# Analyzing Intensive Longitudinal Data

# Overview

- General background on main type of RQ with ILD.
- Discuss common analysis methods for RQs (focusing on their overlap).
- Discuss some common issues you'll likely encounter.
- Work on analyses yourself in R.

# Two main RQ of ILD

- Many things you can do with ILD.

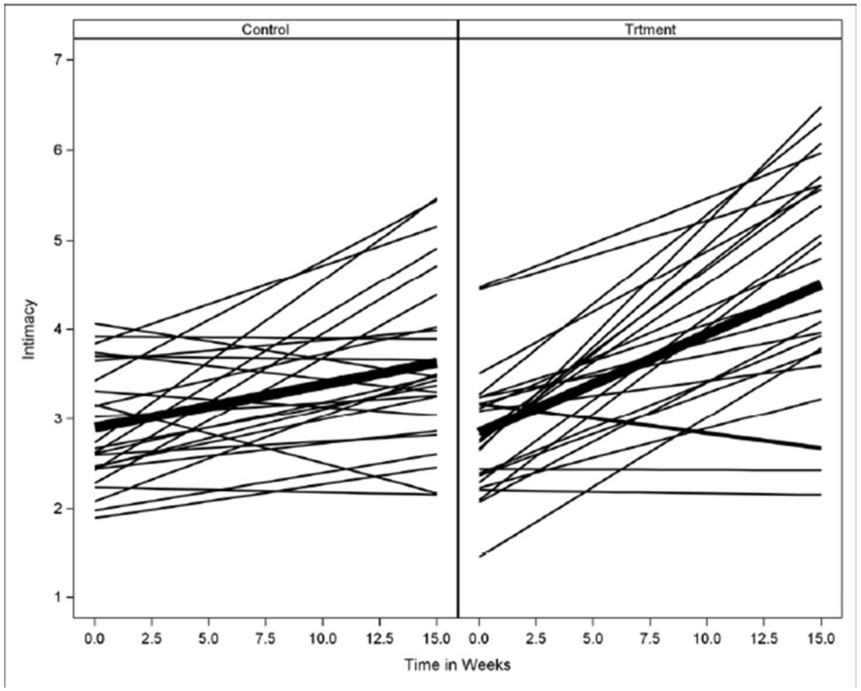
- Can really “dive in” and study processes as they are unfolding over time.
  - How does craving fluctuate and evolve across time?
  - Are there between-person differences in this process?
  - How are (differences in) processes related to other variables (e.g., alcohol abuse)?

- Despite this diversity, most RQ fall into one of two main categories:
  - Studying systematic change in a construct over time.
  - Studying moment-to-moment fluctuations around mean-score (reversible change).

# Systematic Change

# Systematic Change

- Look at increases or decreases across time.
- Also look at individual differences in change (and how these differences are related to other variables)
- Typically studied using:
  - Multilevel regression with “time” as predictor
  - Latent Growth Curve Models (LGCMs)
- Doesn’t have to be linear change, but we’ll focus on that here.

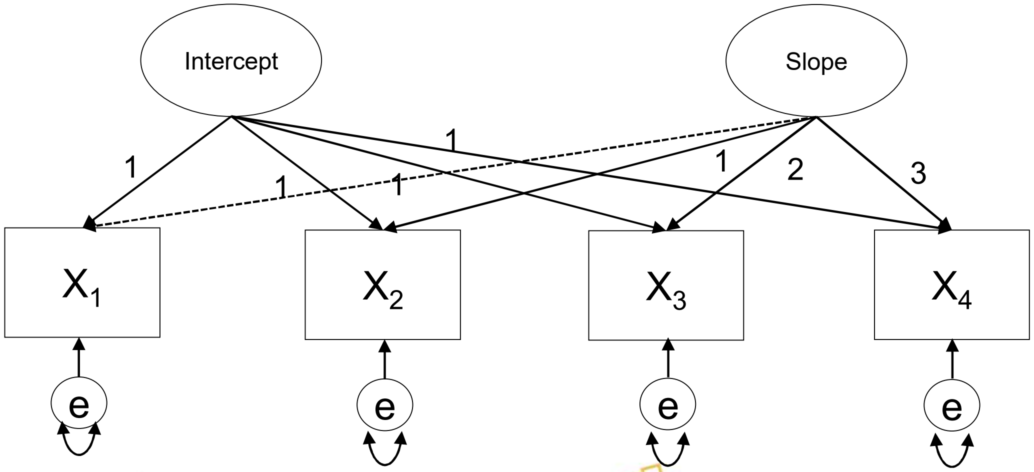


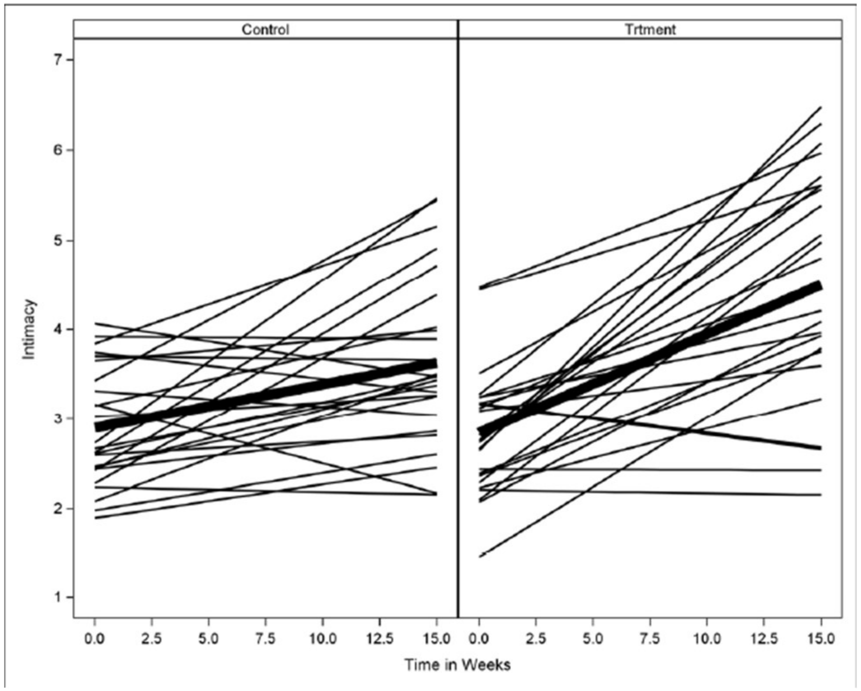
# Systematic Change

- Multilevel regression with “time” as predictor

$$Y_{it} = b_{0i} + \ b_{1i}Time_{ij} + \ \epsilon_{it}$$

- Latent Growth Curve Models (LGCMs)





# Systematic Change

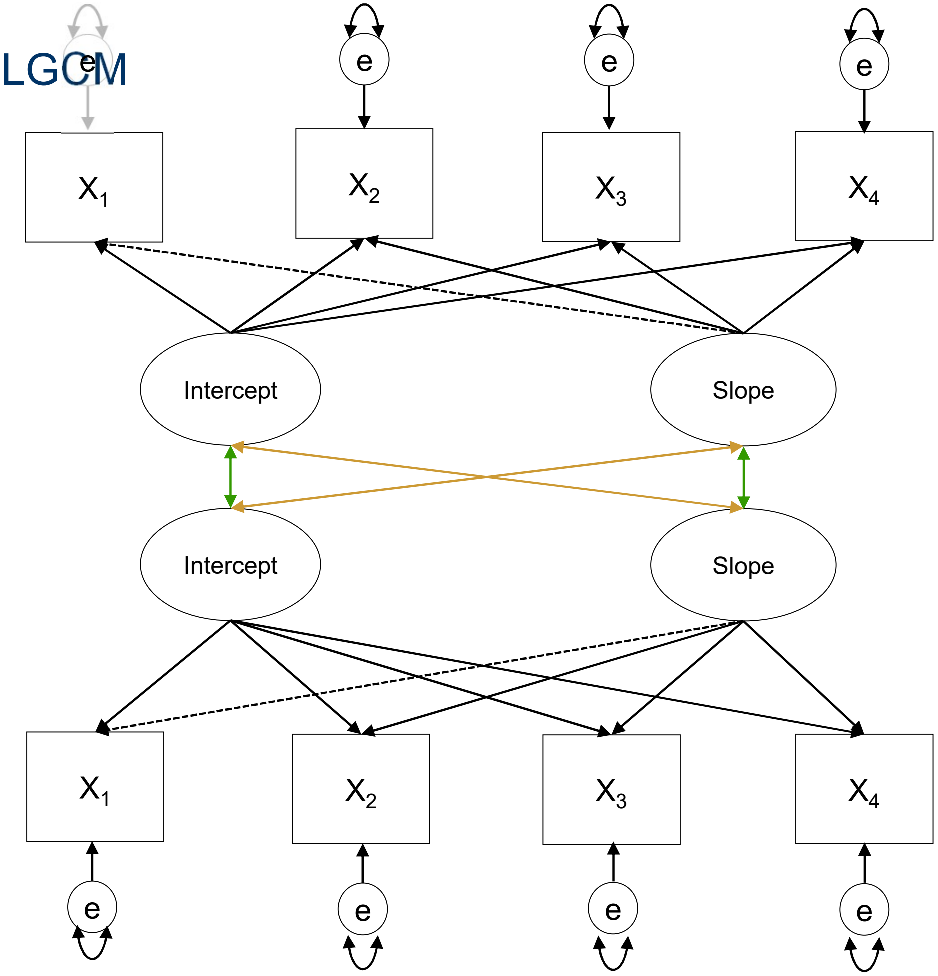
- Multilevel regression

$$\begin{bmatrix} Y_{it} \\ X_{it} \end{bmatrix} = \begin{bmatrix} b_{Y0i} \\ b_{X0i} \end{bmatrix} + \begin{bmatrix} b_{Y1i} \\ b_{X1i} \end{bmatrix} \begin{bmatrix} Time_{it} \\ Time_{it} \end{bmatrix} + \begin{bmatrix} \epsilon_{Yit} \\ \epsilon_{Xit} \end{bmatrix}$$

