Multilevel Analysis of Group and Longitudinal Data

Dr. Joran Jongerling & Dr. Leonie Vogelsmeier

TILBURG UNIVERSITY

tesc

Understanding Society

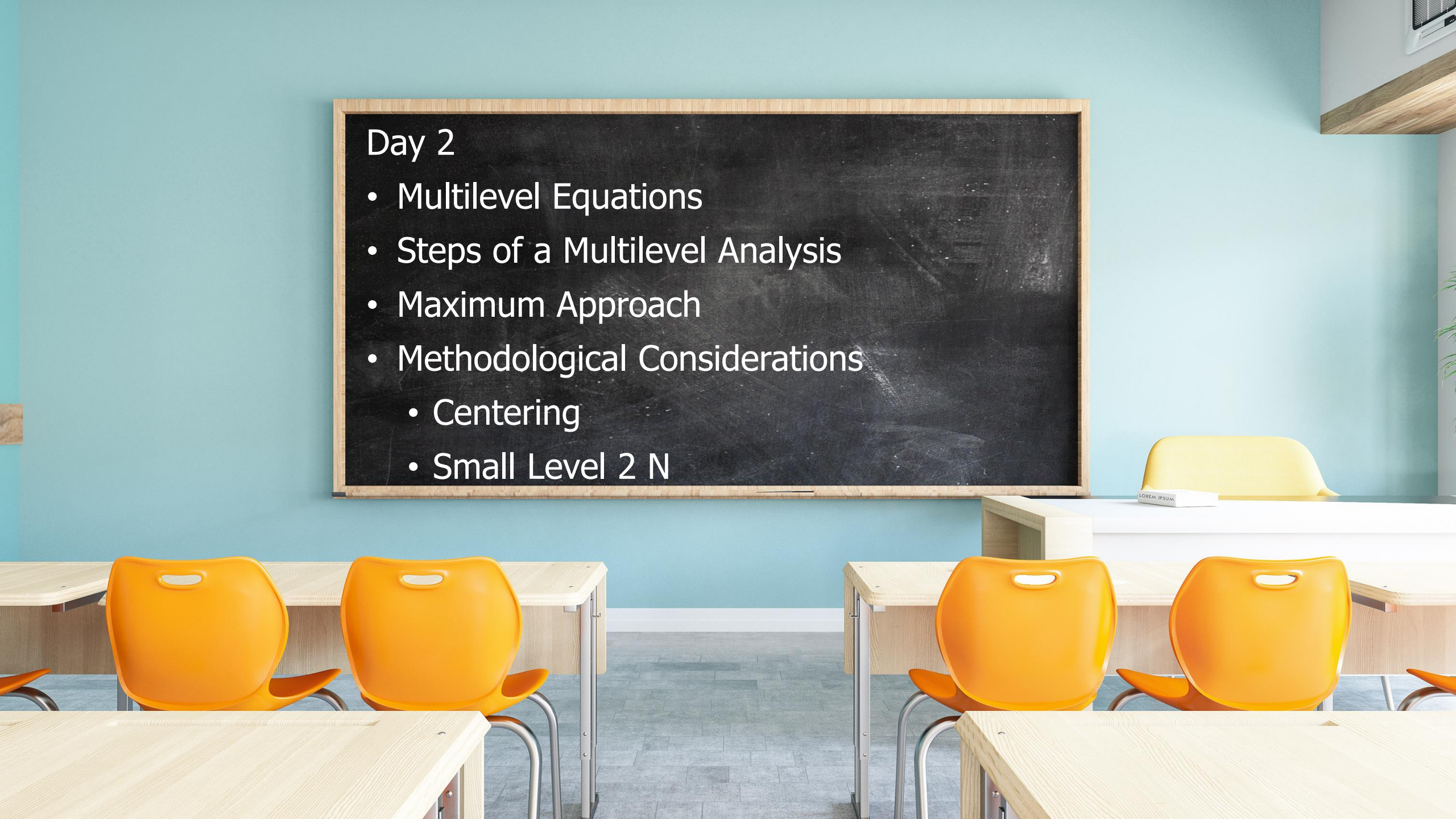
Course Overview

- 3-day course on multilevel analysis
- Video lectures on relevant theory
  - Self-study: either at home or during the (unsupervised) mornings of the course
- Hands-on experience with multilevel analyses in R
  - Supervised practicals in the afternoons of the course
  - Practicals and solutions available on GitHub
- Questions about theory and practicals can be asked during the afternoons

Tilburg University

tesc


Day 1
• Regression Analysis
• Visualization
• Multilevel Data
• Modeling Multilevel Data
LOREM IPSUM


Day 2
• Multilevel Equations
• Steps of a Multilevel Analysis
• Maximum Approach
• Methodological Considerations
• Centering
• Small Level 2 N


Day 3
• Change
• Systematic Mean-Level Change
• Reversible Change

Regression

Regression

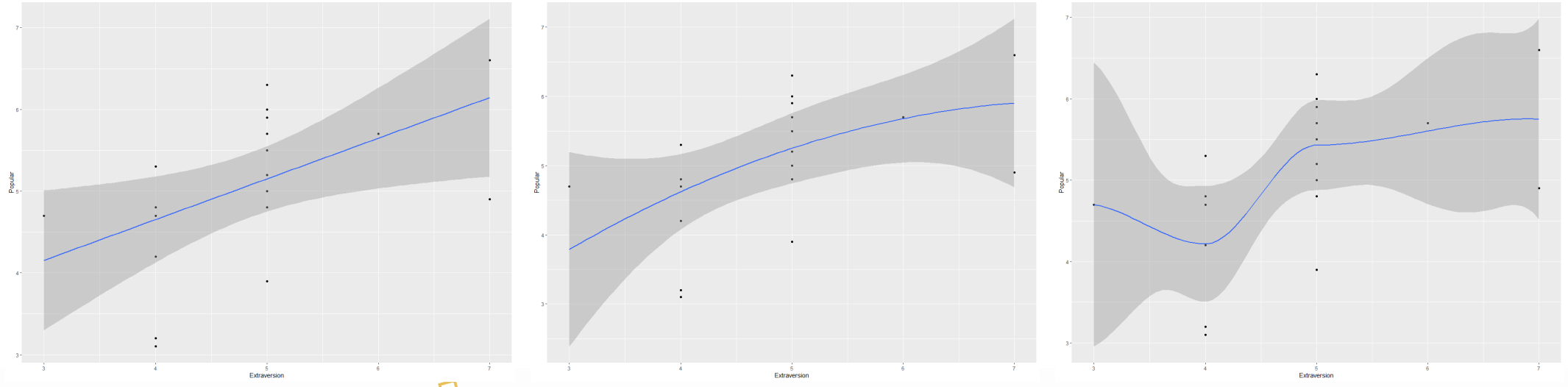
- What can you tell us about multiple regression?
  - That is, what does regression analysis do?

