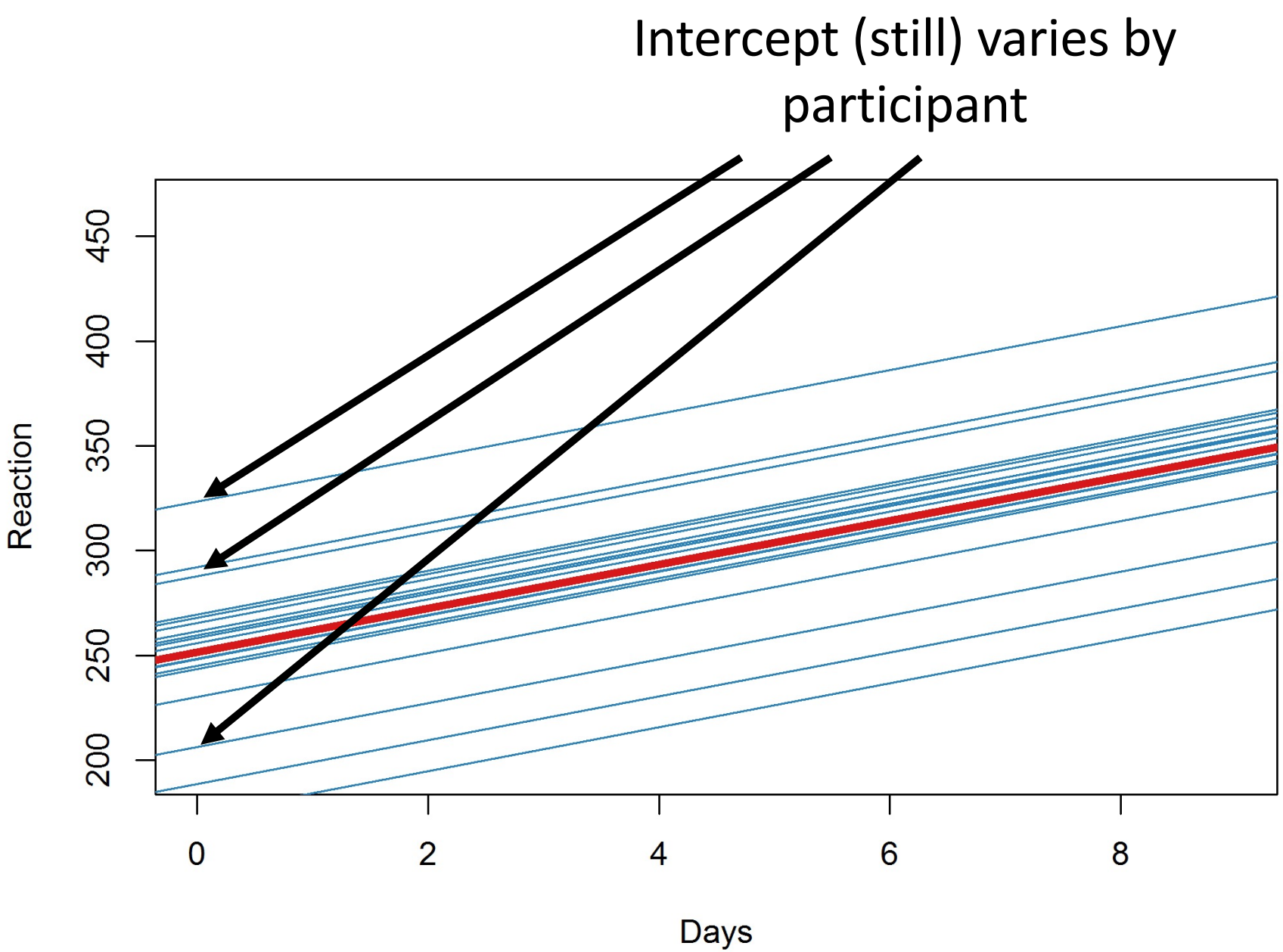

Intercept varies by participant
Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