$$b_{Y0i} = \gamma_{Y00} + \gamma_{Y01} Z_i + u_{Y0}$$
$$b_{X0i} = \gamma_{X00} + \gamma_{X01} Z_i + u_{X0}$$

$$b_{Y1i} = \gamma_{Y10} + \gamma_{Y11} Z_i + u_{Y1}$$
$$b_{X1i} = \gamma_{X10} + \gamma_{X11} Z_i + u_{X1}$$

- LGCM



# Reversible Change

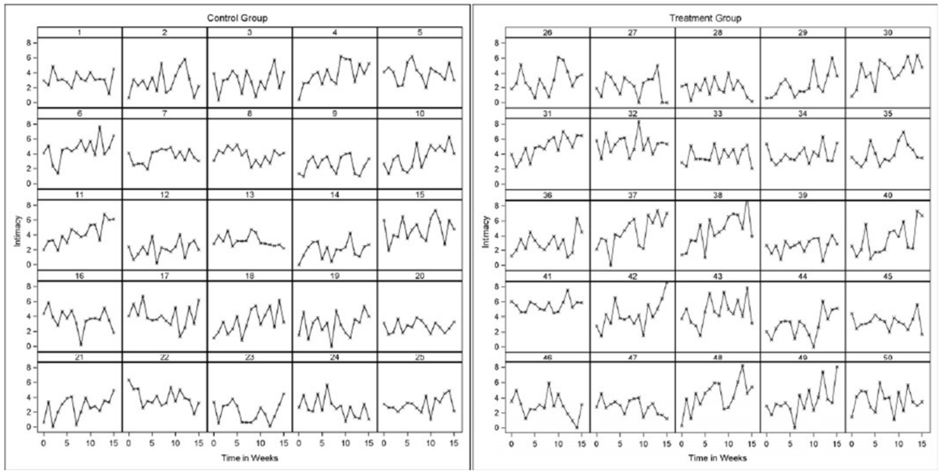
# Systematic Change

- Look at variation around a stable mean.

- Also look at individual differences in variation (and how these differences are related to other variables)

- Typically studied using:
  - (Vector) Autoregressive models (VAR)

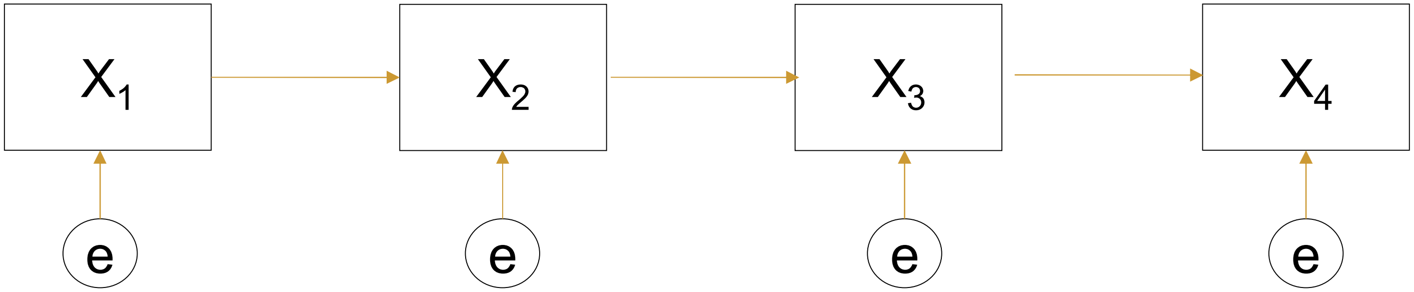
- Can also look at variation around predicted mean (i.e., a regression line).
  - This would be a LGM/multilevel regression with “time”, with a structured residuals.



# Reversible Change

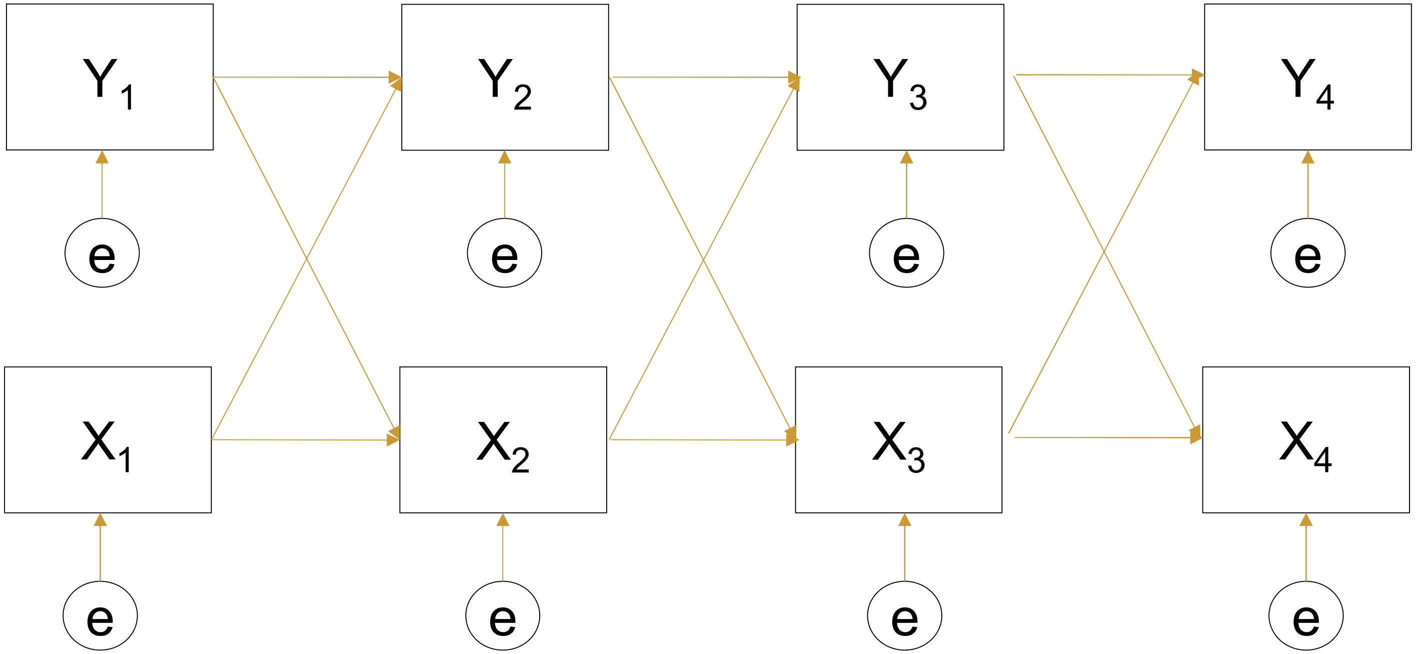
- One type of model used is the AR model

- Don’t focus on overall change but on how current score is predicted from previous ones(s)



# Reversible Change

- Will often be interested in (longitudinal) relation between two or more variables (VAR)



# Reversible Change

- VAR Model

$$\begin{bmatrix} Y_{it} \\ X_{it} \end{bmatrix} = \begin{bmatrix} b_{Y0i} \\ b_{X0i} \end{bmatrix} + \begin{bmatrix} b_{Y1i} & b_{XY} \\ b_{YXi} & b_{X1i} \end{bmatrix} \begin{bmatrix} Y_{it-1} \\ X_{it-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{Yit} \\ \epsilon_{Xit} \end{bmatrix}$$

- Multilevel regression

$$\begin{bmatrix} Y_{it} \\ X_{it} \end{bmatrix} = \begin{bmatrix} b_{Y0i} \\ b_{X0i} \end{bmatrix} + \begin{bmatrix} b_{Y1} \\ b_{X1} \end{bmatrix} \begin{bmatrix} Time_{it} \\ Time_{it} \end{bmatrix} + \begin{bmatrix} \epsilon_{Yi} \\ \epsilon_{Xit} \end{bmatrix}$$

Tilburg University

tesc

# Reversible Change

$$\begin{bmatrix} Y_{it} \\ X_{it} \end{bmatrix} = \begin{bmatrix} b_{Y0i} \\ b_{X0i} \end{bmatrix} + \begin{bmatrix} b_{Y1i} & b_{XYi} \\ b_{YXi} & b_{X1i} \end{bmatrix} \begin{bmatrix} Y_{it-1} \\ X_{it-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{Yit} \\ \epsilon_{Xit} \end{bmatrix}$$

$b_{Y0}/b_{X0}$= Long run tendency → Think “mean”.