Regression

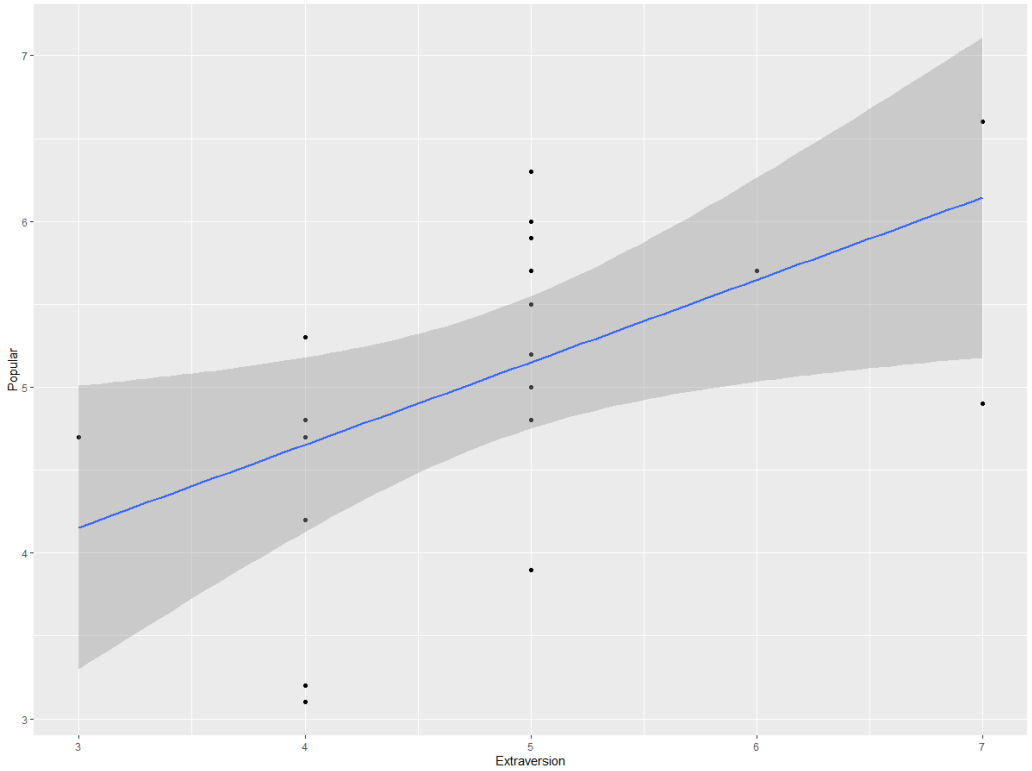
- What can you tell us about multiple regression?
  - That is, what does regression analysis do?

    → A method of curve fitting to explain the relationship between a response variable (Y) and one or more explanatory variables (X).

    → Curves can be straight (linear) or bent (non-linear). It depends on which characteristics you want to capture!



Regression

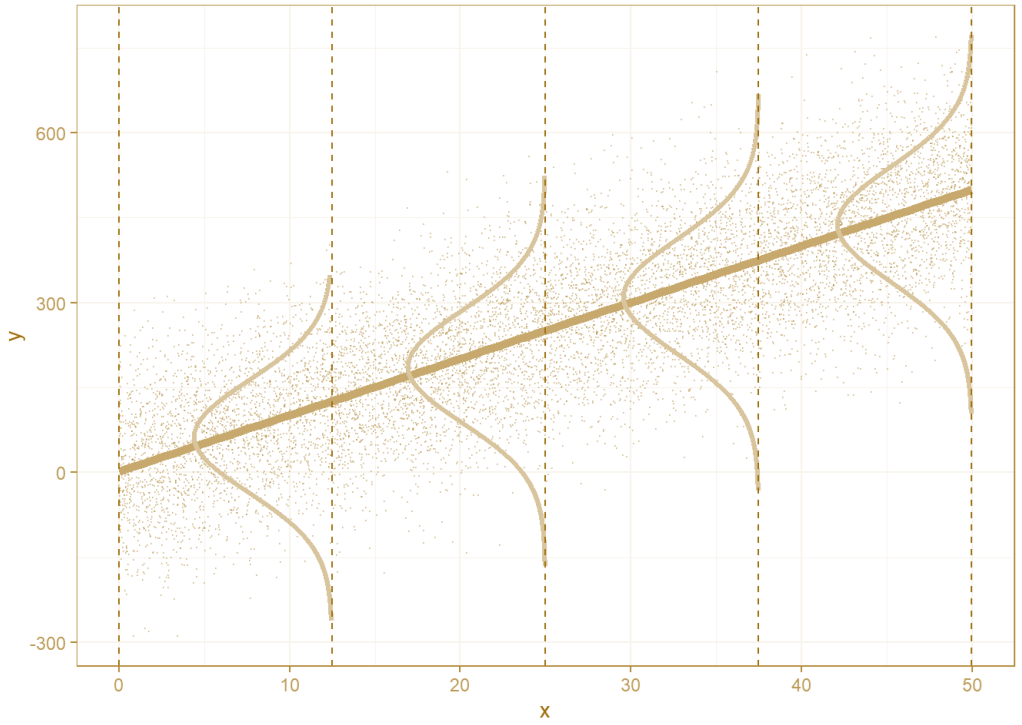


Let's focus on the regular linear model: This comes with assumptions and we need to evaluate whether they make sense!

Tilburg University

tesc

Regression



Assumptions

- **L:** There is a linear relationship between the mean response (Y) and the explanatory variable (X),
- **I:** The errors are independent—there's no connection between how far any two points lie from the regression line,
- **N:** The responses are normally distributed at each level of X, and
- **E:** The error variance or, equivalently, the standard deviation of the responses is equal for all levels of X.