Less silly model

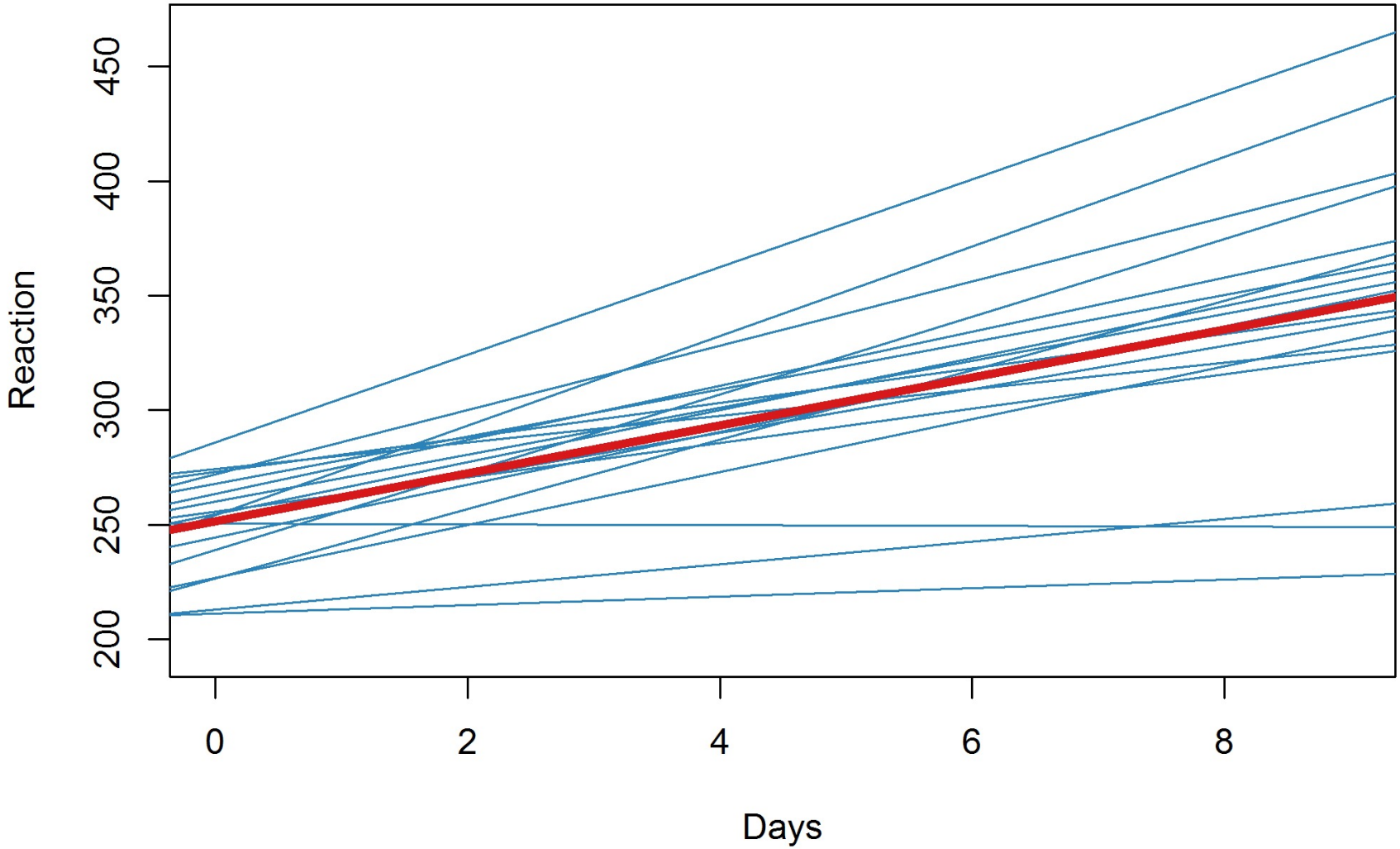