$b_{Y1}/b_{X1}$= Autoregressive parameter → inertia.

$b_{YX}/b_{XY}$= Cross-lagged effects.

$\epsilon_{Yt}/\epsilon_{Xt}$= Residual/Innovation → All variation that can not be  predicted by previous measurement.

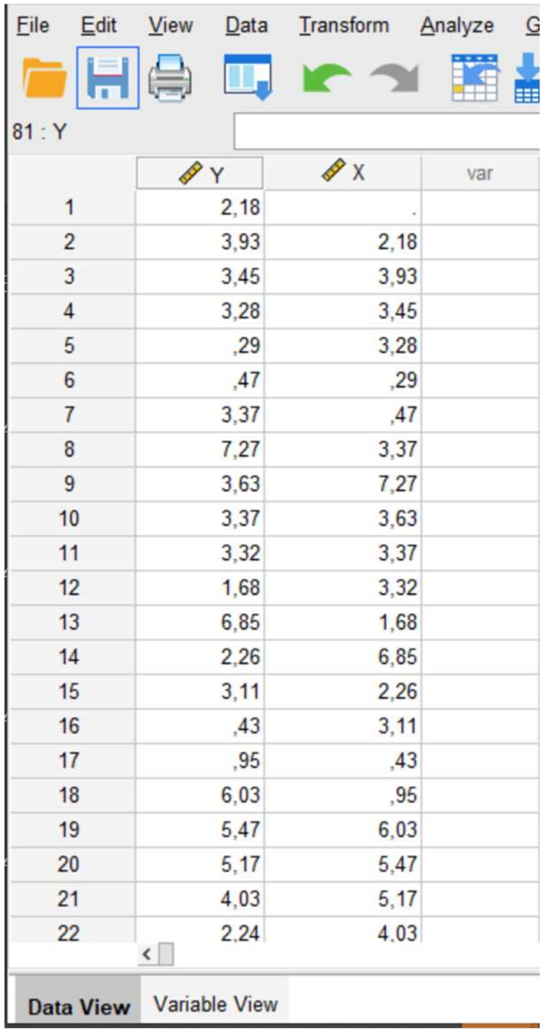
# Reversible Change

- Can run an (V)AR model with multilevel regression too!

$$Y_{it} = b_{0i} + b_{1i}Time_{it} + \epsilon_{it}$$

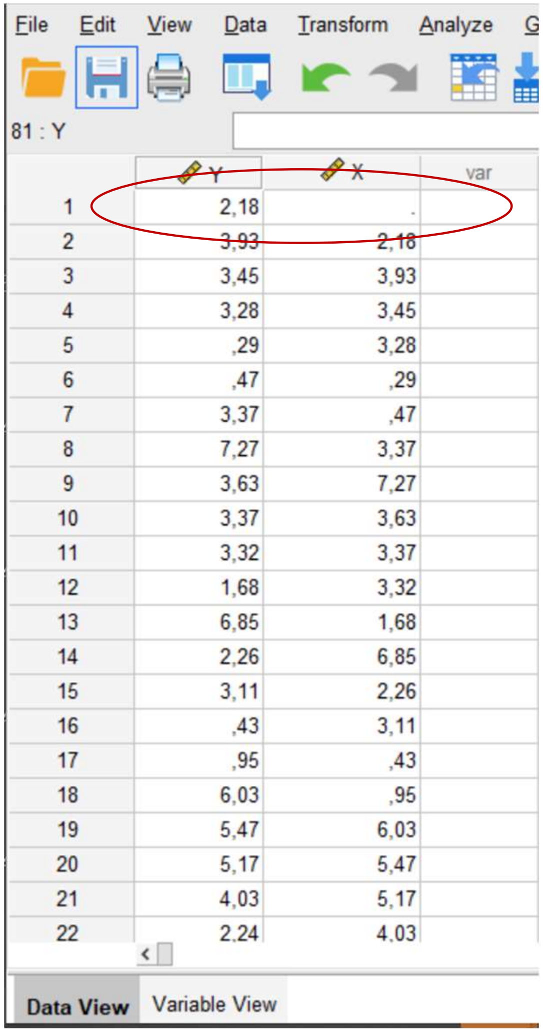
$$Y_{it} = b_{0i} + b_{1i}Y_{i,t-1} + \epsilon_{it}$$

# Reversible Change



| | Y | X | var |
|---|---|---|---|
| 1 | 2,18 | . | |
| 2 | 3,93 | 2,18 | |
| 3 | 3,45 | 3,93 | |
| 4 | 3,28 | 3,45 | |
| 5 | ,29 | 3,28 | |
| 6 | ,47 | ,29 | |
| 7 | 3,37 | ,47 | |
| 8 | 7,27 | 3,37 | |
| 9 | 3,63 | 7,27 | |
| 10 | 3,37 | 3,63 | |
| 11 | 3,32 | 3,37 | |
| 12 | 1,68 | 3,32 | |
| 13 | 6,85 | 1,68 | |
| 14 | 2,26 | 6,85 | |
| 15 | 3,11 | 2,26 | |
| 16 | ,43 | 3,11 | |
| 17 | ,95 | ,43 | |
| 18 | 6,03 | ,95 | |
| 19 | 5,47 | 6,03 | |
| 20 | 5,17 | 5,47 | |
| 21 | 4,03 | 5,17 | |
| 22 | 2,24 | 4,03 | |

# Reversible Change



|  | Y | X | var |
|---|---|---|---|
| 1 | 2,18 | . |  |
| 2 | 3,93 | 2,18 |  |
| 3 | 3,45 | 3,93 |  |
| 4 | 3,28 | 3,45 |  |
| 5 | ,29 | 3,28 |  |
| 6 | ,47 | ,29 |  |
| 7 | 3,37 | ,47 |  |
| 8 | 7,27 | 3,37 |  |
| 9 | 3,63 | 7,27 |  |
| 10 | 3,37 | 3,63 |  |
| 11 | 3,32 | 3,37 |  |
| 12 | 1,68 | 3,32 |  |
| 13 | 6,85 | 1,68 |  |
| 14 | 2,26 | 6,85 |  |
| 15 | 3,11 | 2,26 |  |
| 16 | ,43 | 3,11 |  |
| 17 | ,95 | ,43 |  |
| 18 | 6,03 | ,95 |  |
| 19 | 5,47 | 6,03 |  |
| 20 | 5,17 | 5,47 |  |
| 21 | 4,03 | 5,17 |  |
| 22 | 2,24 | 4,03 |  |

# Reversible Change

$$Y_{it} = b_{0i} + \ b_{1i}Y_{i,t-1} + \ \epsilon_{it}$$

- Couple of things:
  - The intercept is not the mean!

$$\mu = \frac{b_o}{1 - \ b_1^2}$$

  - To get the mean you need to group-mean center the lagged predictor $Y_{i,t-1}$
  - This model assumes there is no trend! If there is, remove it (detrenting) or model it!!

# Modeling

# The “rules” of analysis

- Analyzing your data is all about modeling.

- If your data is linear, model it as such!
  - If it isn’t…don’t.

# The “rules” of analysis

- Analyzing your data is all about modeling.

- If your data is linear, model it as such!
  - If it isn’t…don’t.

- If the residual variance is the same “across” the board, model it as such!
  - If it isn’t…don’t

# The “rules” of analysis

- Analyzing your data is all about modeling.

- If your data is linear, model it as such!
  - If it isn’t…don’t.

- If the residual variance is the same “across” the board, model it as such!
  - If it isn’t…don’t

- If scores are normally distributed around the mean estimates….you get the idea

# Dependent Data

- Hierarchical data, or data with dependence between the observations, is no different that outliers etc.

- If dependence is a characteristic of your data, model it.

- Maybe do it with “multilevel analysis” if needed/appropriate(!!).

# Example

- Let’s start with an example,

- I have 50 participants who I measure several (e.g., five) times

- Interested whether their GPA increases across time and if their GPA is related to the amount of time they spend on their job at each time-point

- List of all the variables:
  - *ID*: participant identification variable
  - *time*: predictor indicating the time-point at which the measurement was taken
  - *job*: predictor indicating the average number of hours worked per week at that measurement occasion
  - *Gender*: participant-level predictor indicating biological sex at birth (0 = male, 1 = female)
  - *GPA:* a participants grade-point average.