Tilburg University

t e s c

Regression

- Your model is your story about your data. Just adapt your model to the situation at hand!

- If your data is linear, model it as such
  - If it isn't…don't
- If the error variance is the same "across" the board (homoscedasticity), model it as such
  - If it isn't…don't
- If scores are normally distributed around the conditional mean estimate, model it as such
  - If it isn't…don't

Regression

The errors in a linear regression model are assumed to be independent. This means there is no connection between how far any two points lie from the regression line.

Key Violation: In hierarchical data (e.g., students in classrooms, patients in hospitals), observations within the same group (cluster) are often correlated, violating the assumption of independence.

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

→This is where "multilevel regression analysis" comes in!

Regression

- Multilevel analysis (also known as hierarchical modeling) addresses dependent error terms by accounting for the structure of hierarchical data.

- Instead of assuming that all errors are independent, multilevel models allow for random effects at each level of hierarchy (e.g., classroom or hospital).

- Multilevel models enable you to explicitly model the dependence within groups, allowing more accurate inferences and better fit to the data.

Tilburg University

t e s c

Visualization

Visualization

- In the video on regression, we saw that:

"Your model is your story about your data."

- Your model needs to capture all relevant characteristics of your data.

- If a model assumption is violated:
  - The data is not "faulty".
  - Your model just cannot describe your (perfectly fine) data properly.

- So instead of transforming data to force it to fit your model, you can also adjust your model so that it describes the data properly.

Visualization

- To make sure our model describes the data well (i.e., captures all relevant characteristics) we need to know our data!

- The key to this is visualization: You should spend a substantial amount of time just visualizing and looking at your data!
  - And do more than just check scatterplots for linearity and outliers.

Visualization

- Typically, researchers mainly "look at their data" through summary statistics…this is problematic.
- Chosen summaries often only useful with symmetric distributions (e.g., means and SDs).
- Even if info on median, mode, kurtosis, skew, etcetera would be provided, this information is difficult to understand for many.
- Summary statistics are cognitively taxing when comparing multiple groups or multiple levels. E.g., two groups can have:
  - Different means but the same mode,
  - Different modes but the same mean,
  - The same mean and standard deviation but a meaningful skew
  - Etcetera (Heino et al., 2019)

Visualization

- Example (based on Heino et al., 2019)
  - Baseline data from the Let's Move It intervention (Hankonen et al., 2016).
  - Aimed at increasing moderate-to-vigorous-intensity PA (MVPA).
  - Students recruited from four educational tracks:
    1. Practical Nurse (Nur)
    2. Hotel, Restaurant and Catering (HRC)
    3. Business and Administration (BA)
    4. Information and Communications Technology (IT)

Visualization

- Data indicates that girls were about as active as boys at baseline (mean 65 vs. 67 min)

	IT	BA	HRC	Nur	Control	Intervention	Boy	Girl	Full sample
n	163	282	213	402	528	638	471	613	1166
Mean daily accelerometer weartime hours	859.1 (67.9)	848.3 (65.3)	839.4 (68.9)	848.0 (69.0)	834.7 (69.2)	861.0 (67.1)	856.9 (71.5)	842.9 (67.0)	848.6 (69.3)
Mean daily breaks in sitting	21.2 (6.8)	25.9 (7.2)	24.7 (6.9)	28.6 (7.8)	24.8 (7.4)	27.0 (7.8)	23.2 (6.8)	27.9 (7.8)	26.0 (7.7)
Mean daily hours spent sitting or lying down	617.3 (92.1)	533.7 (100.2)	511.0 (103.6)	499.2 (84.9)	519.6 (108.4)	534.8 (97.9)	570.5 (102.8)	499.1 (92.4)	527.7 (103.2)
Mean daily MVPA hours	51.2 (24.8)	66.4 (27.2)	58.0 (27.7)	74.4 (30.7)	63.3 (27.8)	68.0 (31.5)	66.9 (32.3)	64.9 (27.9)	65.8 (29.9)

Tilburg University

tesc

Visualization

- Data indicates that girls were about as active as boys at baseline (mean 65 vs. 67 min))…but are means useful here?

	IT	BA	HRC	Nur	Control	Intervention	Boy	Girl	Full sample
n	163	282	213	402	528	638	471	613	1166
Mean daily accelerometer weartime hours	859.1 (67.9)	848.3 (65.3)	839.4 (68.9)	848.0 (69.0)	834.7 (69.2)	861.0 (67.1)	856.9 (71.5)	842.9 (67.0)	848.6 (69.3)
Mean daily breaks in sitting	21.2 (6.8)	25.9 (7.2)	24.7 (6.9)	28.6 (7.8)	24.8 (7.4)	27.0 (7.8)	23.2 (6.8)	27.9 (7.8)	26.0 (7.7)
Mean daily hours spent sitting or lying down	617.3 (92.1)	533.7 (100.2)	511.0 (103.6)	499.2 (84.9)	519.6 (108.4)	534.8 (97.9)	570.5 (102.8)	499.1 (92.4)	527.7 (103.2)
Mean daily MVPA hours	51.2 (24.8)	66.4 (27.2)	58.0 (27.7)	74.4 (30.7)	63.3 (27.8)	68.0 (31.5)	66.9 (32.3)	64.9 (27.9)	65.8 (29.9)

TILBURG UNIVERSITY

t e s c

Visualization

```
# Generate Data Based on Descriptives
IT <- rnorm(1166, 51.2, 24.8)
BA <- rnorm(1166, 66.4, 27.2)
HRC <- rnorm(1166, 58, 27.7)
NUR <- rnorm(1166, 74.4, 30.7)

# Calculate Average Activity Across Educational Tracks
df <- as.data.frame(cbind(IT, BA, HRC, NUR))
MVPA_Full_Sample <- rowMeans(df)

# Calculate Mean and SD for MVPA
mean(MVPA_Full_Sample)
[1] 63.07823

sd(MVPA_Full_Sample)
[1] 14.26248
```

