simon-5502-12-slides

Topics to be covered

- What you will learn
    - Mathematical formulation of random intercepts model
    - Description of HIV-intervention data
    - Random intercepts model using hiv-intervention data
    - Mathematical formulation of random slopes model
    - Assumptions and complications
    - Sample size justification

Longitudinal data

- Measurements taken at different times
  - Emphasis in changes over time

Speaker notes

In the previous module, I talked about hierarchical models and mentioned a particular case, longitudinal data, that I want to talk more about in this presentation.

Longitudinal data is similar to repeated measures data. With both, you measure the same subject repeatedly. With longitudinal data, often the emphasis is in changes that occur over time. Repeated measurements, in contrast, emphasize different treatments with the hope that the time gaps between the measurements are small enough that you don’t see changes over time.

The differences between longitudinal data, repeated measures data, or hierarchical data are subtle. Perhaps these are distinctions without a difference. I decided to separate out longitudinal data for a different module perhaps more out of the desire to split a complex topic into smaller bite-sized pieces.

Random intercepts model, 1

- Simplest pattern for longitudinal data
- $Y_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, k$
  - n subjects, k time points
- t_j, time of jth measurement
  - First time is often zero

Speaker notes

The simplest longitudinal model has n subjects and k time points. The first time point is often set to zero. The times are often evenly spaced, but they don’t have to be.

Random intercepts model, 2

- $Y_{ij} = \beta_0 + u_{0i} + \beta_1 t_j + \epsilon_{ij}$
    - β_0 and β_1 are unknown constants
    - u_{0i} and ϵ_{ij} are normally distributed
        - $SD(u_{0i}) = \sigma_{intercept}$
        - $SD(\epsilon_{ij}) = \sigma_{error}$

Speaker notes

There are two sources of random variation in the random intercepts model, u_{0i} and ϵ_{ij}.

Random intercepts model, 3

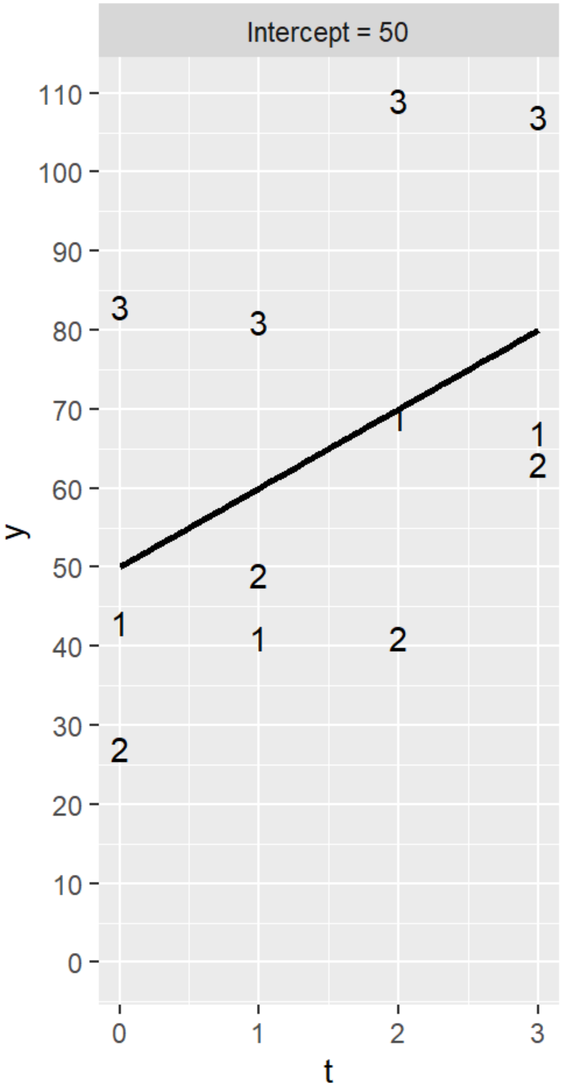
- $SD(Y_{ij}) = \sqrt{\sigma^2_{intercept}\ +\ \sigma^2_{error}}$
- $Corr(Y_{ij}, Y_{im}) = \frac{\sigma^2_{intercept}}{\sigma^2_{intercept} + \sigma^2_{error}}$

Speaker notes

The standard deviation for any individual observation combines the standard deviation for the random intercepts and the standard deviation for the error terms. They combine in a Pythagorean way.

The correlation of two measurements on the same patient is comparable to a measure we defined in the last module, the intraclass correlation.

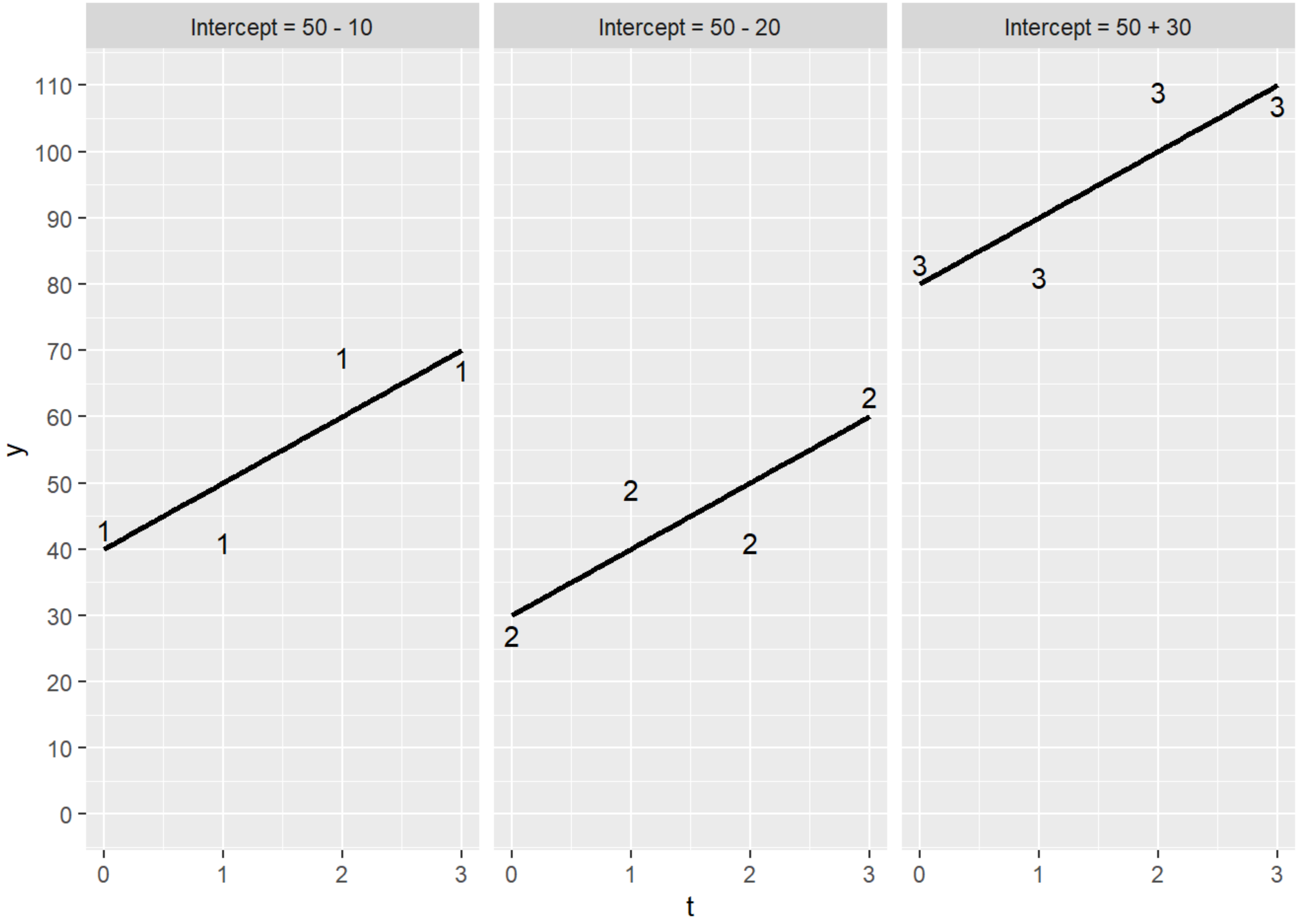
Random intercepts illustrated, 1



Speaker notes

This graph shows a single line. It does reasonably well, but there is a fair amount of variation. There is something worth examining more closely. The third group has values well above the regression line. The first two groups have values slightly below the regression line.