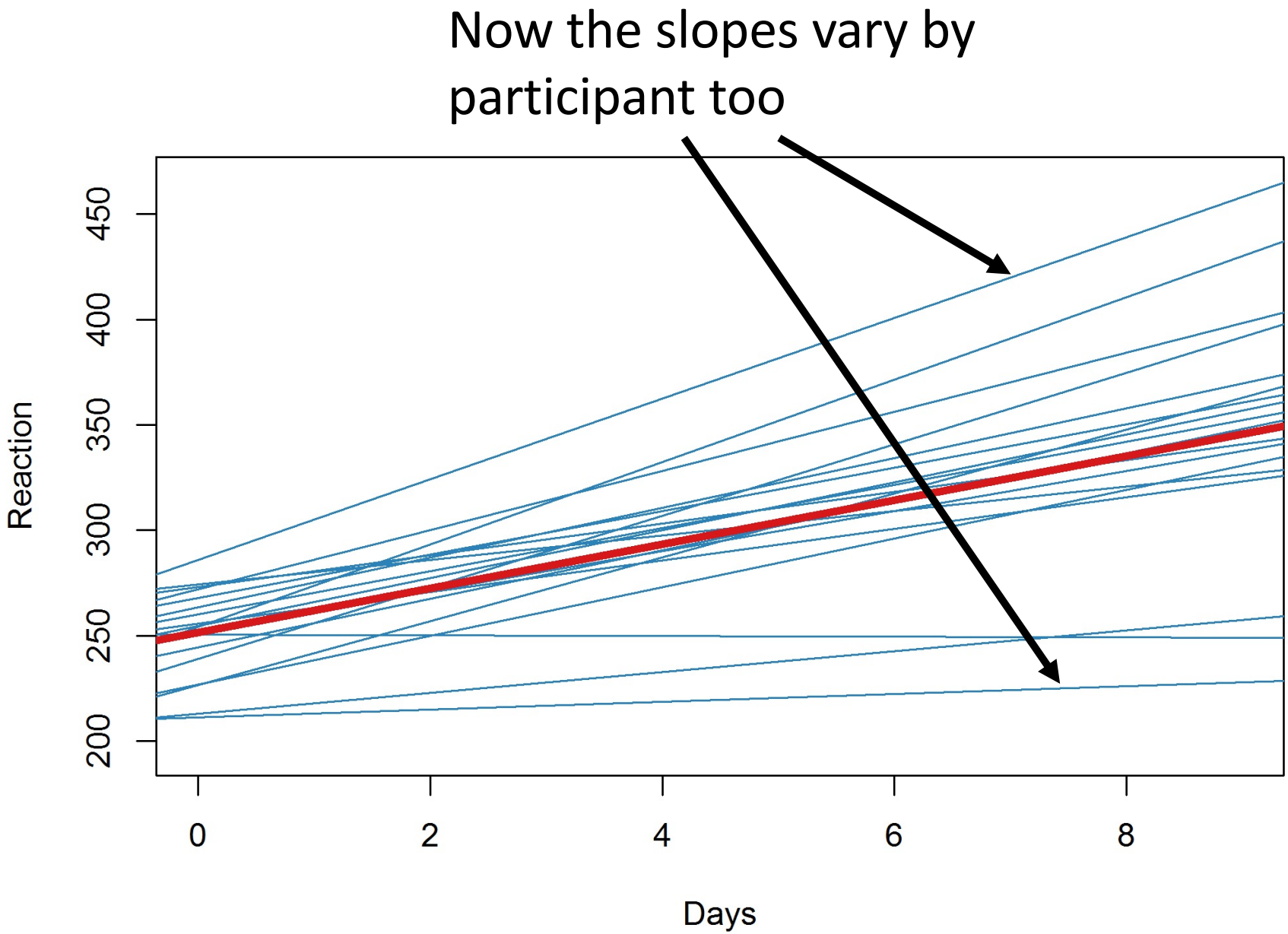
Average slope (slower reactions over time)