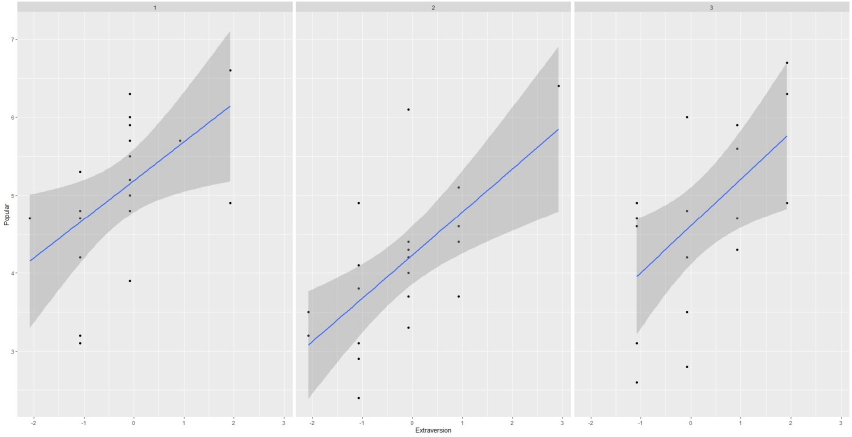
# Example

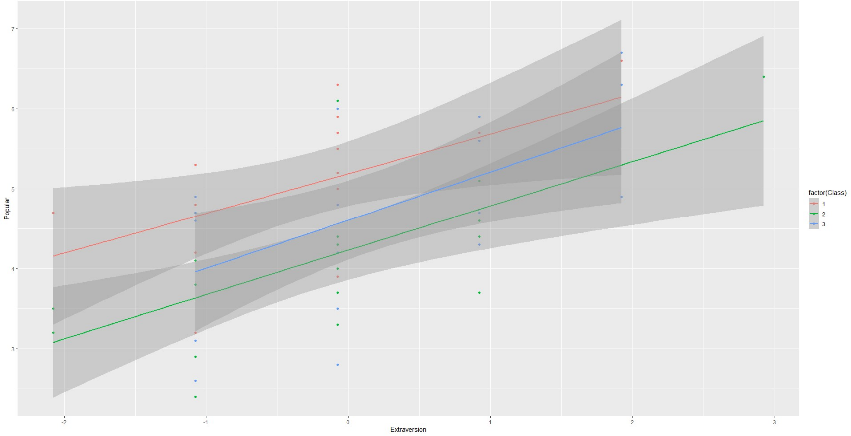
- How would you analyze this?
- Just mention all the different ways you can think of!
- No wrong answers!

| ID | Time | Job | Gender | GPA |
|---|---|---|---|---|
| 1 | 1 | 5 | 1 | 3,30 |
| 1 | 2 | 4 | 1 | 3,90 |
| 1 | 3 | 8 | 1 | 4,30 |
| 1 | 4 | 2 | 1 | 4,70 |
| 1 | 5 | 5 | 1 | 4,00 |
| 2 | 1 | 2 | 0 | 2,70 |
| 2 | 2 | 2 | 0 | 2,90 |
| 2 | 3 | 3 | 0 | 3,20 |
| 2 | 4 | 4 | 0 | 3,20 |
| 2 | 5 | 3 | 0 | 3,90 |

# Some Options: Option 1

- Analyze all participants separately



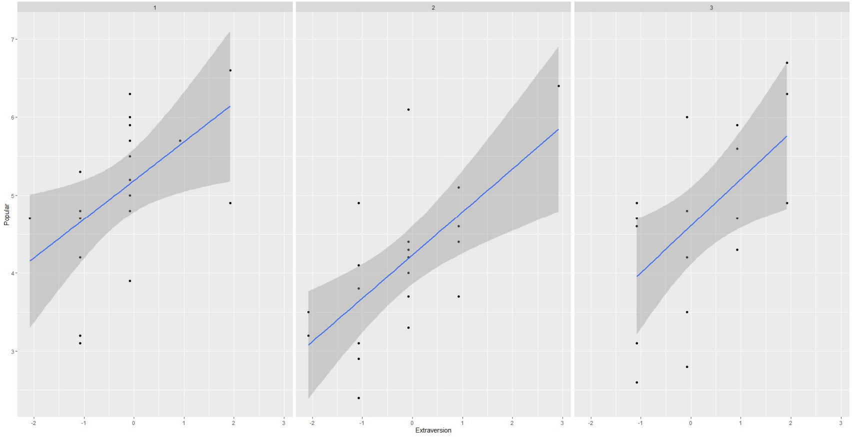


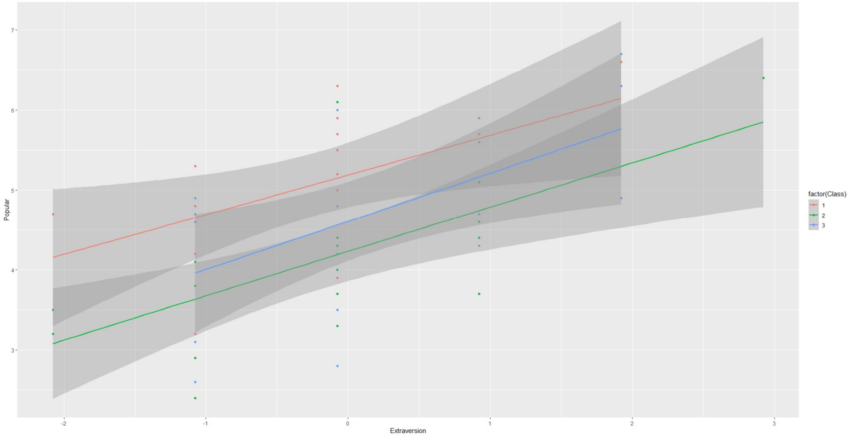
# Some Options: Option 1

- Analyze all participants separately

- Benefits?

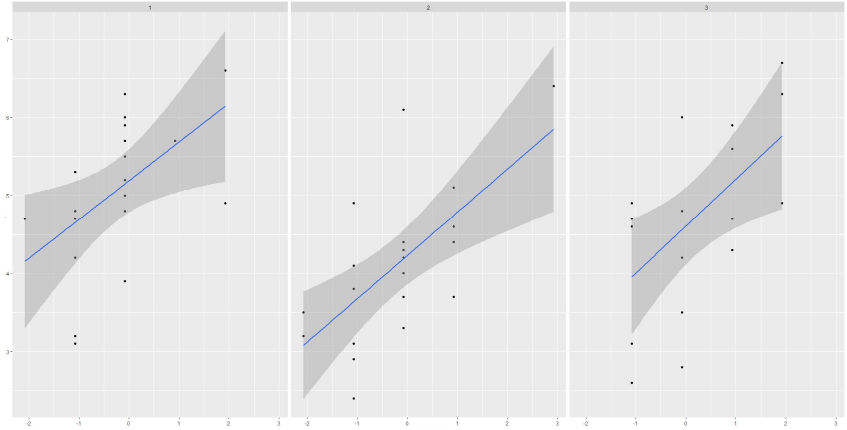
- Disadvantages?

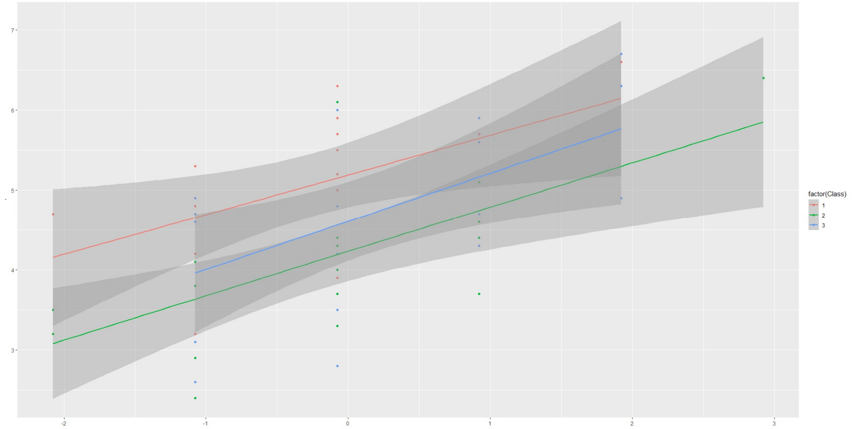




# Some Options : Option 1

- Analyze all participants separately
- Benefits?
- Disadvantages?
- How did we deal with the dependence in our data?