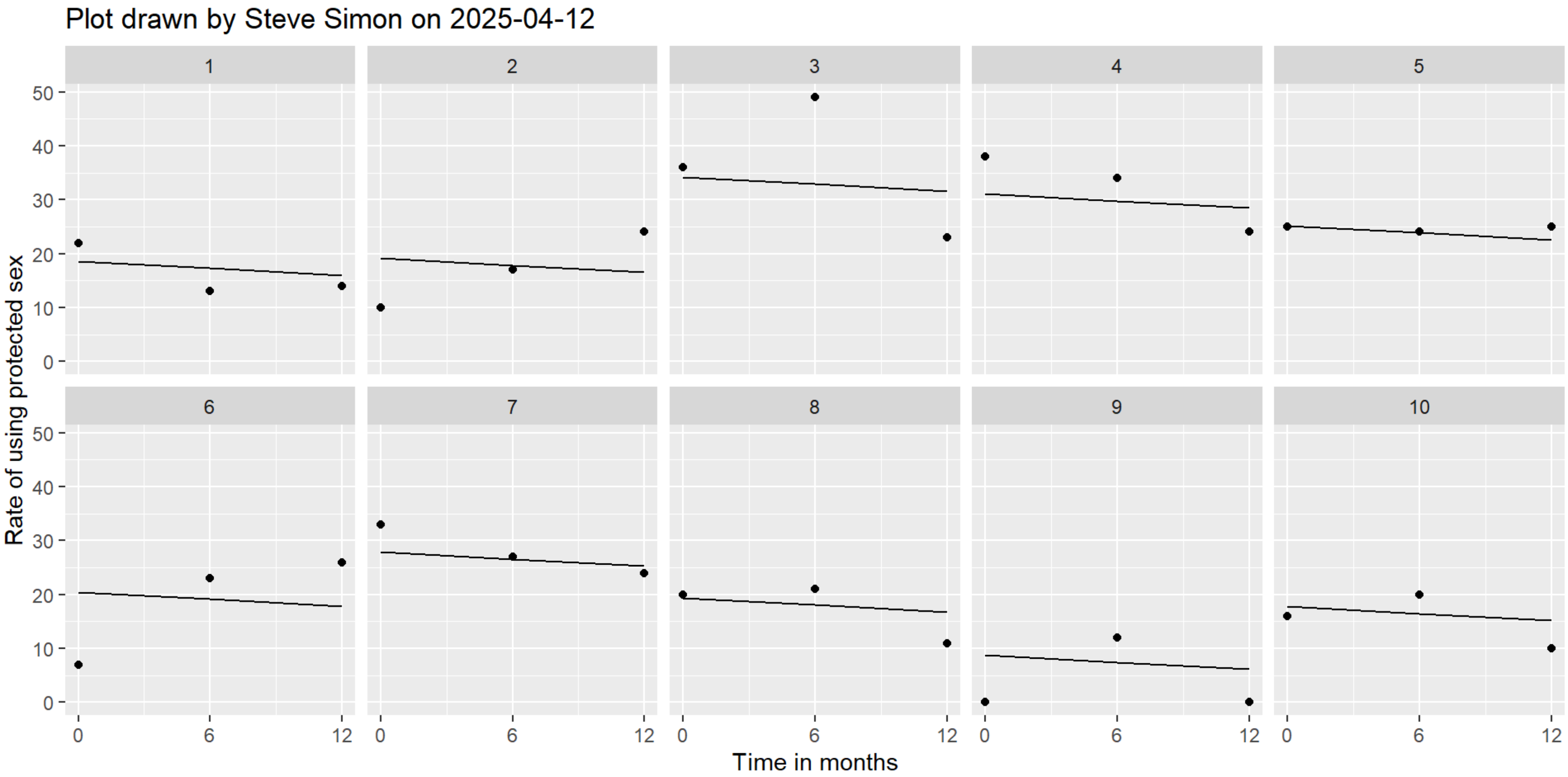
Random intercepts illustrated, 2



Speaker notes

If you fit a separate intercept for each line, you get a lot closer to the data.

Illustration of random intercepts with real data



Break #1

- What you have learned
  - Mathematical formulation of random intercepts model
- What's coming next
  - Description of HIV-intervention data

Description of hiv-intervention data, 1

```
data_dictionary: hiv-intervention.txt

source: OzDASL website

description: |
  This is a longitudinal study of an intervention in 14-18 adolescents
intended to increase the frequency of condom protected sex. Subjects  were
allocated randomly to treatment or control groups. All were evaluated prior to
the intervention, immediately  after the intervention, 6 months and  12 months
after the intervention.The outcome variable is the logarithm-transformed
frequency of condom-protected sex ( log(Y+1) )."
```

Speaker notes

Here is a dataset I will use to illustrate the random intercepts model. It actually might require a more sophisticated model than the random intercepts, but it is always a good idea to start with the simplest model, even if you know it is an oversimplification. Slowly add layers of complexity, and don’t fit the final model too early. You want to wade in from the shallow end of the pool rather than jump right away into the deep end.

Description of hiv-intervention data, 2

```
BST:
  label: treatment group
  values:
    '1': BST intervention
    '0': control
Pre:
  label: Log-frequency of protected sex before the intervention
Post:
  label: Log-frequency of protected sex after the intervention
FU6:
  label: Log-frequency of protected sex reported at the 6 months follow-up
FU12:
  label: Log-frequency of protected sex reported at the 12 months follow-up
```

Here are the variables. The actual times associated with the Pre and Post measurements are unclear. It turns out that it will be best to hold the Pre measurement back for the time being and start the clock at time=0 for the Post measurement. Remember that you are wading in from the shallow end of the pool.

The remaining two variables FU6 and FU12 represent time=6 and time=12, respectively.