	IT	BA	HRC	Nur	Control	Intervention	Boy	Girl	Full sample
n	163	282	213	402	528	638	471	613	1166
Mean daily accelerometer weartime hours	859.1 (67.9)	848.3 (65.3)	839.4 (68.9)	848.0 (69.0)	834.7 (69.2)	861.0 (67.1)	856.9 (71.5)	842.9 (67.0)	848.6 (69.3)
Mean daily breaks in sitting	21.2 (6.8)	25.9 (7.2)	24.7 (6.9)	28.6 (7.8)	24.8 (7.4)	27.0 (7.8)	23.2 (6.8)	27.9 (7.8)	26.0 (7.7)
Mean daily hours spent sitting or lying down	617.3 (92.1)	533.7 (100.2)	511.0 (103.6)	499.2 (84.9)	519.6 (108.4)	534.8 (97.9)	570.5 (102.8)	499.1 (92.4)	527.7 (103.2)
Mean daily MVPA hours	51.2 (24.8)	66.4 (27.2)	58.0 (27.7)	74.4 (30.7)	63.3 (27.8)	68.0 (31.5)	66.9 (32.3)	64.9 (27.9)	65.8 (29.9)

Tilburg University

tesc

Visualization

```
# Generate Data Based on Descriptives
IT <- rnorm(1166, 51.2, 24.8)
BA <- rnorm(1166, 66.4, 27.2)
HRC <- rnorm(1166, 58, 27.7)
NUR <- rnorm(1166, 74.4, 30.7)

# Calculate Average Activity Across Educational Tracks
df <- as.data.frame(cbind(IT, BA, HRC, NUR))
MVPA_Full_Sample <- rowMeans(df)

# Calculate Mean and SD for MVPA
mean(MVPA_Full_Sample)
[1] 63.07823

sd(MVPA_Full_Sample)
[1] 14.26248
```

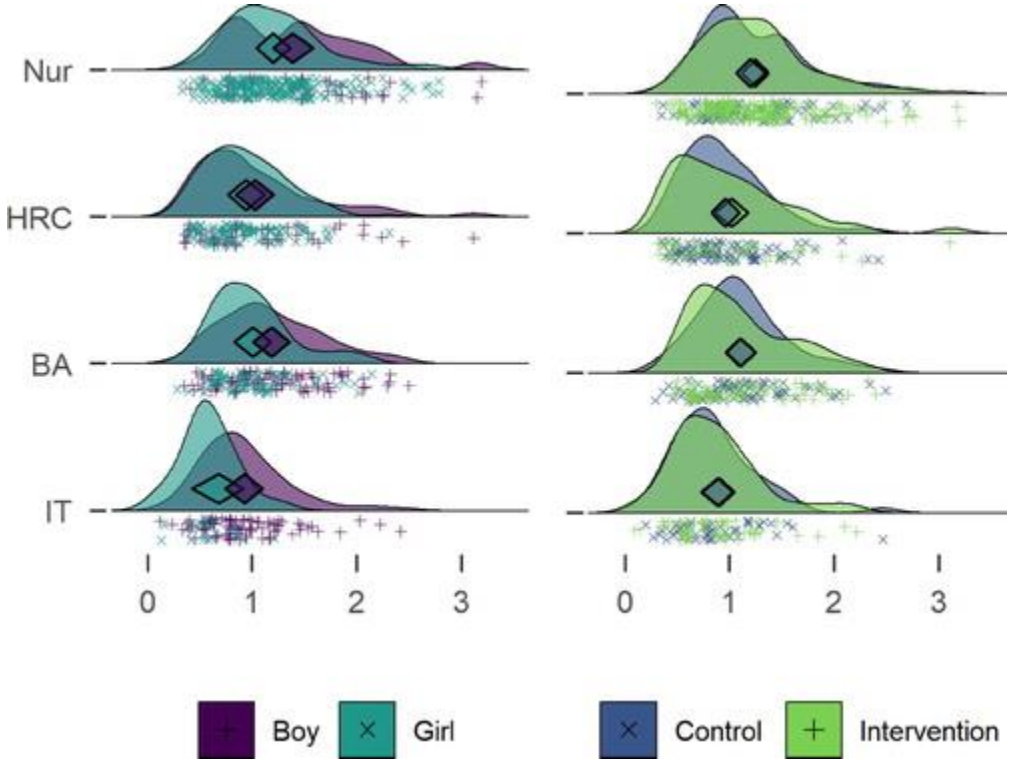
	IT	BA	HRC	Nur	Control	Intervention	Boy	Girl	Full sample
n	163	282	213	402	528	638	471	613	1166
Mean daily accelerometer weartime hours	859.1 (67.9)	848.3 (65.3)	839.4 (68.9)	848.0 (69.0)	834.7 (69.2)	861.0 (67.1)	856.9 (71.5)	842.9 (67.0)	848.6 (69.3)
Mean daily breaks in sitting	21.2 (6.8)	25.9 (7.2)	24.7 (6.9)	28.6 (7.8)	24.8 (7.4)	27.0 (7.8)	23.2 (6.8)	27.9 (7.8)	26.0 (7.7)
Mean daily hours spent sitting or lying down	617.3 (92.1)	533.7 (100.2)	511.0 (103.6)	499.2 (84.9)	519.6 (108.4)	534.8 (97.9)	570.5 (102.8)	499.1 (92.4)	527.7 (103.2)
Mean daily MVPA hours	51.2 (24.8)	66.4 (27.2)	58.0 (27.7)	74.4 (30.7)	63.3 (27.8)	68.0 (31.5)	66.9 (32.3)	64.9 (27.9)	65.8 (29.9)

- Mean and especially SD don't match! Data is not symmetrical, but this was impossible to see from the table.
- Also, what if we want to see a pattern across the tracks? (e.g., do boys always score higher?)
- This can't be easily done with means and SD anymore…we need to visualize.

TILBURG UNIVERSITY

t e s c

Visualization



TILBURG UNIVERSITY

tesc

Visualization

- Data visualizations are crucial supplements to large numerical tables of descriptive statistics (Tay, Parrigon, Huang, & LeBreton, 2016).
- They are more straightforward ways to provide lots of important information (including uncertainty).
  - Require much less statistical/mathematical knowledge from the reader.
- Also, providing extensive visualizations (including the raw data) will aid open-science and replication/reproducibility.

Visualization

- Study on the popularity of high school students.
  - Total of 246 students from 12 different classes.
  - Determined how extraversion, gender, and teacher experience influenced a student's popularity.