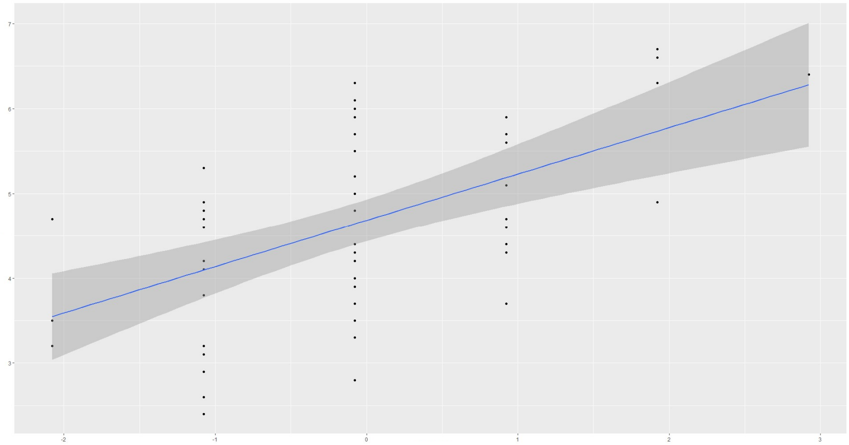


# Richard McElreath

- *Many statistical models also have anterograde amnesia. As the models move from one cluster—individual, group, location—in the data to another, estimating parameters for each cluster, they forget everything about the previous clusters. They behave this way, because the assumptions force them to. Any of the models from previous chapters that used dummy variables to handle categories are programmed for amnesia. These models implicitly assume that nothing learned about any one category informs estimates for the other categories—the parameters are independent of one another and learn from completely separate portions of the data. This would be like forgetting you had ever been in a café, each time you go to a new café. Cafés do differ, but they are also alike*

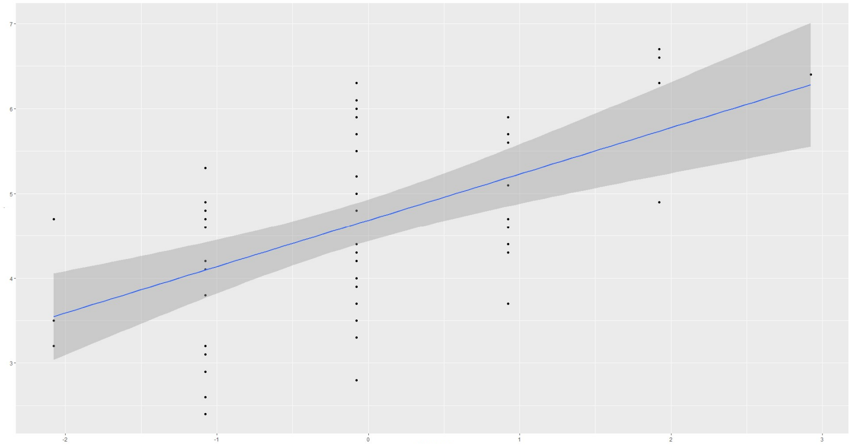
# Some Options: Option 2

- Ok…so…analyze together then?



# Some Options: Option 2

- Ok…so…analyze together then?
- What is the main issue in doing so?



# Some Options: Option 2

- Ok…so…analyze together then?
- What is the main issue in doing so?
- What is the problem with Gender specifically?

| ID | Time | Job | Gender | GPA |
|---|---|---|---|---|
| 1 | 1 | 5 | 1 | 3,30 |
| 1 | 2 | 4 | 1 | 3,90 |
| 1 | 3 | 8 | 1 | 4,30 |
| 1 | 4 | 2 | 1 | 4,70 |
| 1 | 5 | 5 | 1 | 4,00 |
| 2 | 1 | 2 | 0 | 2,70 |
| 2 | 2 | 2 | 0 | 2,90 |
| 2 | 3 | 3 | 0 | 3,20 |
| 2 | 4 | 4 | 0 | 3,20 |
| 2 | 5 | 3 | 0 | 3,90 |

# Some Options: Option 2

- Hint: How much data do we have?

$$n_{eff} = \frac{n}{1 + (n_{clus} - 1)\rho}$$

- Ok, so we have less data than data points.

- Now what?

| ID | Time | Job | Gender | GPA |
|---|---|---|---|---|
| 1 | 1 | 5 | 1 | 3,30 |
| 1 | 2 | 4 | 1 | 3,90 |
| 1 | 3 | 8 | 1 | 4,30 |
| 1 | 4 | 2 | 1 | 4,70 |
| 1 | 5 | 5 | 1 | 4,00 |
| 2 | 1 | 2 | 0 | 2,70 |
| 2 | 2 | 2 | 0 | 2,90 |
| 2 | 3 | 3 | 0 | 3,20 |
| 2 | 4 | 4 | 0 | 3,20 |
| 2 | 5 | 3 | 0 | 3,90 |

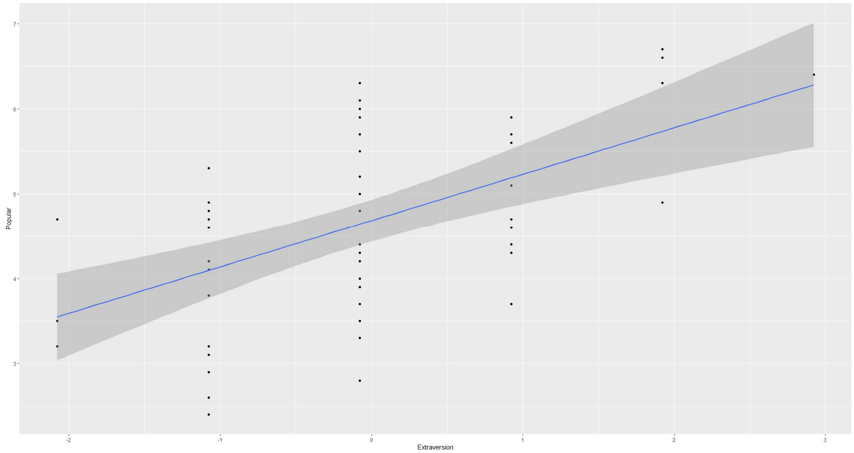
# Some Options: Option 2

- Adjust se’s

$$v_{eff} = v(1 + (n_{clus} - 1)\rho)$$

- Fortunately, smart ways to do this.

- Cluster robust s.e’s



# Some Options: Option 2

- Pro-tip: Use the formula below in combination with G*power for power analyses!

$$n_{eff} = \frac{n}{1+(n_{clus}-1)\rho} \quad \longrightarrow \quad n = n_{eff} \ * (1 + (n_{clus} - 1)\rho)$$

- This way you can do a power analysis even if you don’t have all the information needed for a simulation study.

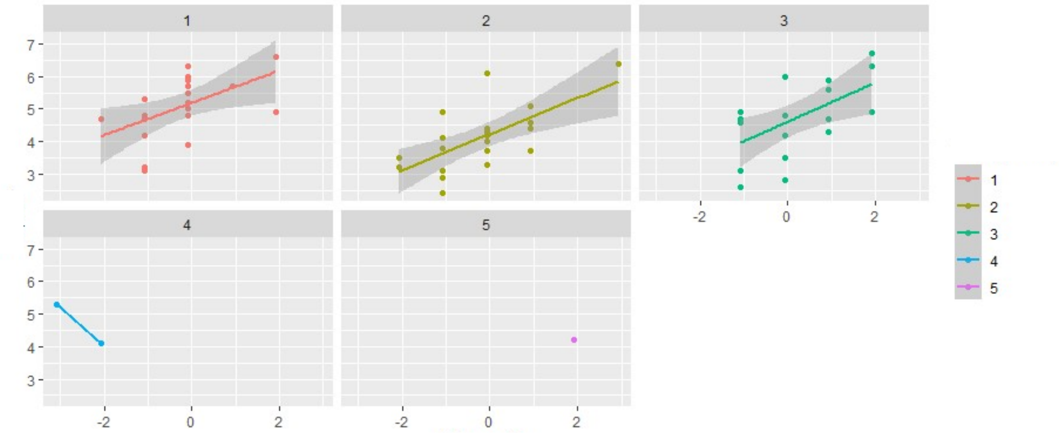
- G*power gives you the $n_{eff}$ you need

# Some Options: Option 3 ---- Pool(ing) Party

- Who ever heard of partial-pooling?

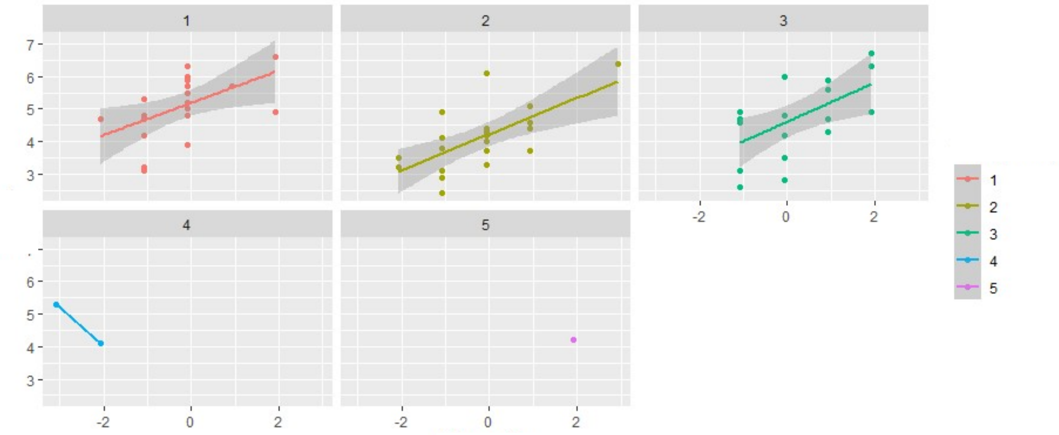
- I'll add two participants to our dataset to illustrate it better
  - One participant with two observations
  - One participant with one observation

# Pool(ing) Party



- What would be a problem if I analyzed everyone as N=1?