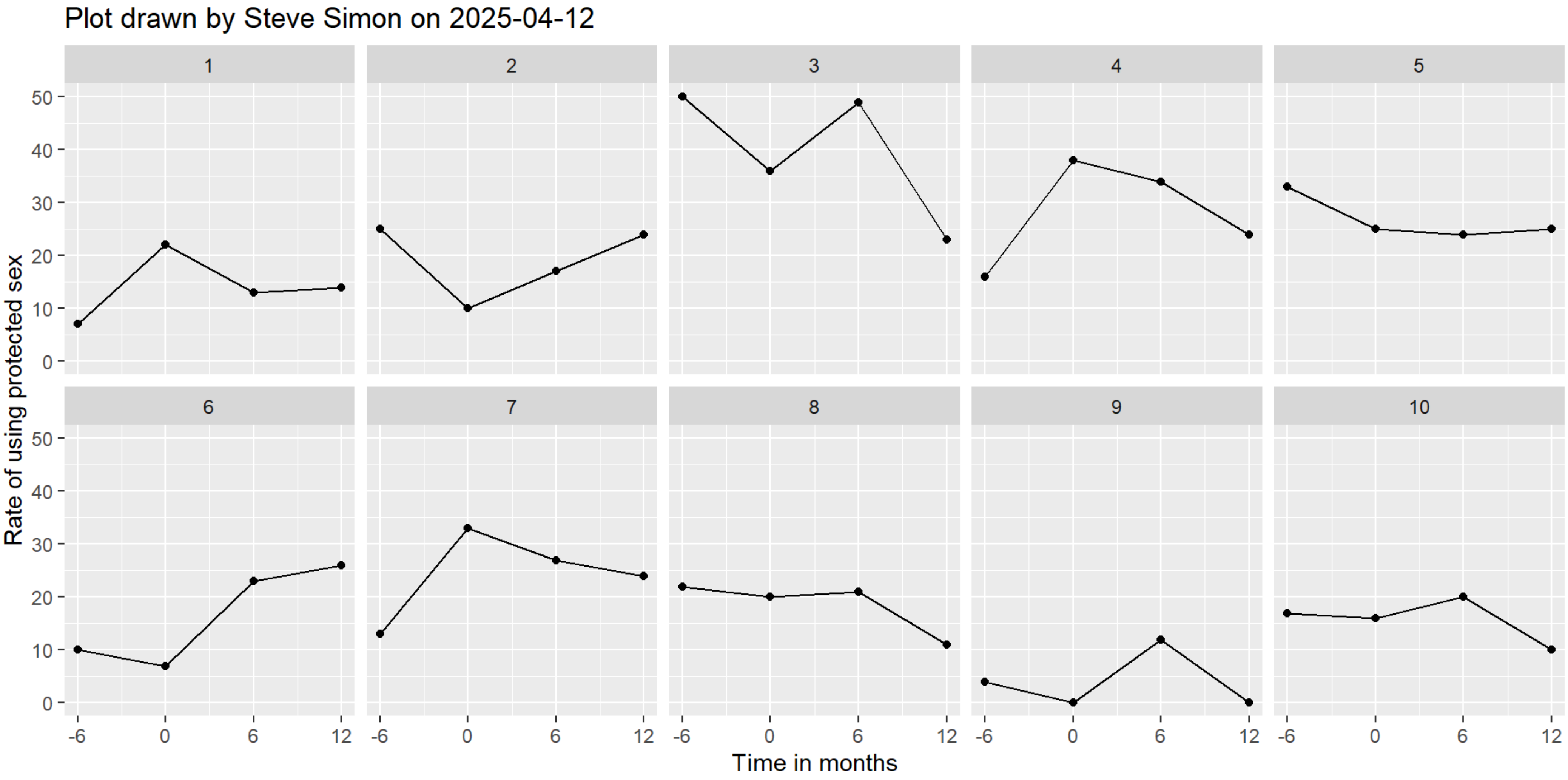
Glimpse of the hiv-intervention data

```
Rows: 20
Columns: 5
$ BST  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
$ Pre  <dbl> 7, 25, 50, 16, 33, 10, 13, 22, 4, 17, 0, 69, 5, 4, 35, 7, 51,
25,…
$ Post <dbl> 22, 10, 36, 38, 25, 7, 33, 20, 0, 16, 0, 56, 0, 24, 8, 0, 53, 0,
…
$ FU6  <dbl> 13, 17, 49, 34, 24, 23, 27, 21, 12, 20, 0, 14, 0, 0, 0, 9, 8, 0,
…
$ FU12 <dbl> 14, 24, 23, 24, 25, 26, 24, 11, 0, 10, 0, 36, 5, 0, 0, 37, 26,
15…
```

Speaker notes

The data uses a wide format with one row per subject and individual columns for the measurements prior to the intervention, after the intervention, 6 months later, and 12 months later.

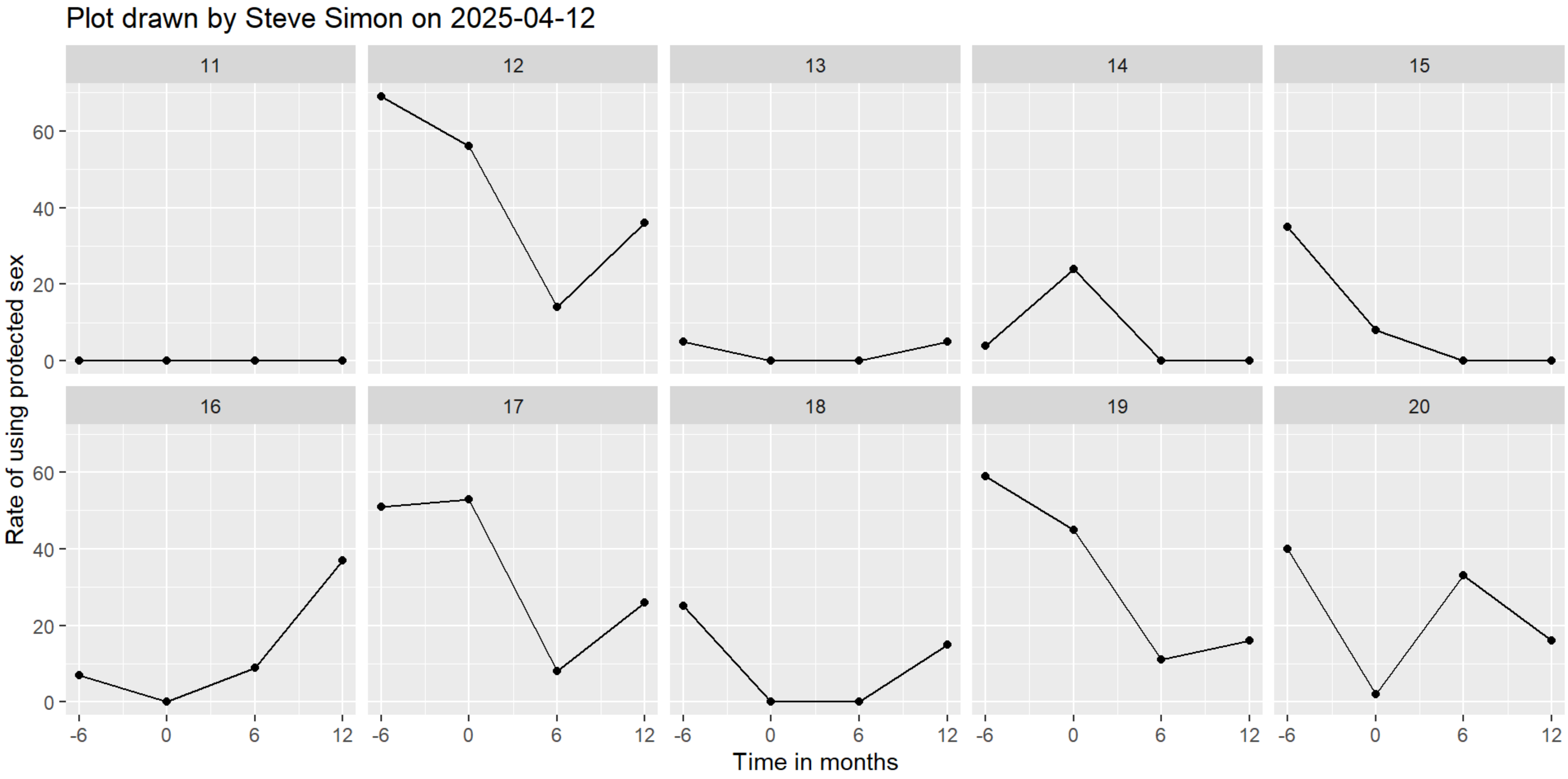
Plot of the data, 1



Speaker notes

This is a series of plots, one for each subject, in the treatment group (BST=1). It uses the facet_wrap function.

Plot of the data, 2



Speaker notes

This is a similar series of plots for the control group (BST=0).

Glimpse of the restructured and simplified data

```
Rows: 30
Columns: 4
$ BST           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,…
$ id            <int> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6,
7,…
$ t             <dbl> 0, 6, 12, 0, 6, 12, 0, 6, 12, 0, 6, 12, 0, 6, 12, 0, 6,
…
$ protected_sex <dbl> 22, 13, 14, 10, 17, 24, 36, 49, 23, 38, 34, 24, 25, 24,
…
```

Break #2

- What you have learned
  - Description of HIV-intervention data
- What’s coming next
  - Random intercepts model using hiv-intervention data

Random intercepts analysis, 1

```
Linear mixed model fit by REML ['lmerMod']
Formula: protected_sex ~ t + (1 | id)
   Data: hiv_3

REML criterion at convergence: 215.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8385 -0.6028  0.0364  0.5183  2.1998
```

Speaker notes

See simon-5505-12-demo for a detailed interpretation of this output.