- List of all the variables:
  - pupil: pupil identification variable, not needed in the analysis
  - class: class identification variable, the linking variable to define the 2-level structure
  - student-level independent variables: extraversion (continuous; higher scores mean higher extraversion) and gender (dichotomous; 0=male, 1 =female)
  - class-level independent variables: teacher experience (in years)
  - outcome variable: popular (continuous outcome variable at the student level, higher scores indicate higher popularity)

Visualization

- Study on the popularity of highschool students.
  - Total of 246 students from 12 different classes.
  - Determined how extraversion, gender, and teacher experience influenced a student's popularity.

- List of all the variables:
  - pupil: pupil identification variable, not needed in the analysis
  - class: class identification variable, the linking variable to define the 2 - level structure
  - student-level independent variables: extraversion (continuous; higher scores mean higher extraversion) and gender (dichotomous; 0=male, 1 =female)
  - class-level independent variables: teacher experience (in years)
  - outcome variable: popular (continuous outcome variable at the student-level, higher scores indicate higher popularity)

TILBURG UNIVERSITY

t e s c

Visualization

```
# Visualize data: Show first lines of code and get descriptive statistics for
# all variables
head(Total)
describe(Total, fast=TRUE)
```

```
  Pupil Class Extraversion Gender teacherExp Popular
1     1     1  -0.07723577      1         24     6.3
2     2     1   1.92276423      0         24     4.9
3     3     1  -1.07723577      1         24     5.3
4     4     1  -2.07723577      1         24     4.7
5     5     1  -0.07723577      1         24     6.0
6     6     1  -1.07723577      0         24     4.7
```

```
             vars   n  mean   sd median   min   max range  skew kurtosis   se
Pupil           1 246 10.83 6.07  11.00  1.00 24.00  23.0  0.08    -1.11 0.39
Class           2 246  6.52 3.41   7.00  1.00 12.00  11.0 -0.03    -1.23 0.22
Extraversion    3 246  0.00 1.11  -0.08 -3.08  2.92   6.0 -0.01    -0.28 0.07
Gender          4 246  0.48 0.50   0.00  0.00  1.00   1.0  0.08    -2.00 0.03
teacherExp      5 246 14.28 5.52  14.00  5.00 24.00  19.0  0.11    -0.99 0.35
Popular         6 246  4.67 1.11   4.70  1.50  7.30   5.8 -0.16    -0.23 0.07
```

TILBURG UNIVERSITY

t e s c

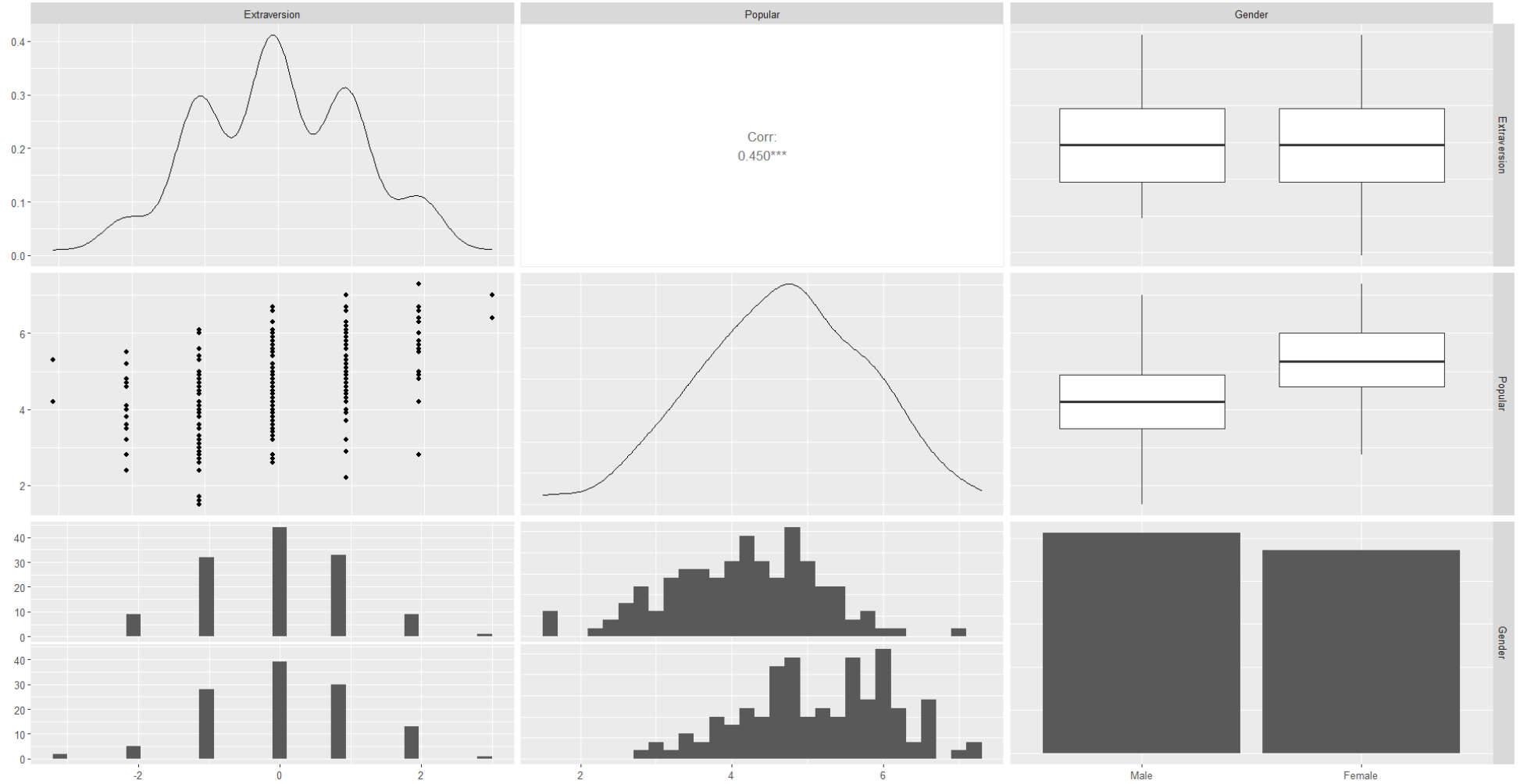
Visualization

```
# Check assumptions for student level variables: Only look at the actual
# variables, not the ID variables Class and Student

subset <- Total %>%
          select(Extraversion, Popular, Gender)


# Visualize (combination of) variables
ggpairs(subset)
```