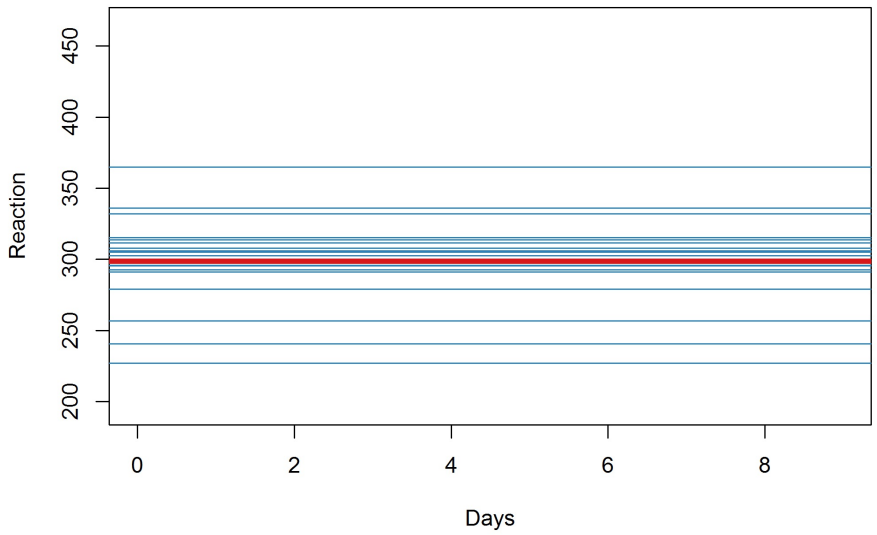


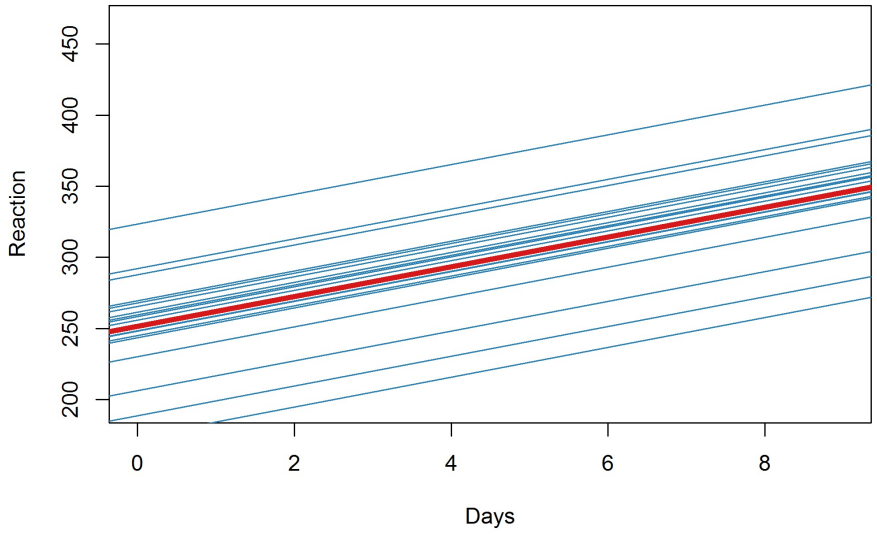

Intercept (still) varies by participant
Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

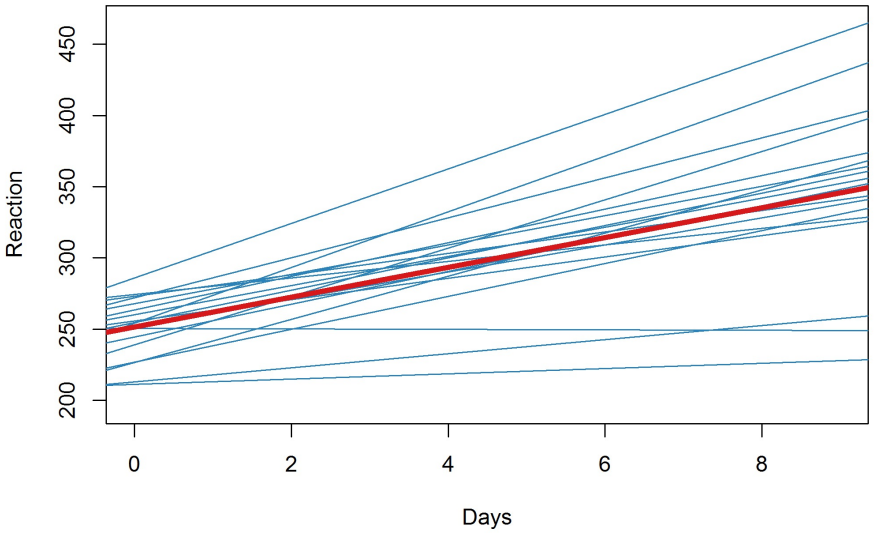
Better model




Now the slopes vary by participant too
Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