Random intercepts analysis, 2

```
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 69.57    8.341
 Residual             53.34    7.304
Number of obs: 30, groups:  id, 10
```

Random intercepts analysis, 3

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.2333     3.3768   6.584
t            -0.2167     0.2722  -0.796
```

Random intercepts analysis, 4

```
Correlation of Fixed Effects:
  (Intr)
t -0.484
```

Live demo, fitting a random intercepts model

Break #3

- What you have learned
  - Random intercepts model using hiv-intervention data
- What’s coming next
  - Mathematical formulation of random slopes model

Random slopes model, 1

- Same notation for the time and outcome variables
- $Y_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, k$
  - n subjects, k time points
- t_j, time of jth measurement

Speaker notes

The random slopes model has the same basic notation for the time and outcome variables. The outcome variable has two subscripts on for the individual patient and one for each time measurement.

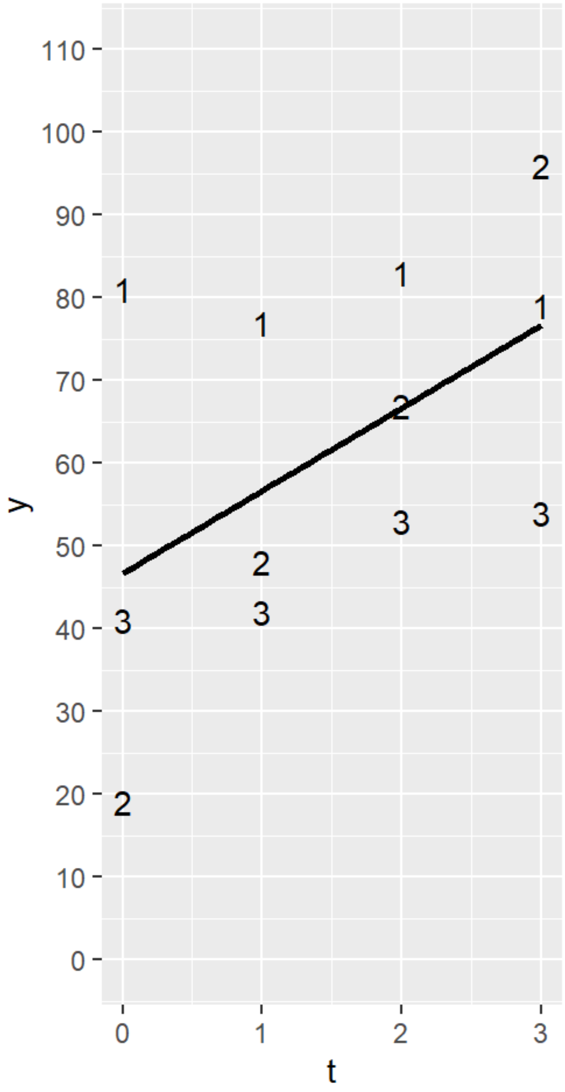
Random slopes model, 2

- $Y_{ij} = \beta_0 + u_{0i} + \beta_1 t_j + u_{1i} t_j + \epsilon_{ij}$
  - β_0 and β_1 are unknown constants
  - u_{0i}, u_{1i}, and ϵ_{ij} are normally distributed
    - $SD(u_{0i}) = \sigma_{intercept}$
    - $SD(u_{1i}) = \sigma_{slope}$
    - $SD(\epsilon_{ij}) = \sigma_{error}$

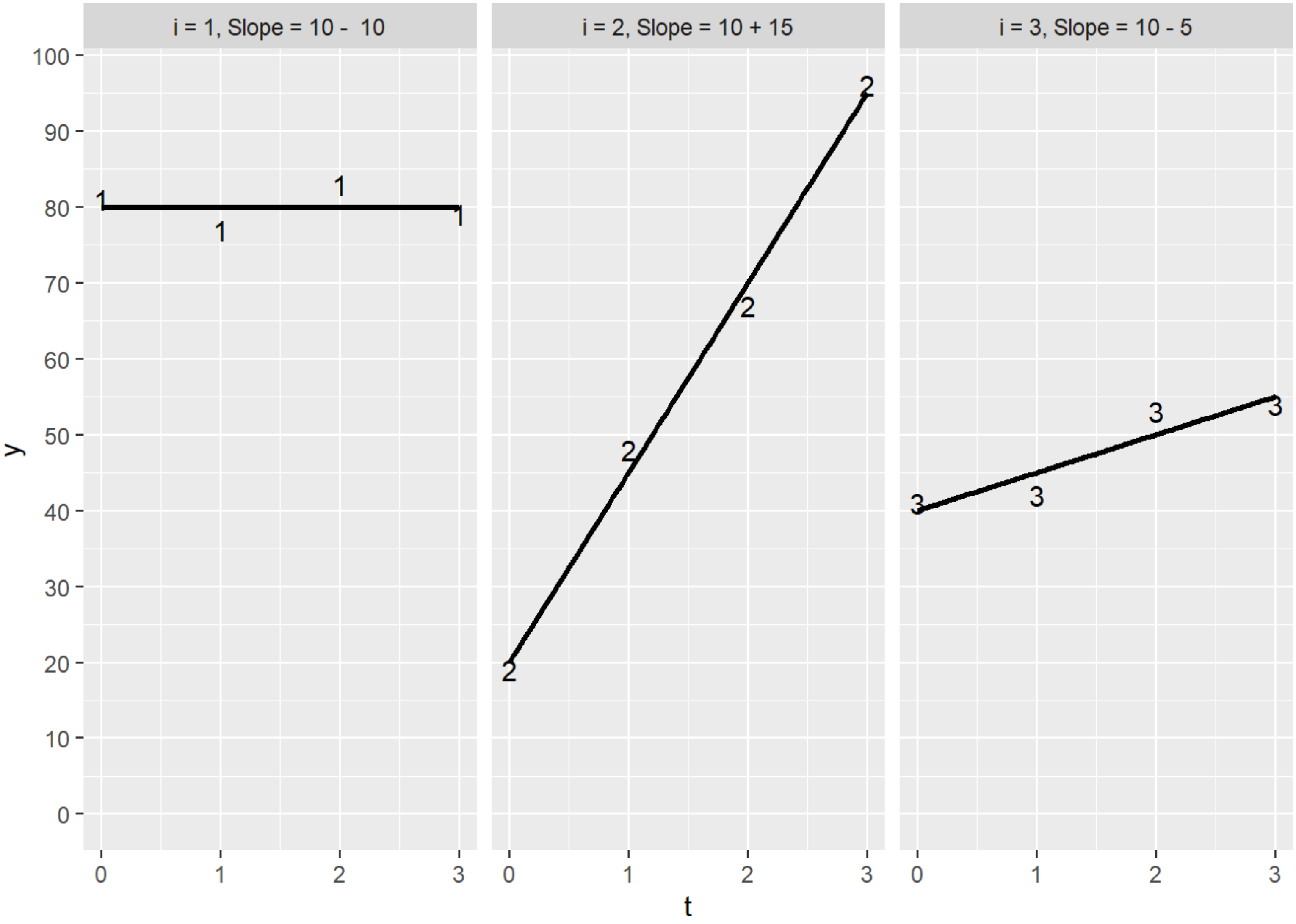
Speaker notes

There are three sources of random variation in the random slopes model, u_{0i}, u_{1i}, and ϵ_{ij}.

Random slopes illustrated, 1



Random slopes illustrated, 2



Break #4

- What you have learned
  - Mathematical formulation of random slopes model
- What's coming next
  - Assumptions and complications

Assumptions

- Independence
  - Only between subjects
- Normality
  - Residuals
  - Random intercepts and/or slopes
- Linearity

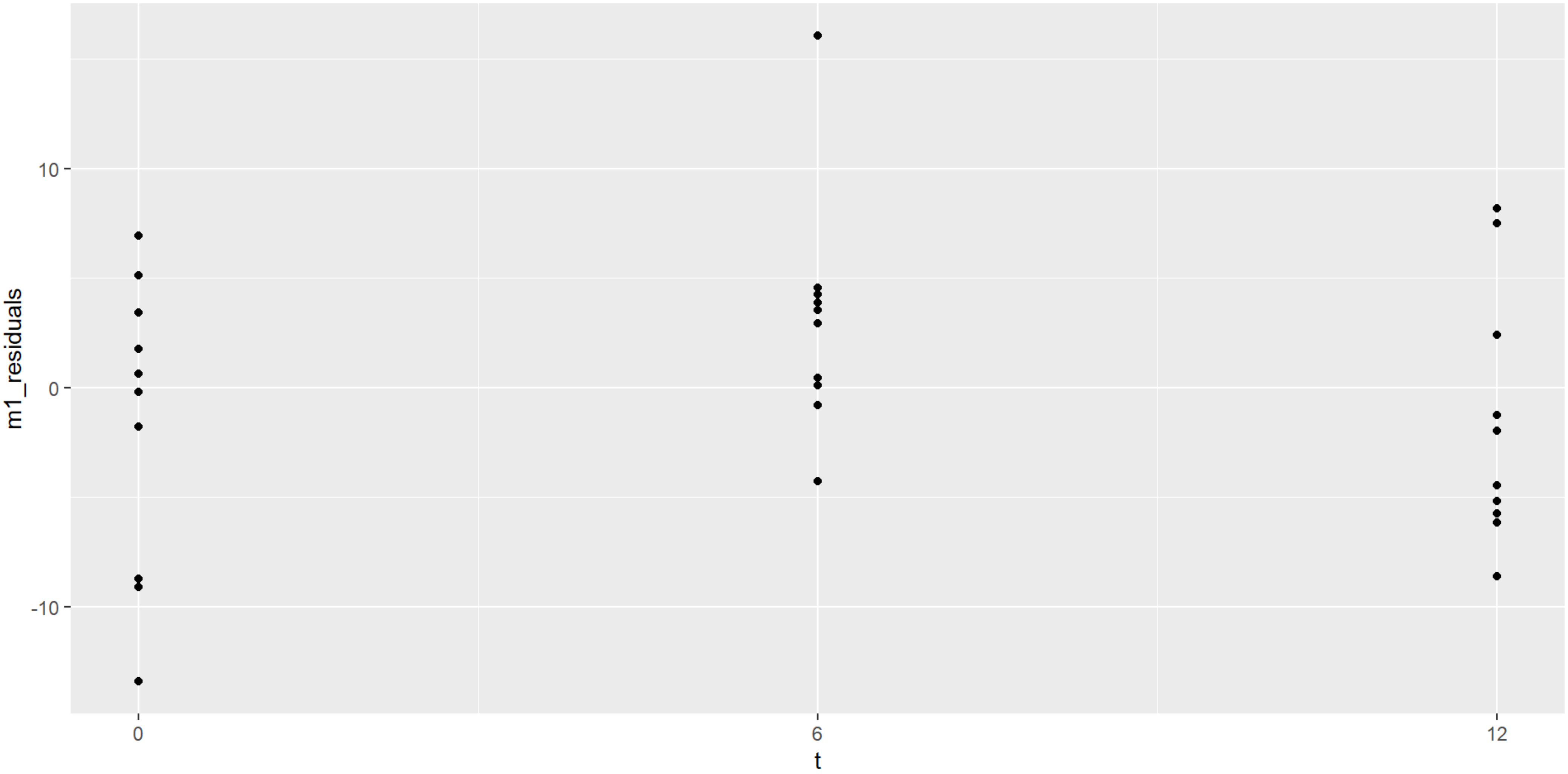
Speaker notes

There are three assumptions: independence, normality, and linearity. The independence assumption is only for between cluster observations. The multiple measurements within a cluster are correlated. In fact, we are glad that the within measurements are correlated. The only time you worry about independence is when observations from one cluster are correlated with observations of a different cluster.

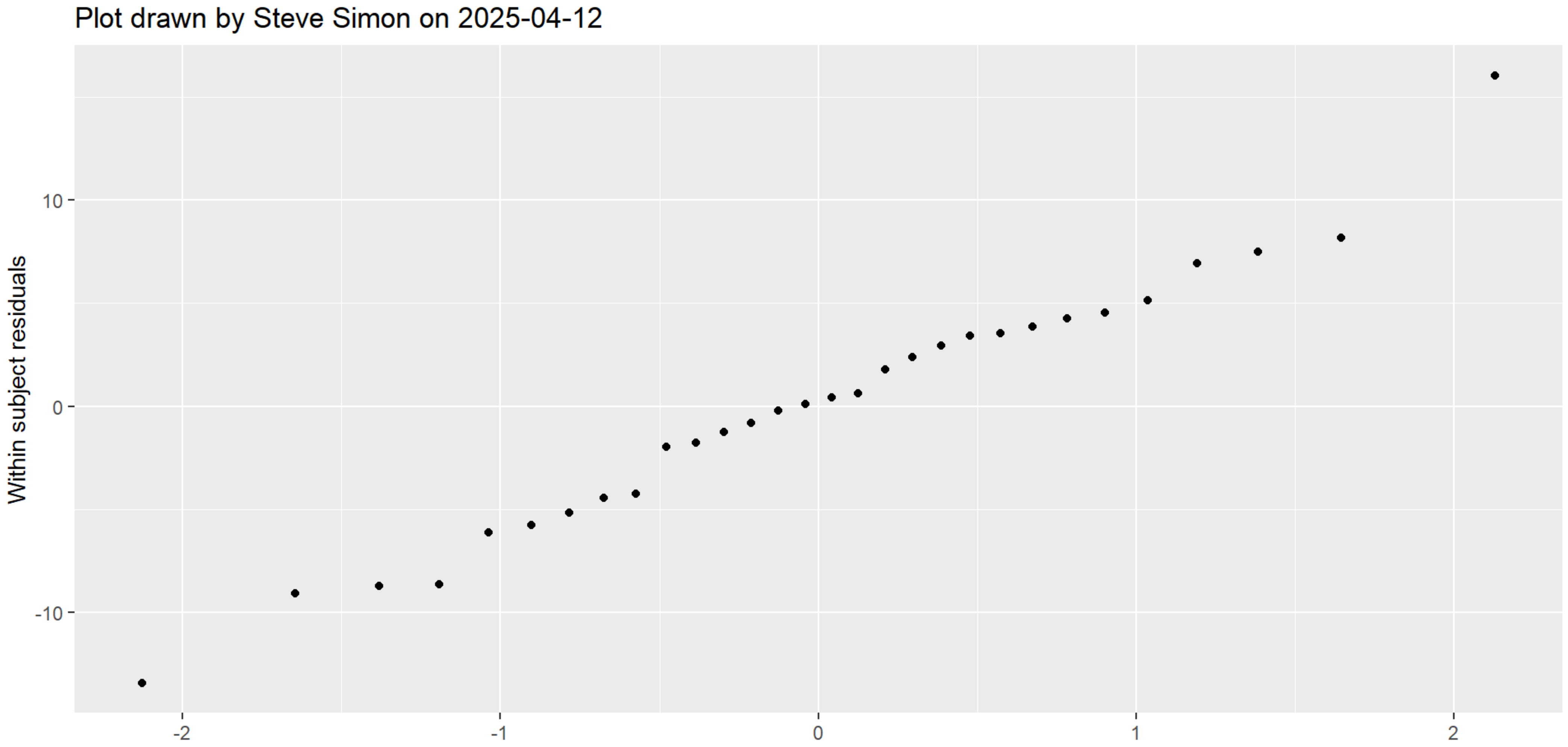
There are two or more normality checks. You need to look at the residuals within each cluster and see if they are normally distributed. Then you have to assess normality of the random intercepts. If there are random slopes, you have to assess normality there are well.

The random intercepts and random slopes model assumes that the relationship between time and the outcome variable are linear.

Linearity check



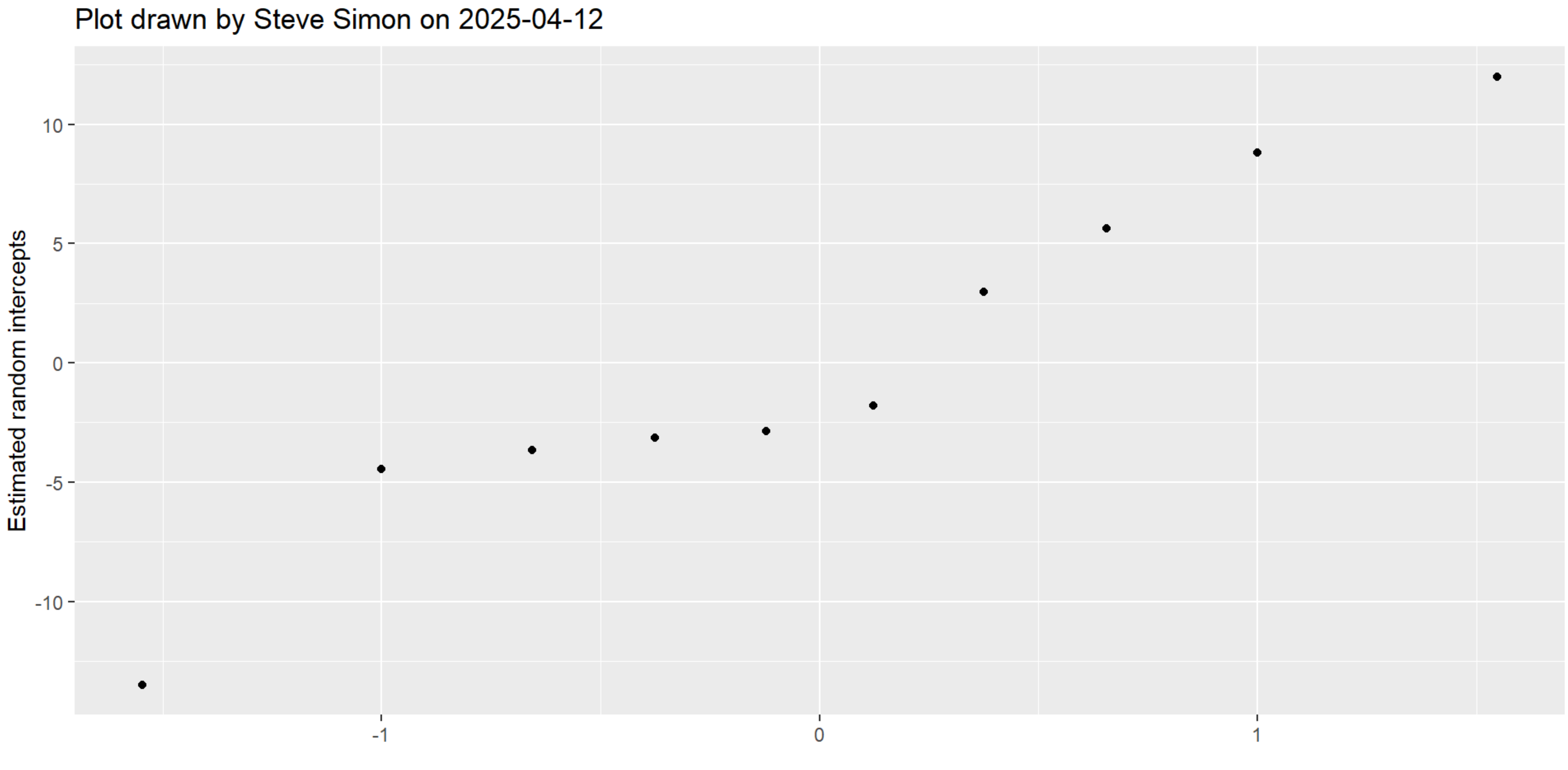
Normality check within clusters



Speaker notes