# Pool(ing) Party

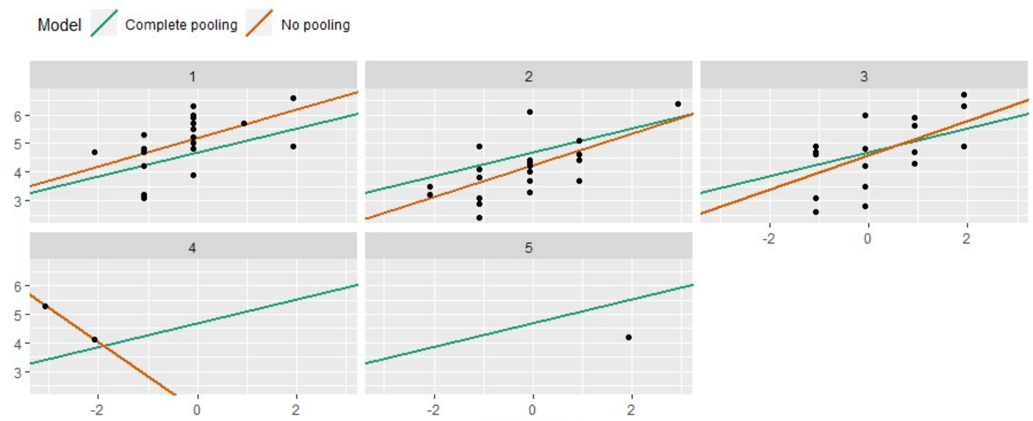


- Each panel shows an independently estimated regression line.

- This approach of fitting a separate line for each participant is sometimes called the no pooling model

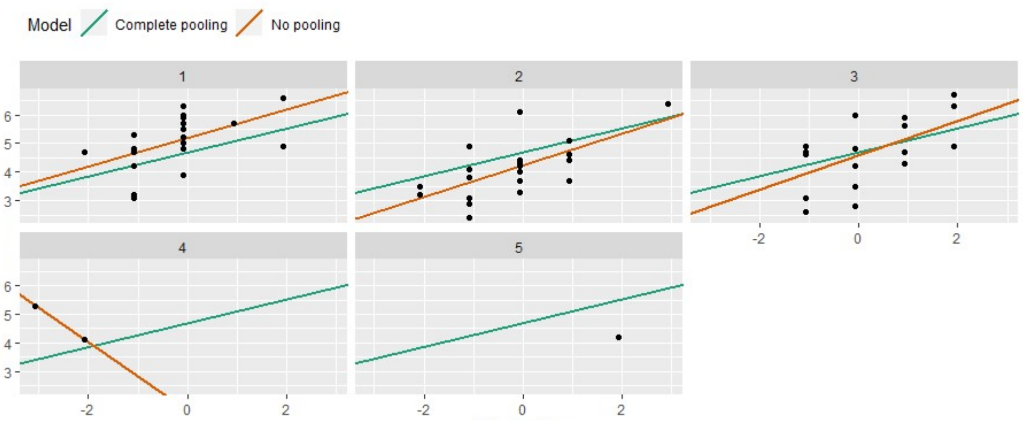
- Information from different participants is NOT combined or pooled together.

# Pool(ing) Party



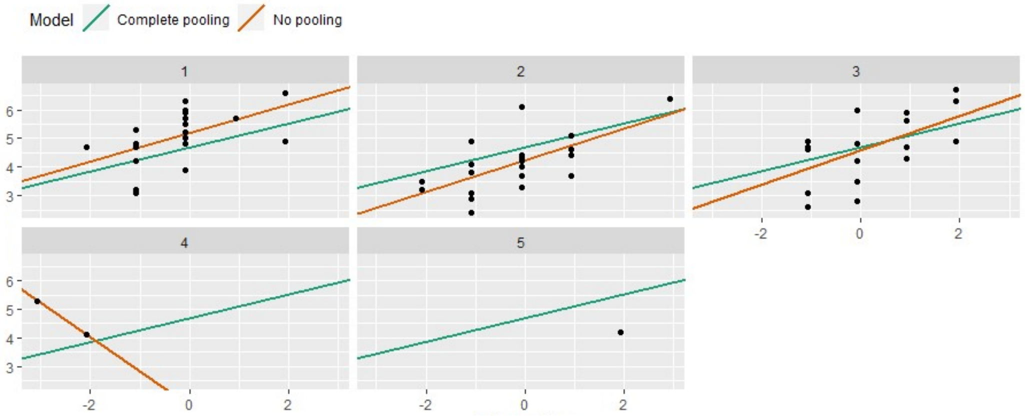
- Could also combine information from all participants (like we saw earlier).
- Fit a single line for the combined data set, unaware that the data came from different participants.

# Pool(ing) Party



- Notice anything?

# Pool(ing) Party



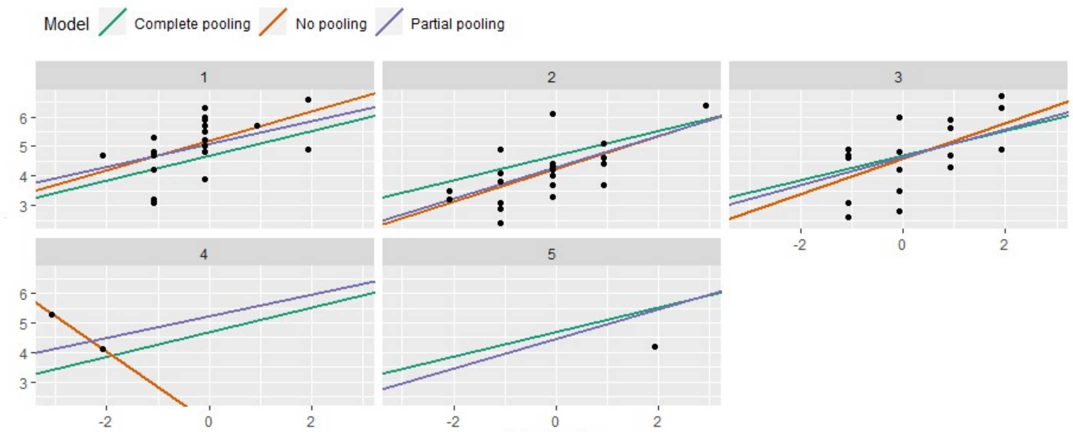
- Notice anything?

- Also get estimates for participant 5 now!

- And the estimate for participant 4 is not as different from those of the other participants

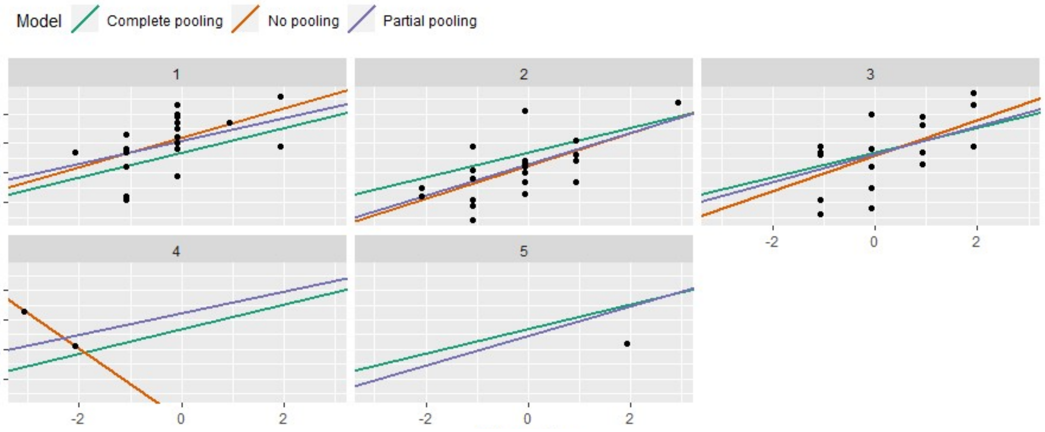
- Everyone get’s the same line now however.

# Pool(ing) Party



- We pool information from all the lines together to improve our estimates of each individual line.
- After seeing the 3 trend lines for the classes with complete data, we can make an informed guess about the trend lines for the two classes with incomplete data.
- But we still allow for differences between people!!

# Pool(ing) Party



- Most of the time, the no pooling and partial pooling lines are almost the same.

- When the two differ, it’s because the partial pooling line is pulled slightly towards the complete-pooling line.

- Amount of pull depends on:
  - amount of data.
  - how extreme a person is.