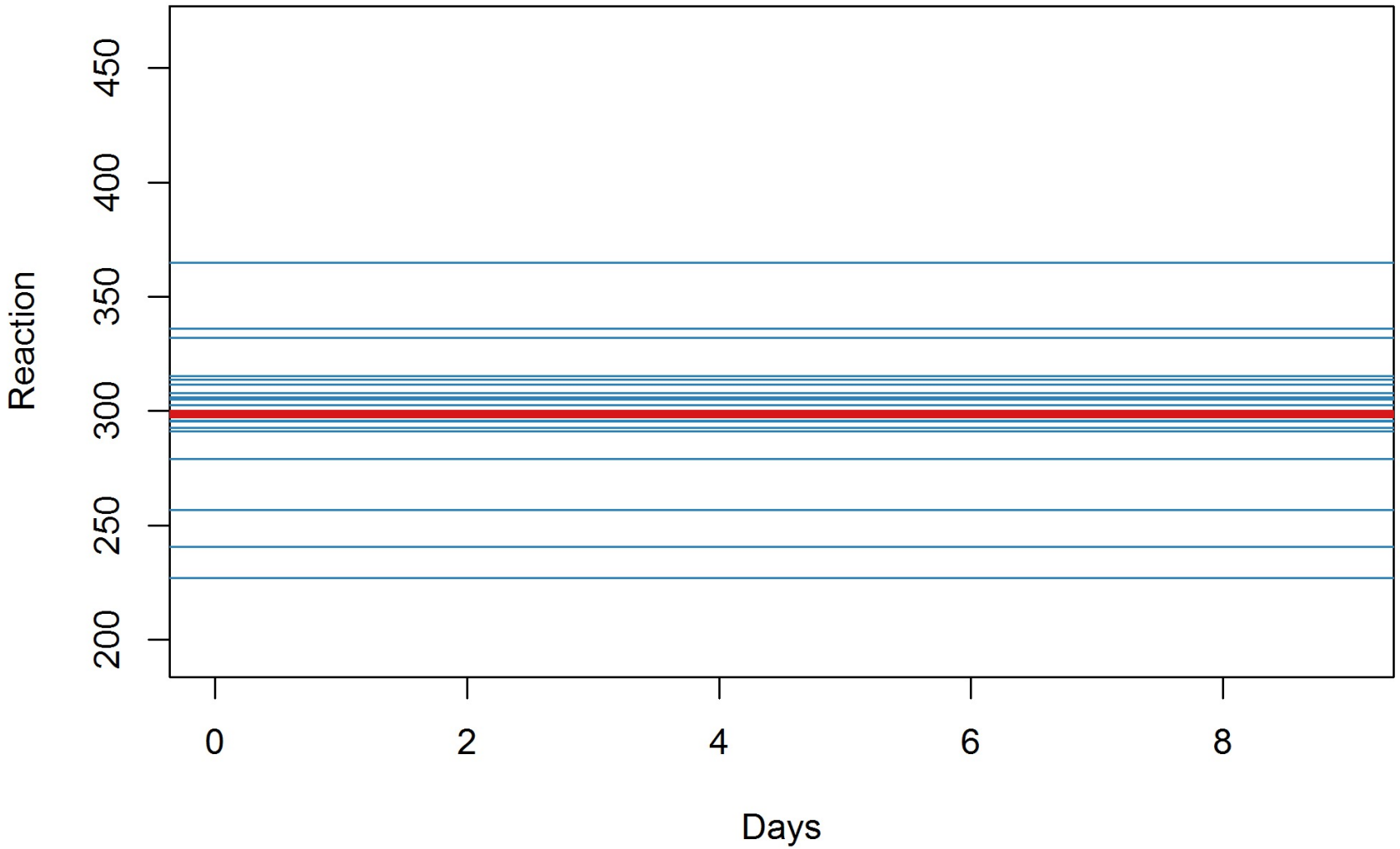

Reaction
200
250
300
350
400
450
0
2
4
6
8
Days


Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