The residuals represent the deviations of individual time measurements from the regression line for an individual patient. That means the deviation from a trend line that has been randomly shifted up or down (random intercepts) and/or a randomly steeper or flatter trend line (random slopes). In this dataset, the normal probability plot looks fairly close to a straight line.

Normality check for random intercepts



Complications

- Not a problem
    - Missing values
    - Better than Last Observation Carried Forward
- Problems (more tedious than difficult)
    - Interactions
    - Nonlinear trends
    - Covariates
        - Between patients
        - Within patients

Speaker notes

Missing values are normally a big headache, but less so for the random intercepts or slopes models. If a patient missed a visit at a particular time point, you can easily extrapolate from the visits before and after. You are assuming linearity, after all.

In some clinical trials, if a patient came at the intermediate evaluation but did not show up at the final evaluation, the research team would replace the missing final outcome with the intermediate outcome. This technique, called Last Observation Carried Forward (LOCF) was not very popular when it was introduced and has been pretty much discredited. The random intercepts or slopes model would extrapolate the trend from the intermediate value, which works much better.

Interactions are always difficult, and when the interactions involve time, it can get a bit messy. Nonlinear trends over time are also a bit of a problem. In both cases, though, the work is more tedious than difficult. Interactions and nonlinearity mean that your interpretation of the results will require a bit more thought and you can’t come up with as simple a story to tell.

A covariate is a variable which is not of direct interest in the research, but one that you must take account of in order to produce a credible analysis. In just about any cancer study, you should track whether the patient is a smoker. It’s not something you’re interested in testing. The role of smoking in lung cancer and most other types of cancer was established many decades ago. You still have to account for smoking though because it can explain so much of the variation in your outcome. Failure to account for smoking would greatly reduce your power and precision.

There are two types of covariates. The ones that are fixed and do not change over time are called time constant covariates or between subject covariates. Patient demographics are time constant. Measurements done at baseline to assess how ill the patient was at the start of the study are time constant.

Covariates that change over time are called time varying covariates or within subject covariates. The extent to which a patient complies with taking his/her medication is a time varying covariate. Seasonal changes in temperature, humidity, or pollen counts are time varying covariates.

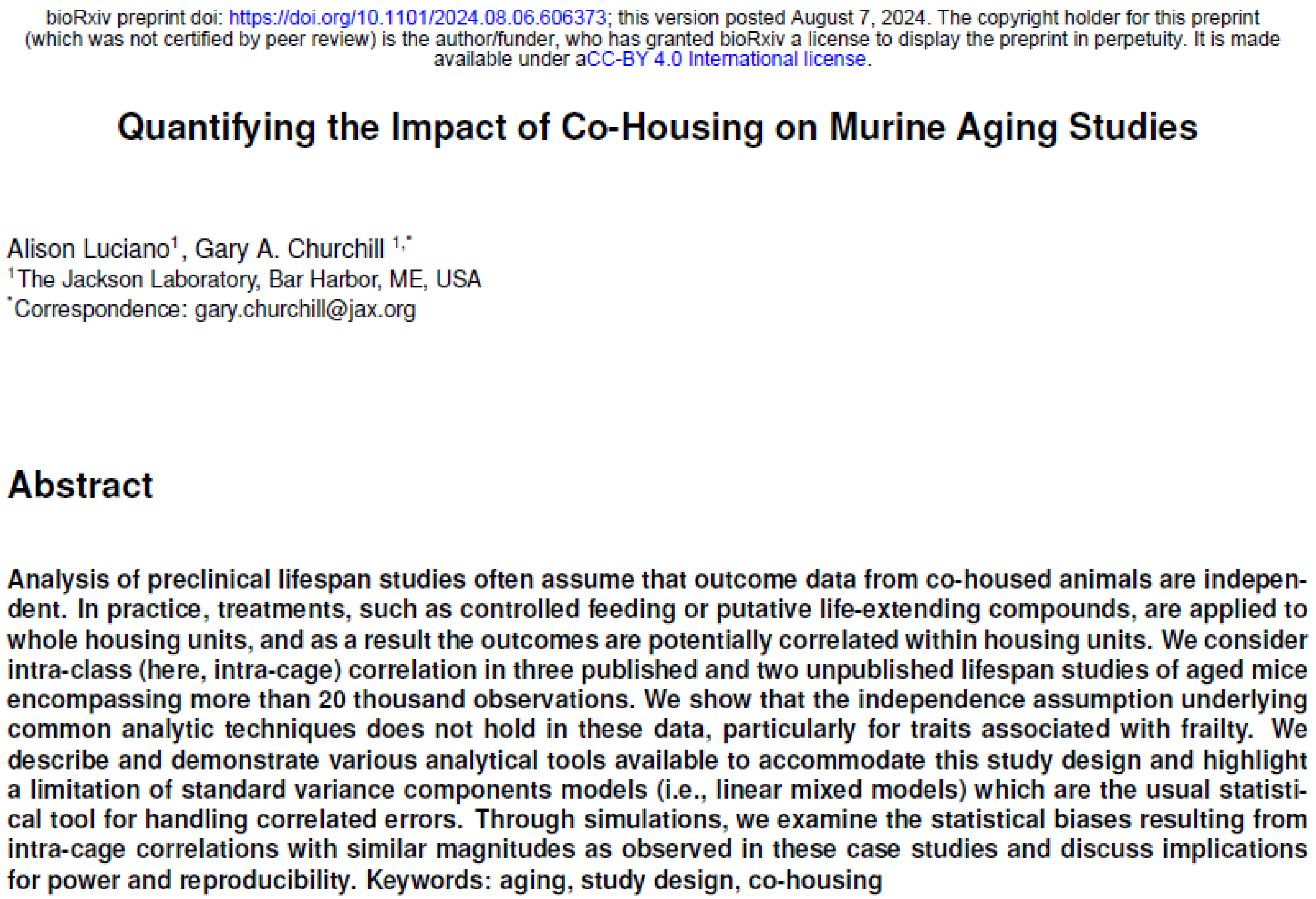