Visualization



TILBURG UNIVERSITY

tesc

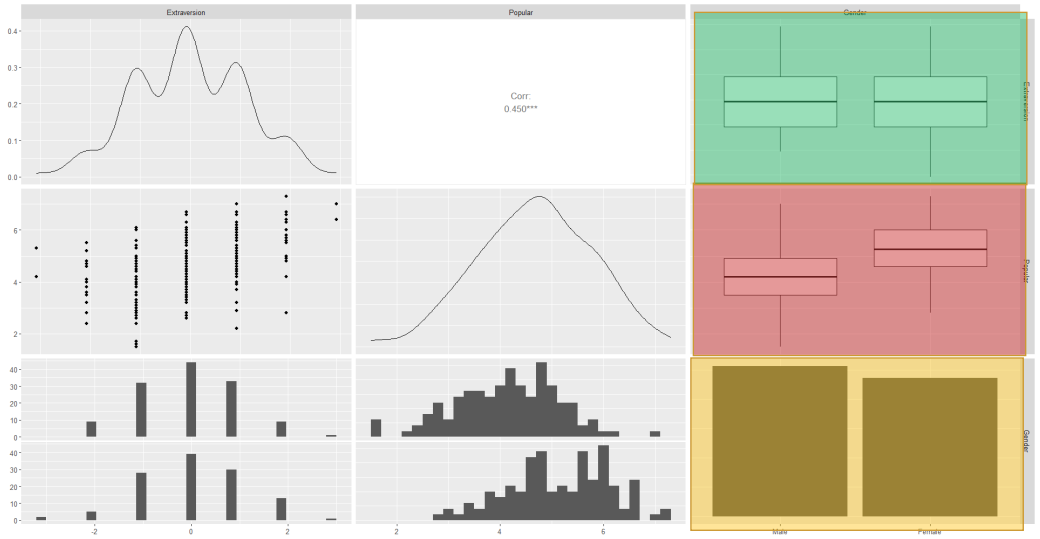
Visualization

- Popularity looks relatively normal overall.
  - But scores slightly skewed within genders

TILBURG UNIVERSITY

tesc

Visualization

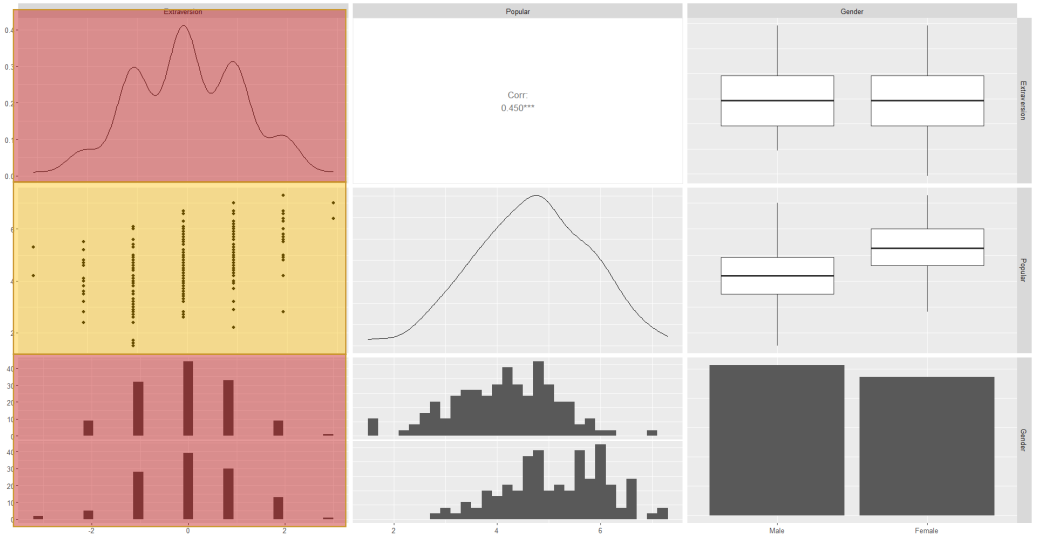


- Popularity looks relatively normal overall.
  - But scores slightly skewed within genders
- Proportion of boys and girls relatively equal.
  - No apparent Gender difference in extraversion
  - Girls appear slightly more popular.

TILBURG UNIVERSITY

tesc

Visualization



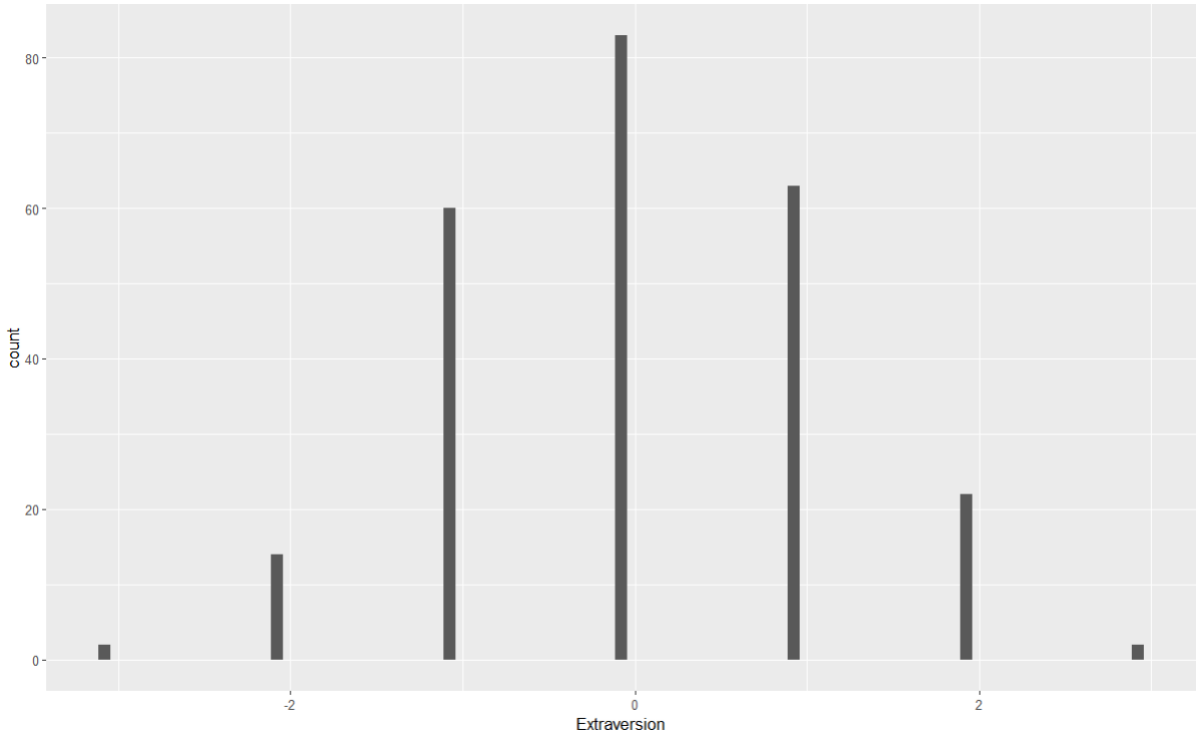
- Popularity looks relatively normal overall.
  - But scores slightly skewed within genders.
- Proportion of boys and girls relatively equal.
  - No apparent Gender difference in extraversion
  - Girls appear slightly more popular.
- Relationship between Extraverison and Popular looks linear
  - But distribution of Extraversion looks odd…more ordinal than continuous.