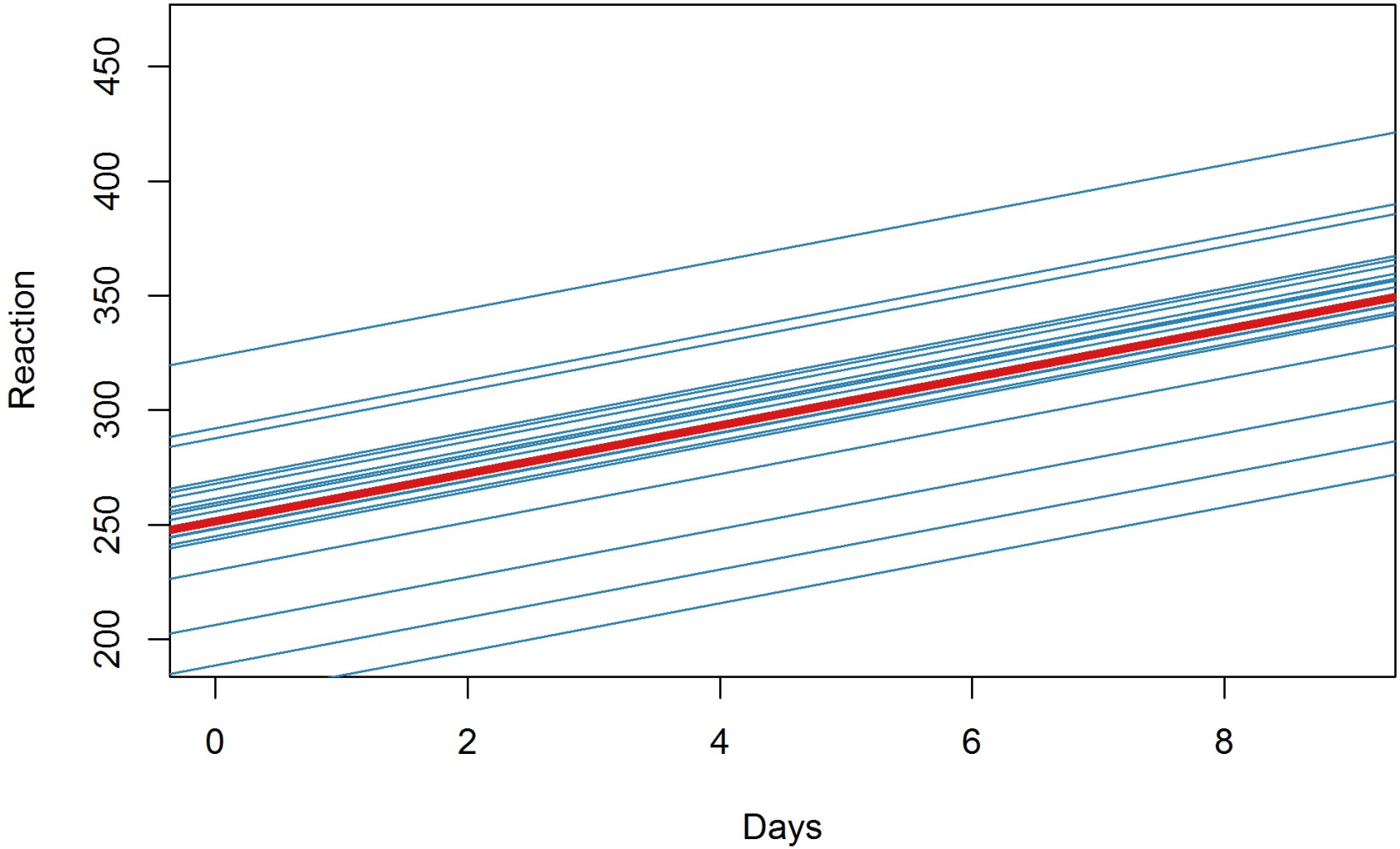

Reaction
200
250
300
350
400
450
0
2
4
6
8
Days

Next step: R code!

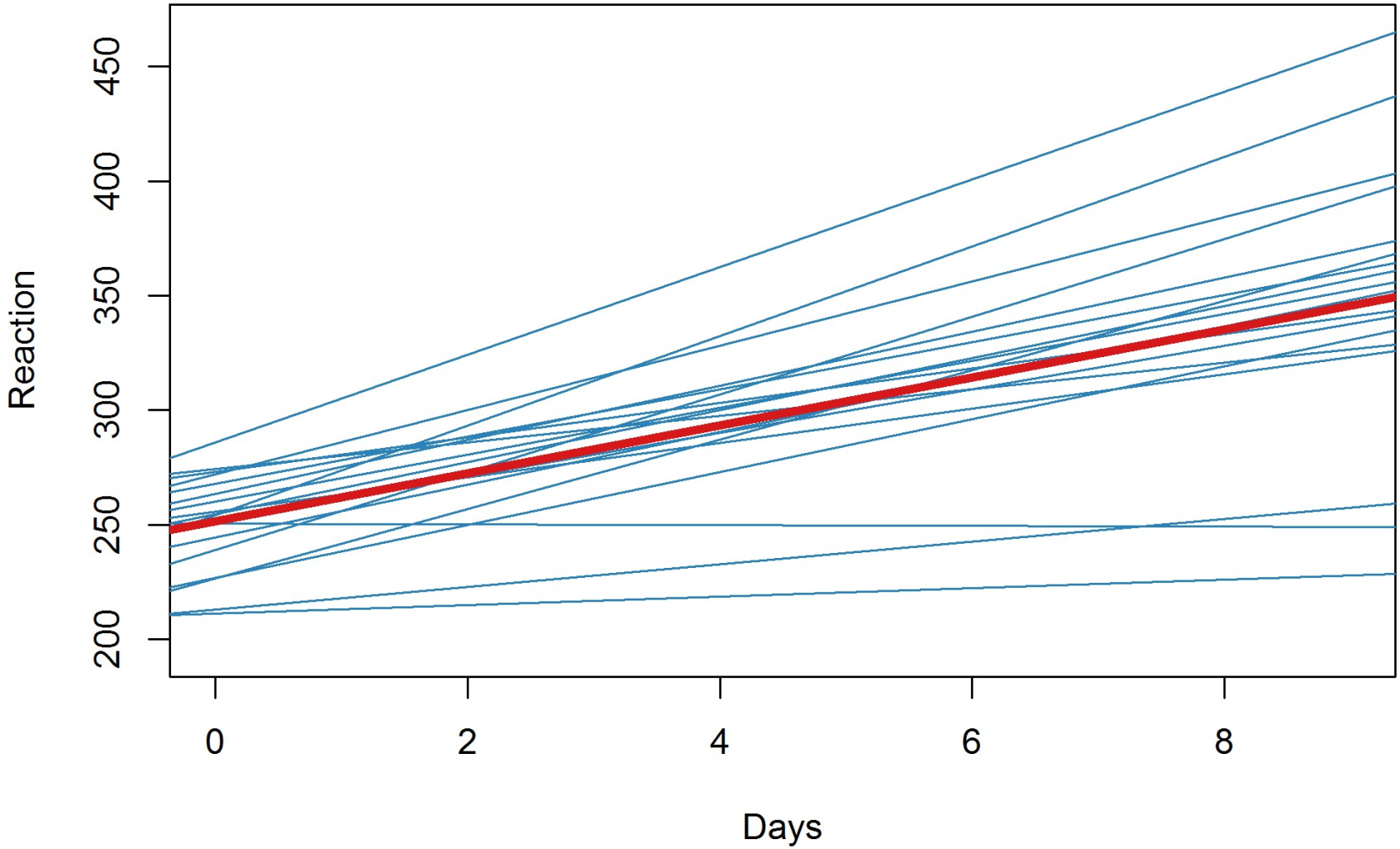
Reaction ~ 1 + (1|Subject)