There is one important distinction between time constant and time varying covariates. The latter are much better at removing variation from your outcome, and can greatly improve your power and precision.

Live demo, checking assumptions

Break #5

- What you have learned
  - Assumptions and complications
- What’s coming next
  - Sample size justification

Effect of co-housing on sample size calculations, 1



Quantifying the Impact of Co-Housing on Murine Aging Studies

Alison Luciano[1], Gary A. Churchill [1,*]
[1] The Jackson Laboratory, Bar Harbor, ME, USA
[*] Correspondence: gary.churchill@jax.org

Abstract

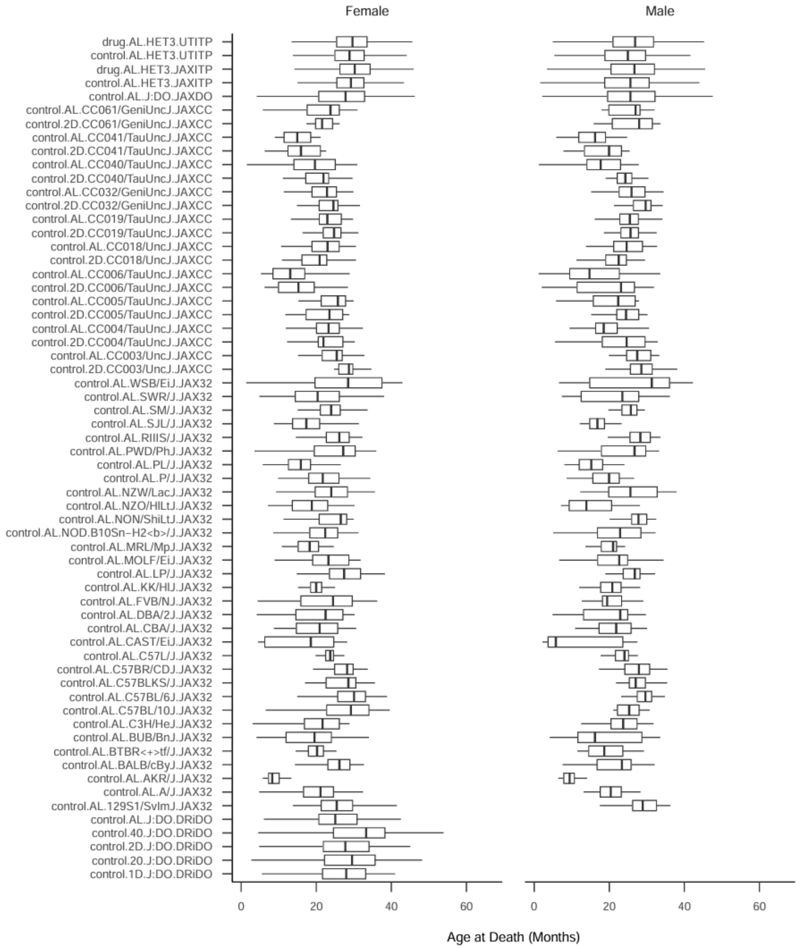
Analysis of preclinical lifespan studies often assume that outcome data from co-housed animals are independent. In practice, treatments, such as controlled feeding or putative life-extending compounds, are applied to whole housing units, and as a result the outcomes are potentially correlated within housing units. We consider intra-class (here, intra-cage) correlation in three published and two unpublished lifespan studies of aged mice encompassing more than 20 thousand observations. We show that the independence assumption underlying common analytic techniques does not hold in these data, particularly for traits associated with frailty. We describe and demonstrate various analytical tools available to accommodate this study design and highlight a limitation of standard variance components models (i.e., linear mixed models) which are the usual statistical tool for handling correlated errors. Through simulations, we examine the statistical biases resulting from intra-cage correlations with similar magnitudes as observed in these case studies and discuss implications for power and reproducibility. Keywords: aging, study design, co-housing

Speaker notes

Read a bit from this abstract:

“Analysis of preclinical lifespan studies often assume that outcome data from co-housed animals are independent. In practice, treatments, such as controlled feeding or putative life-extending compounds, are applied to whole housing units, and as a result the outcomes are potentially correlated within housing units. We consider intra-class (here, intra-cage) correlation in three published and two unpublished lifespan studies of aged mice encompassing more than 20 thousand observations. We show that the independence assumption underlying common analytic techniques does not hold in these data, particularly for traits associated with frailty. We describe and demonstrate various analytical tools available to accommodate this study design and highlight a limitation of standard variance components models (i.e., linear mixed models) which are the usual statistical tool for handling correlated errors. Through simulations, we examine the statistical biases resulting from intra-cage correlations with similar magnitudes as observed in these case studies and discuss implications for power and reproducibility.”

Effect of co-housing on sample size calculations, 2



Speaker notes