# Partial Pooling = Multilevel

- So that’s the beauty of Multilevel:
  - We get information about individual units (like in separate analyses).
  - But, we don’t have “amnesia”.

- And can statistically compare for differences between units.
  - Is job equally influential for everyone?
  - Does gender play a role?

# Residuals

# Longitudinal Multilevel Regression

- Remember that you always need to think about what your model says about the data.

- And a standard multilevel regression is saying something that might be a bit weird with longitudinal data:
  - Assumes that the variance at each measurement occasion is the same.
  - Assumes that the covariance between each successive set of measurements is the same.

# Multilevel Repeated Measures

$$\boldsymbol{\Sigma}(\mathbf{Y}) = \begin{pmatrix} \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_e^2 + \sigma_{u_0}^2 \end{pmatrix}.$$

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}$$

- Related to assumption of Sphericity in RM ANOVA.
- This structure for the covariance matirx is called “Compound Symmetry”.
- If violated → Inflated Type I errors.

- Likely?
- And what do we do if violated?

# Multilevel Repeated Measures

$$\mathbf{\Sigma}(\mathbf{Y}) = \begin{pmatrix} \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_e^2 + \sigma_{u_0}^2 & \cdots & \sigma_{u_0}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_0}^2 & \sigma_{u_0}^2 & \sigma_{u_0}^2 & \cdots & \sigma_e^2 + \sigma_{u_0}^2 \end{pmatrix}.$$

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}$$

- Related to assumption of Sphericity in RM ANOVA.
- This structure for the covariance matirx is called “Compound Symmetry”.
- If violated → Inflated Type I errors.

- Likely?
- And what do we do if violated? → MODEL IT

# Multilevel Repeated Measures

- Use dummies for all measurement occasions
  - Remove Intercept
  - Each dummie has random slope

$$
\begin{aligned}
GPA_{it} &= \pi_{1i}T1_{it} + \pi_{2i}T2_{it} + \pi_{3i}T3_{it} + \pi_{4i}T4_{it} + \pi_{5i}T5_{it} + \pi_{6i}T6_{it} \\
\pi_{1i} &= \beta_{10} + u_{1t} \\
\pi_{2i} &= \beta_{20} + u_{2t} \\
&\vdots \\
\pi_{6i} &= \beta_{60} + u_{6t}
\end{aligned}
$$

- This is basically a MANOVA

# Multilevel Repeated Measures

- Disadvantages:
  - no time variable, so no structure on change
    - But can use contrasts
  - no time varying covariates

# Alternative: Model the residuals!

- If the variance changes/fluctuates over time -> Model the variances a such!!
  - Typical patterns in residual (autoregression) can be included in most packages.
  - Mixed-Effects Location-Scale models: don’t just specify regression equation for predicted mean, but also for predicted variance.

- Easy to do using brms package in R

- If unsure, use an unstructured covariance matrix in which every (co)variance can have a unique value
  - This is a more complicated model and needs more data

# Methodological Considerations: Small Level 2 N

# Fixed Effects Models

- Multilevel is great, but we are modeling on level 2 as well!
- Would you run a regression on N=4?
- Would you calculate a variance on N=2?
- Same with multilevel, if N on level 2 is small (let’s say less than 10), modeling distributions there is probably not a great idea.
- Fortunately, there is a solution! ☺.

# Fixed Effects Models

- Hint, the solution is not cluster robust se’s as is sometimes suggested
    - Guess what, that correction also need to be estimated ;).

- How have you corrected for things in regression in the past?

# Fixed Effects Models

- Hint, the solution is not cluster robust se's as is sometimes suggested
  - Guess what, that correction also need to be estimated ;).

- How have you corrected for things in regression in the past?

- Add dummies! One for each class (and remove the intercept)

$$GPA_{it} = b_{01}Participant1 + b_{02}Participant2 + \ \dots + b_{0x}ParticpantX + b_1 Job_{it} + e_{it}$$

# Fixed Effects Models

$$GPA_{it} = b_{01}Participant1 + b_{02}Participant2 + \ \ldots + b_{0x}ParticpantX + b_1Job_{it} + e_{it}$$

- This is called a Fixed Effects model and is used often in economics.
- It works reaaaaaly well, as the dummies take care of all the level 2 differences.
- Estimates of level 1 predictors unbiased.
- Also deal with “unmodeled” level 2 influences, so is very robust.

# Fixed Effects Models

$$GPA_{it} = b_{01}Participant1 + b_{02}Participant2 + \ \dots + b_{0x}ParticpantX + b_1Job_{it} + e_{it}$$

- But! No free lunch.

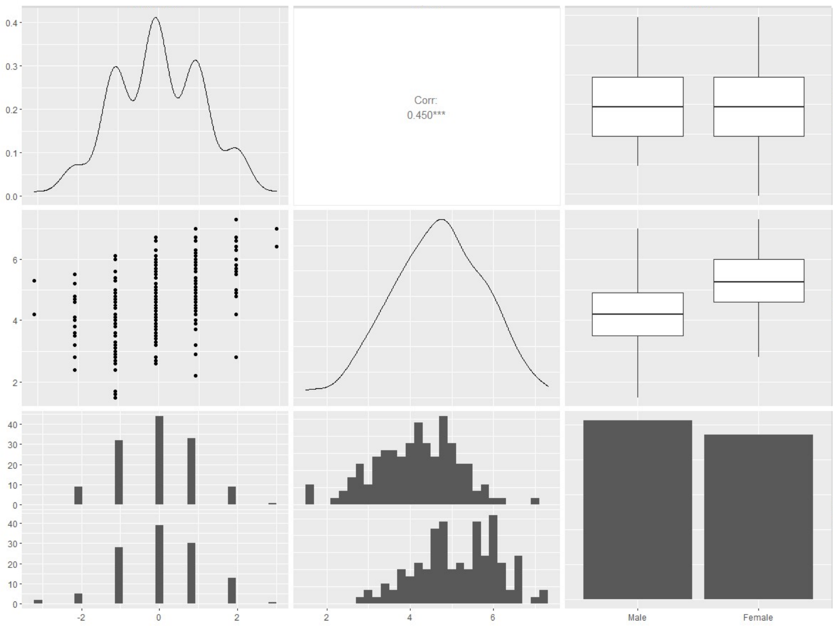
- You can’t model level 2 variables!
  - Since all level 2 variance is in the dummies they are perfectly colinear with level 2 predictors.
  - Not a big problem, if level 2 N is small what do you hope to find there anyway?

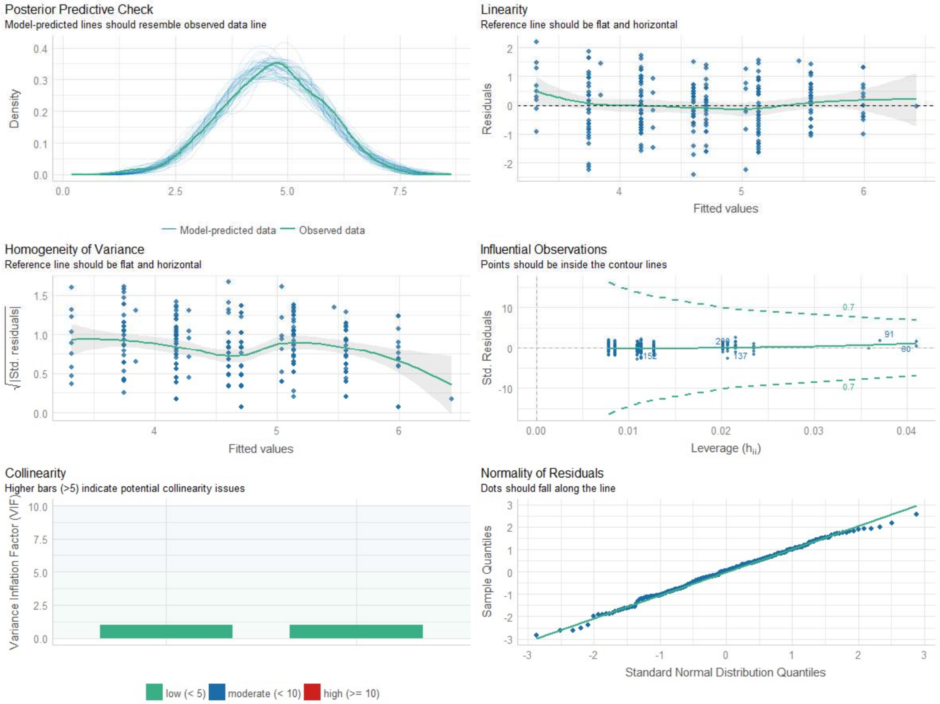
- Can model interactions between level 1 and level 2 predictors though.

# Assumptions & Model Building

# Assumptions & Model Building

- You should always start with visualizing your data!!!