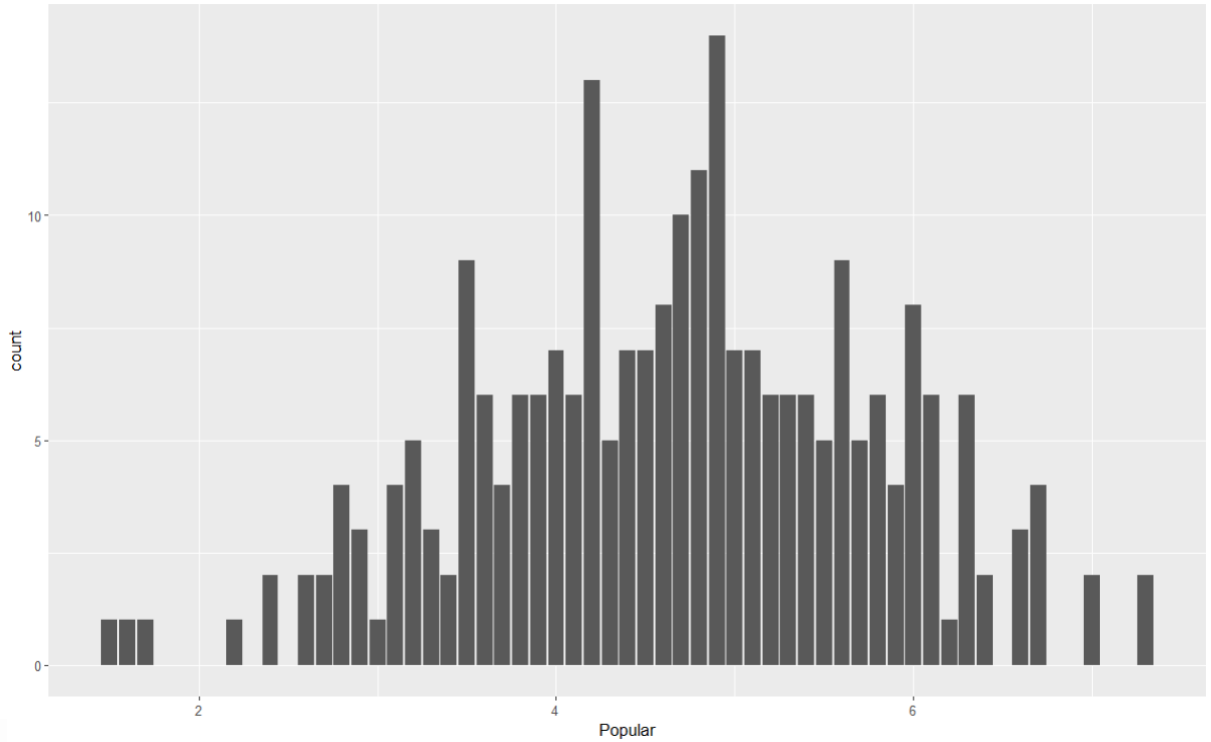
TILBURG UNIVERSITY

tesc

Visualization

```
# Scatterplots for Popularity and Extraversion
ggplot(Total, aes(x = Extraversion)) + geom_bar()
ggplot(Total, aes(x = Popular)) + geom_bar()
```