There are around 120 different conditions across five major studies. The total number of mice in all these studies is roughly twenty thousand.

If you are planning a study, find a boxplot similar to yours (in particular, look at the sex and the strain of mouse).

I just picked the bottom most graph for females. The median is about 28 months and the interquartile range is about 15.

Effect of co-housing on sample size calculations, 3

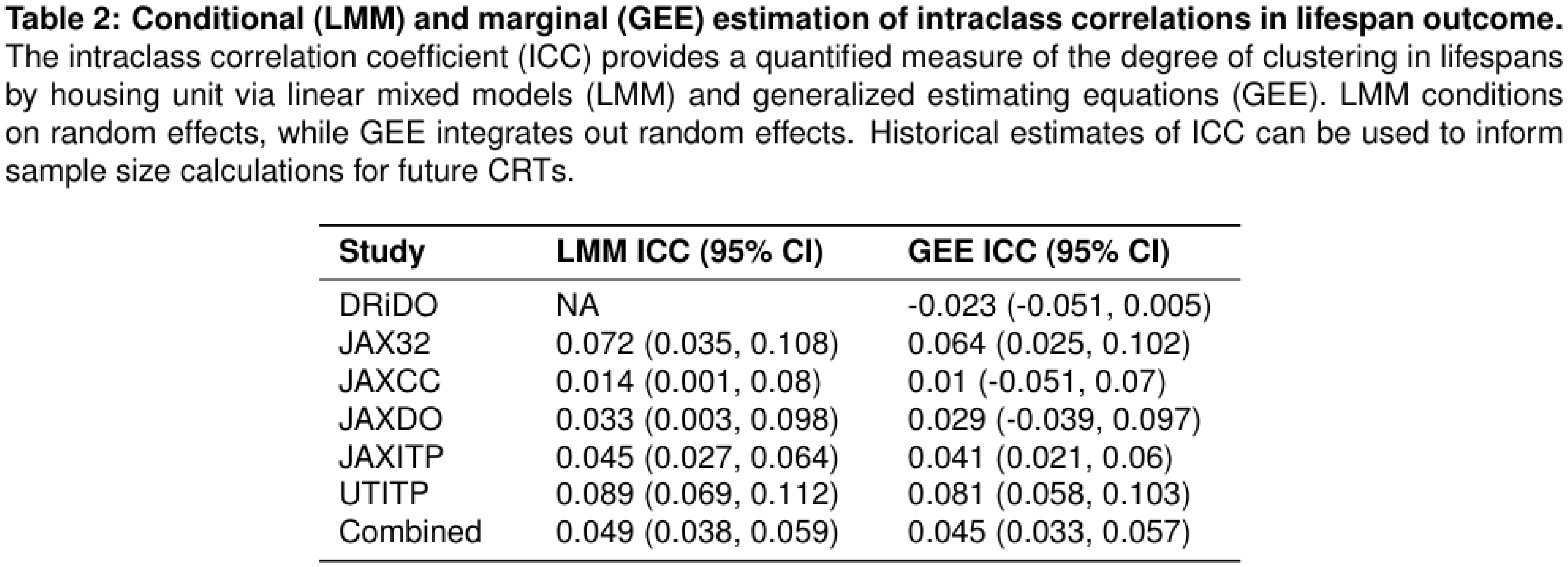


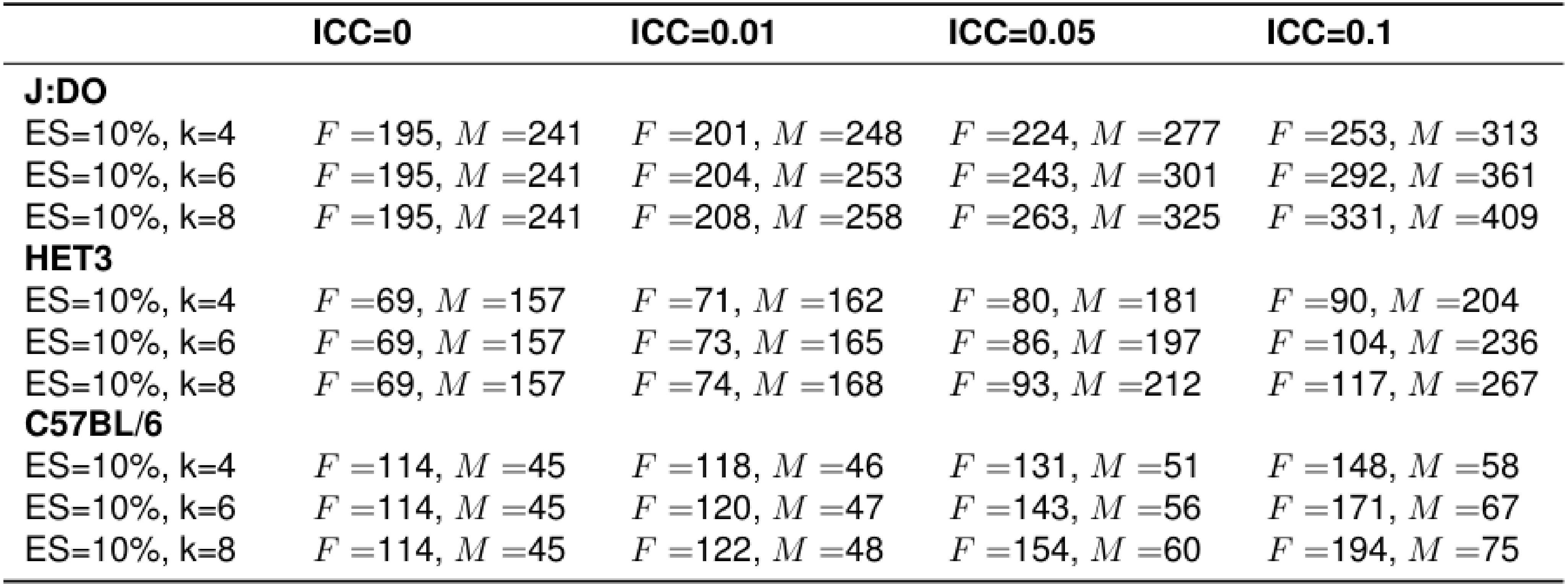
Table 2: Conditional (LMM) and marginal (GEE) estimation of intraclass correlations in lifespan outcome. The intraclass correlation coefficient (ICC) provides a quantified measure of the degree of clustering in lifespans by housing unit via linear mixed models (LMM) and generalized estimating equations (GEE). LMM conditions on random effects, while GEE integrates out random effects. Historical estimates of ICC can be used to inform sample size calculations for future CRTs.

Study	LMM ICC (95% CI)	GEE ICC (95% CI)
DRiDO	NA	-0.023 (-0.051, 0.005)
JAX32	0.072 (0.035, 0.108)	0.064 (0.025, 0.102)
JAXCC	0.014 (0.001, 0.08)	0.01 (-0.051, 0.07)
JAXDO	0.033 (0.003, 0.098)	0.029 (-0.039, 0.097)
JAXITP	0.045 (0.027, 0.064)	0.041 (0.021, 0.06)
UTITP	0.089 (0.069, 0.112)	0.081 (0.058, 0.103)
Combined	0.049 (0.038, 0.059)	0.045 (0.033, 0.057)

Speaker notes

There are two different ways to estimate the intraclass correlation. The combined estimate from a linear mixed model (bottom left) is 0.049. This is a reasonable choice, but if you know some details of these experiments, you may prefer an estimate from an experiment that is similar to the one you plan.

Effect of co-housing on sample size calculations, 4



	ICC=0	ICC=0.01	ICC=0.05	ICC=0.1
J:DO				
ES=10%, k=4	F =195, M =241	F =201, M =248	F =224, M =277	F =253, M =313
ES=10%, k=6	F =195, M =241	F =204, M =253	F =243, M =301	F =292, M =361
ES=10%, k=8	F =195, M =241	F =208, M =258	F =263, M =325	F =331, M =409
HET3				
ES=10%, k=4	F =69, M =157	F =71, M =162	F =80, M =181	F =90, M =204
ES=10%, k=6	F =69, M =157	F =73, M =165	F =86, M =197	F =104, M =236
ES=10%, k=8	F =69, M =157	F =74, M =168	F =93, M =212	F =117, M =267
C57BL/6				
ES=10%, k=4	F =114, M =45	F =118, M =46	F =131, M =51	F =148, M =58
ES=10%, k=6	F =114, M =45	F =120, M =47	F =143, M =56	F =171, M =67
ES=10%, k=8	F =114, M =45	F =122, M =48	F =154, M =60	F =194, M =75

Speaker notes

The researchers did some sample size calculations for a between cluster comparison. The general trend is that you need a larger sample size as the intraclass correlation increases. Notice that even a very small intraclass correlation has an impact on the sample size. Also notice that housing more mice per cage also causes an increase in sample size. This is fairly intutive if you think about it. More animals in a cage means more positive correlation spread across a larger number of animals.

Sample size estimate without clustering