# Assumptions & Model Building

- What we look for with multilevel regression:
  - Normally distributed residuals (QQ-plot or histograms)
  - Linear relation between continuous variables (Scatter plot)

- What we don’t look for:
  - Independence
  - Homogeneity

- Remember, you need to check assumptions for each level separately

# Assumptions & Model Building

- Also two hidden assumptions:
  - There is variance on the higher level (check with ICC and significance of level 2 variance using intercept-only model).
  - Data at each level was obtained by (successive) random sampling!
    - Second assumption often violated
      - When measuring participants at fixed time-point for example
      - Can cause problem with estimate of explained variance (more on this later)

# Maximum Approach

- Often you hear that you should build multilevel models bottom-up

- Go through 5 step:
  1. Run an empty model with no predictors and calculate ICC (intercept-only model)
  2. Run a model with all level 1 predictors (all effects fixed)
  3. Run a model with all level 2 predictors
  4. Add random slopes to your level 1 predictors
  5. If random slopes are present, add cross-level interactions.

# Maximum Approach

- Going through the steps is a good way to see what is going on, but…

- We always know that things will differ between people/units?
  - So why assume fixed effects at first?
  - Why not go for a full model straight away? (Maximum Approach)

- What are possible advantages of staring with a full model?

- And what are disadvantages?

# Maximum Approach

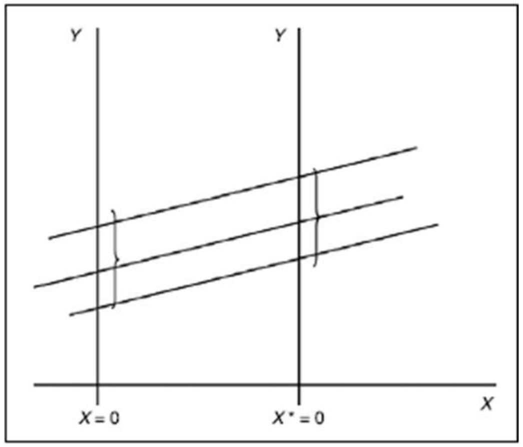


*Figure 4.1* Parallel regression lines, with shift on *X*.

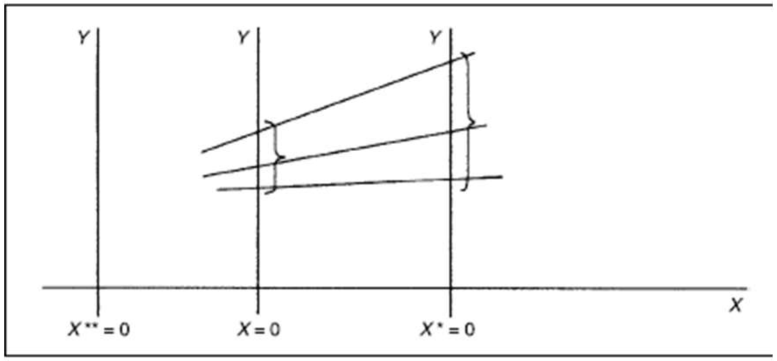
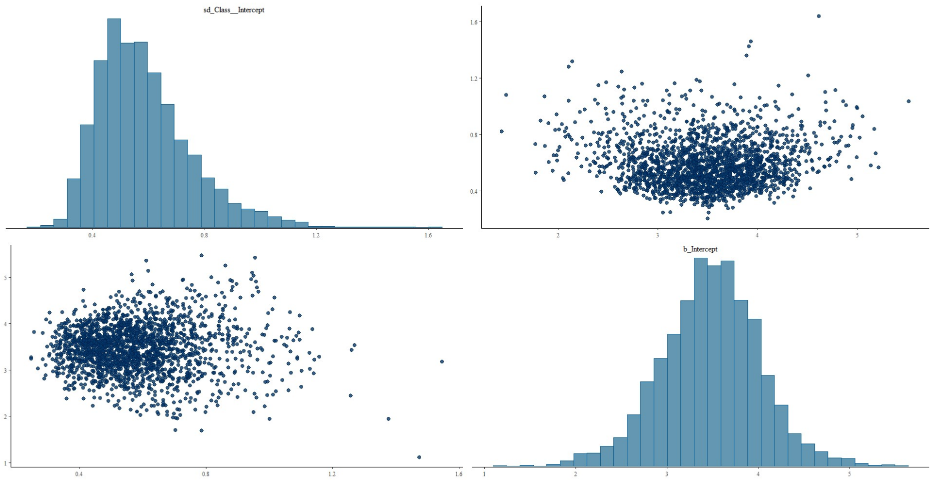
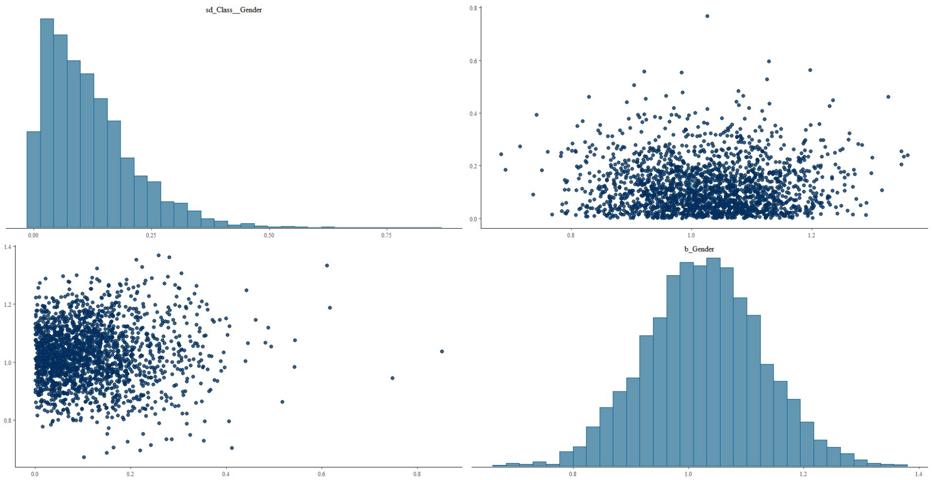
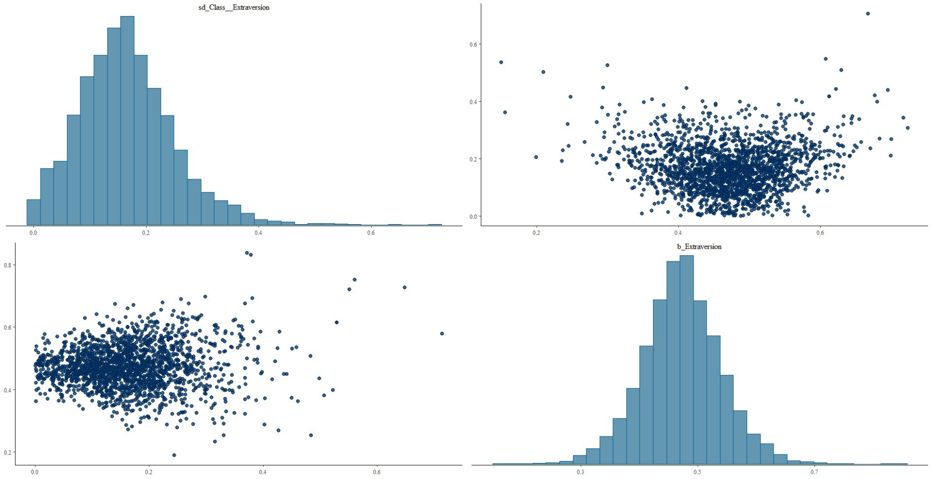


*Figure 4.2* Varying regression lines, with shifts on *X*.

# Maximum Approach







# Assumptions & Model Building

- After fitting the model, you obviously check for significance, but(!!)

- Need to check for relevance too!
  - In multilevel done by looking at explained variance on level 1 and level 2.

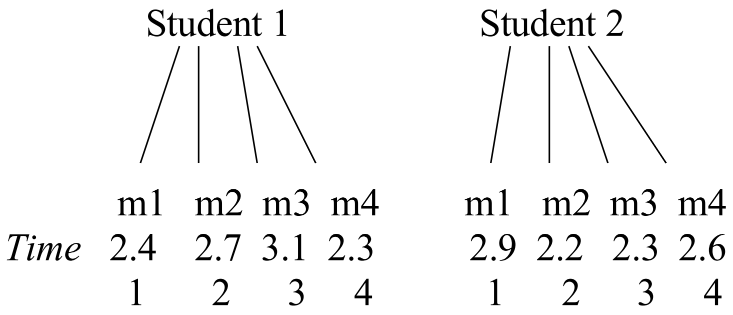
- Usually done by comparing the unexplained variance on level 1 and level 2 for successive models.

- But, when comparing a model with predictors to a model with-out predictors (intercept-only) this approach can lead to a weird result.

# Interpretation and explained variance

- When your sampling is not random, for instance with fixed time-points, there is less between student variance than expected.

- Consequence: the first level variance is overestimated and the second level variance underestimated (because the total is known).



Student 1

m1 m2 m3 m4

*Time* 2.4 2.7 3.1 2.3

1 2 3 4

Student 2

m1 m2 m3 m4

2.9 2.2 2.3 2.6

1 2 3 4