Reaction ~ 1 + Days + (1|Subject)



Reaction ~ 1 + Days + (1+Days|Subject)



What information do we get?

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

Gelman (2005)

‘We prefer to sidestep the overloaded terms “fixed” and “random” with a cleaner distinction [...]. We define effects (or coefficients) in a multilevel model as constant [fixed] if they are identical for all groups in a population and varying [random] if they are allowed to differ from group to group.’

Gelman A. (2005). Analysis of variance—
why it is more important than ever. *Annals of Statistics*, *33*, 1–53

How to think about random effects

They are like residuals

(in fact residuals are random effects)

But they can be at higher levels in the dataset, e.g.,

slopes

intercept

by school, class, participant, …

Random effects: tells you about the variation at different levels

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
```

Number of obs: 180, **groups**: *Subject*, 18

Fixed effects – interpret as for regression: mean change

```
             Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

Interactive entertainment

1. Read in the sleepstudy.csv dataset
2. Make the “Subject” variable a factor
3. Try the following

```
library(lattice)
xyplot(Reaction ~ Days | Subject, data = dat)
```

More entertainment

4. Use the `lmer` command to fit the models depicted in previous pictures:
   a) Average intercept and allow the intercept to vary by Subject
   b) Average intercept and average slope for Days; again allow the intercept to vary by Subject
   c) Same as *b*, but additionally allow the slope for Days to vary by subject

   What changes and what stays the same across the model summaries…?

[R] lmer, p-values and all that

Douglas Bates bates at stat.wisc.edu
Fri May 19 22:40:27 CEST 2006

- Previous message: [R] Fast update of a lot of records in a database?
- Next message: [R] lmer, p-values and all that
- **Messages sorted by:** [date] [thread] [subject] [author]

Users are often surprised and alarmed that the summary of a linear mixed model fit by lmer provides estimates of the fixed-effects parameters, standard errors for these parameters and a t-ratio but no p-values. Similarly the output from anova applied to a single lmer model provides the sequential sums of squares for the terms in the fixed-effects specification and the corresponding numerator degrees of freedom but no denominator degrees of freedom and, again, no p-values.

Because they feel that the denominator degrees of freedom and the corresponding p-values can easily be calculated they conclude that failure to do this is a sign of inattention or, worse, incompetence on the part of the person who wrote lmer (i.e. me).

Solution 0

The *anova* command still works and gives you a **log-likelihood ratio test** with p-values

```
refitting model(s) with ML (instead of REML)
Data: dat
Models:
lmm0: Reaction ~ 1 + (1 | Subject)
lmm1: Reaction ~ Days + (1 | Subject)
     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
lmm0  3 1916.5 1926.1 -955.27   1910.5
lmm1  4 1802.1 1814.8 -897.04   1794.1 116.46      1  < 2.2e-16 ***
```

Write this as: $\chi^2(1) = 116.4$, $p < .001$

Solution 1

Use **confidence intervals** instead.
These do work for lmer!

```
confint(model name, oldNames = FALSE)
```

Try with and without the `oldNames` option to see what that does

Remember this command also works for *lm* and *glm* and *polr*

Solution 2

Use Alexandra Kuznetsova et al.'s

lmerTest package

This fixes the standard *summary* command so it includes an approximate p-value. So just do:

```
library(lmerTest)
mod <- lmer(...)
summary(mod)
```

You can allow the intercept and slopes to vary at multiple levels…

```
lmer(outcome ~ 1 + (1|ID) + (1|Classroom) …)
```

Intercept varies by ID and by classroom

```
lmer(outcome ~ 1 + Days +
                (1 + Days | ID) +
                (1 + Days | Classroom) …)
```