```r
delta <- 4
sd_independence <- 11.1

power.t.test(
    n=NULL,
    delta=4,
    sd=sd_independence,
    sig.level=0.05,
    power=0.8,
    type="two.sample") |>
    tidy() -> sample_size_1

sample_size_1
```

```
# A tibble: 1 × 5
      n delta    sd sig.level power
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1  122.     4  11.1      0.05   0.8
```

Sample size estimate with clustering, 4 animals per cage

```r
deff <- 1+(4-1)*0.049
sd_correlated <- sd_independence * sqrt(deff)

power.t.test(
    n=NULL,
    delta=4,
    sd=sd_correlated,
    sig.level=0.05,
    power=0.8,
    type="two.sample") |>
    tidy()
```

```
# A tibble: 1 × 5
      n delta    sd sig.level power
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1  140.     4  11.9      0.05   0.8
```

Sample size estimate with clustering, 8 animals per cage

```r
deff <- 1+(8-1)*0.049
sd_correlated <- sd_independence * sqrt(deff)

power.t.test(
    n=NULL,
    delta=4,
    sd=sd_correlated,
    sig.level=0.05,
    power=0.8,
    type="two.sample") |>
    tidy()
```

```
# A tibble: 1 × 5
      n delta    sd sig.level power
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1  163.     4  12.9      0.05   0.8
```

You could also just multiply the sample size by the design effect

```
sample_size_1$n * (1+(4-1)*0.049)
```

```
[1] 139.7624
```

```
sample_size_1$n * (1+(8-1)*0.049)
```

```
[1] 163.6451
```

Summary

- What you have learned
    - Mathematical formulation of random intercepts model
    - Description of HIV-intervention data
    - Random intercepts model using hiv-intervention data
    - Mathematical formulation of random slopes model
    - Assumptions and complications
    - Sample size justification