Intercept and slope for Days varies by ID and classroom

Diagnostics

Linear regression assumptions
(Gelman & Hill, 2007, pp.45-46)

1. **Validity.** The data map to the research question
2. **Additivity** and **linearity**

   $y = B_0 + B_1x_1 + B_2x_2 + \ldots$

   (Transforming the *x*s and *y* might help, if not)
3. **Independence** of errors
4. **Equal variance** of errors ("Homoscedasticity")
5. **Normal distribution** of errors

Linear multilevel regression assumptions

1. **Validity.** The data map to the research question
2. **Additivity** and **linearity**

   $y = B_0 + B_1x_1 + B_2x_2 + \ldots$

   Transforming the *x*s and *y* might help, if not
3. **Independence** of errors

   We have a new trick for dealing with dependence
4. **Equal variance** of errors ("Homoscedasticity")
5. **Normal distribution** of errors

Now also have higher level "errors":

random slopes & intercepts

What's the same?

VIFs – as before (use **vif** in the **car** package)

Get the residuals and predicted values as before:

resid (mod)
predict(mod)

} **Where *mod* is your model's name**

These commands give numbers, which can be scatter **plot**ted (to check for homogeneity of variance) and **hist**ogrammed (to check for normal distributed residuals)

The random effects (here just for Subject; you could have more, e.g., for schools or teachers)

```
> ranef(myMod)

$Subject
    (Intercept)         Days
308   2.2585654   9.1989719
309 -40.3985769  -8.6197032
310 -38.9602458  -5.4488799
330  23.6904985  -4.8143313
331  22.2602027  -3.0698946
332   9.0395259  -0.2721707
…
```

Select the subject ones

```
> ranef(myMod)$Subject

    (Intercept)        Days
308   2.2585654   9.1989719
309 -40.3985769  -8.6197032
310 -38.9602458  -5.4488799
330  23.6904985  -4.8143313
331  22.2602027  -3.0698946
332   9.0395259  -0.2721707
…
```

The parentheses in “(Intercept)” confuse R, so you need to use backquotes:

```
> ranef(myMod)$Subject$`(Intercept)`
```

2.2585654 -40.3985769 -38.9602458 23.6904985 22.2602027 9.0395259 16.8404311 -7.2325792 -0.3336958 34.8903508 -25.2101104

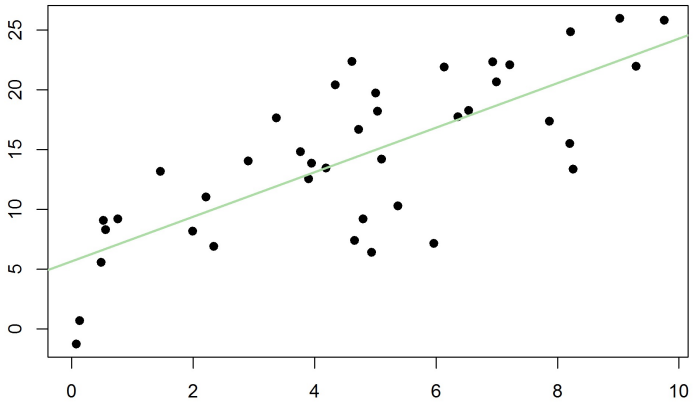
…

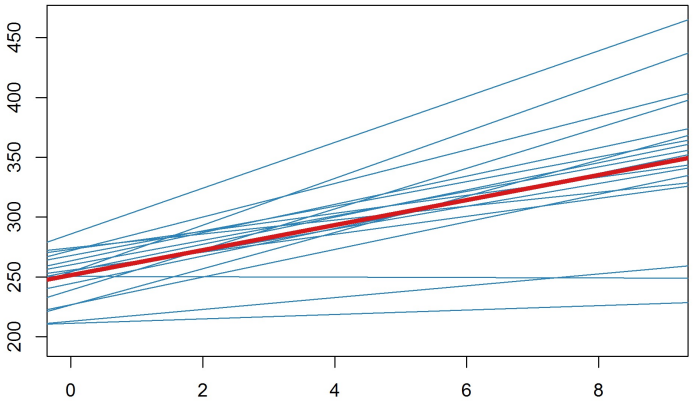
Alternatively, ask for the first column

```
> ranef(myMod)$Subject[,1]
```

2.2585654 -40.3985769 -38.9602458 23.6904985
22.2602027 9.0395259 16.8404311 -7.2325792 -
0.3336958 34.8903508 -25.2101104
…

Summary





Multilevel models are

extensions of regression

Allow multilevel data, like students in classrooms (in schools (in counties (in countries)))…

Describing

variation at different levels

+ average relationships

The end