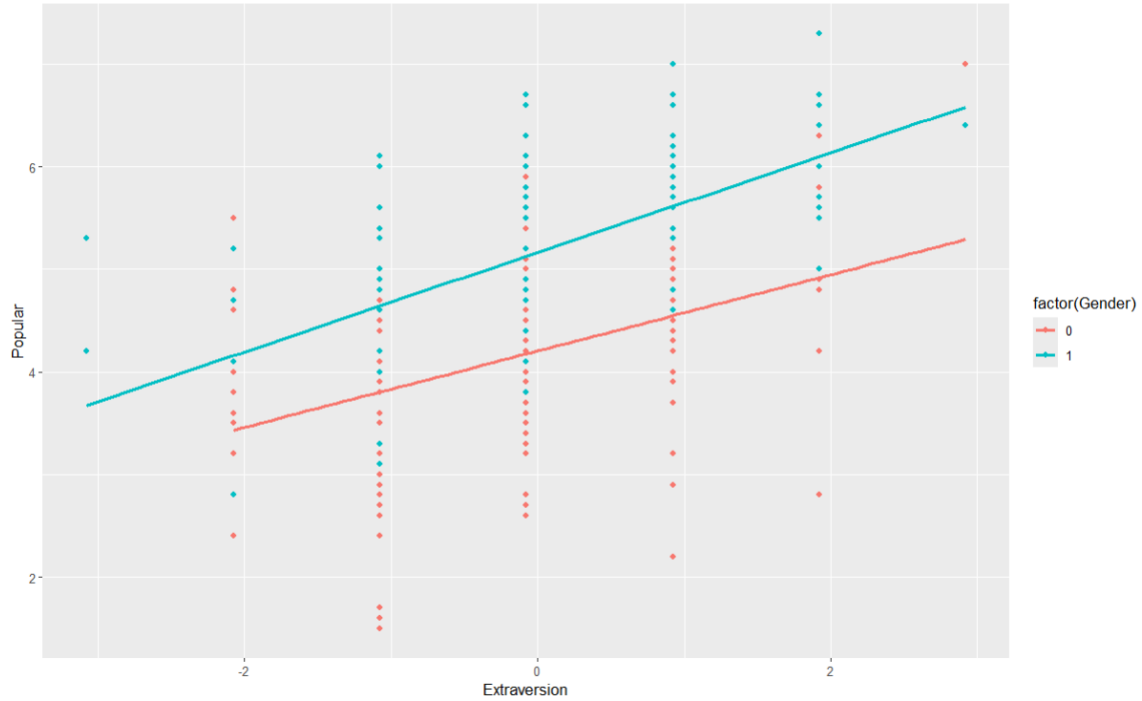
TILBURG UNIVERSITY

tesc

Visualization

```
# Also check for possible interactions, using the "fill" option from GGplot
c <- ggplot(Total, aes(x = Extraversion, y = Popular, col=factor(Gender)))
     + geom_point()

c + geom_smooth(method='lm', se=FALSE)
```



Visualization

- Conclusions:
  - Extraversion might be more ordinal than continuous (how we model depends on theory and measurement).
  - Proportion of boys and girls pretty similar (important for interpretation of effects and interactions)
  - No clear indications of interaction effect.

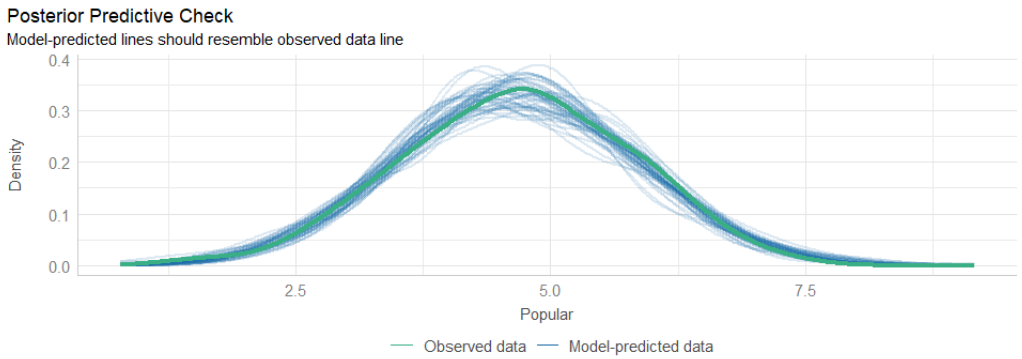
- These are all the checks of the data that we need to (and can do) before fitting the model.
  - Fitting a linear model seems ok (for now).
  - Let's see if it indeed fits the data well.

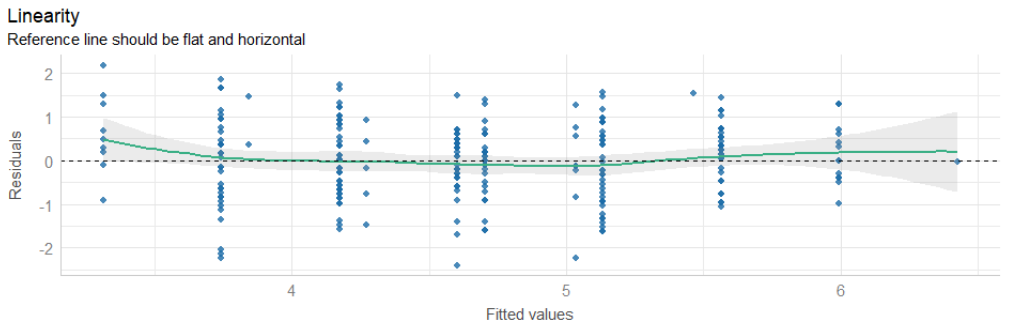
Visualization

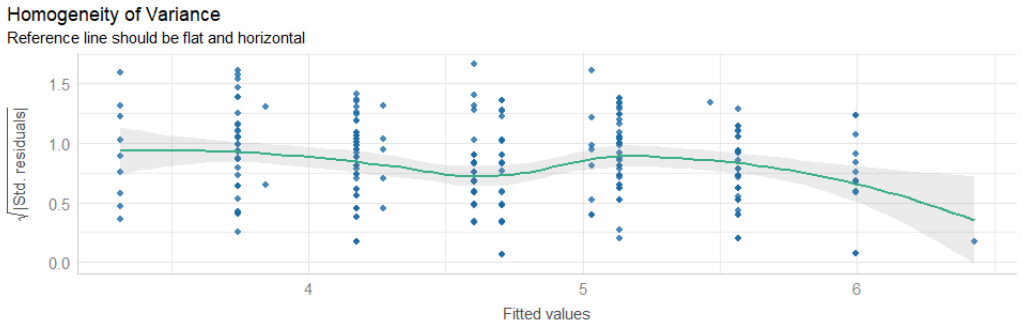
```
# Run model checks using the check_model command
Regr1 <- lm(Popular ~ 1 + Extraversion + Gender, data = Total)
check_model(Regr1)
```

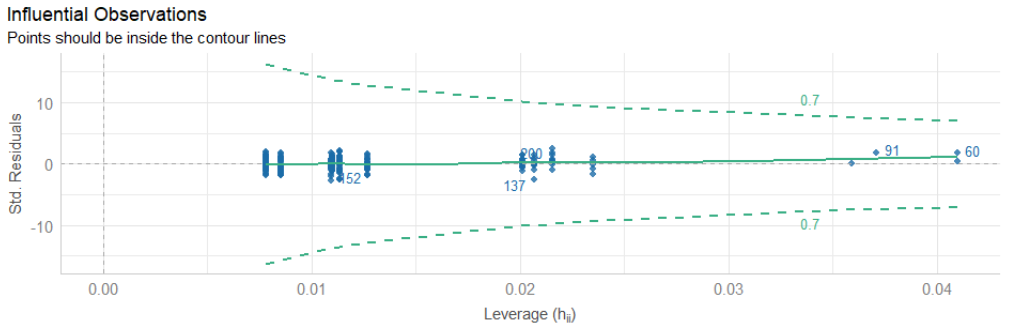
$$Popular = b_{intercept} + b_{ext}Extraversion + b_{gen}Gender + e$$

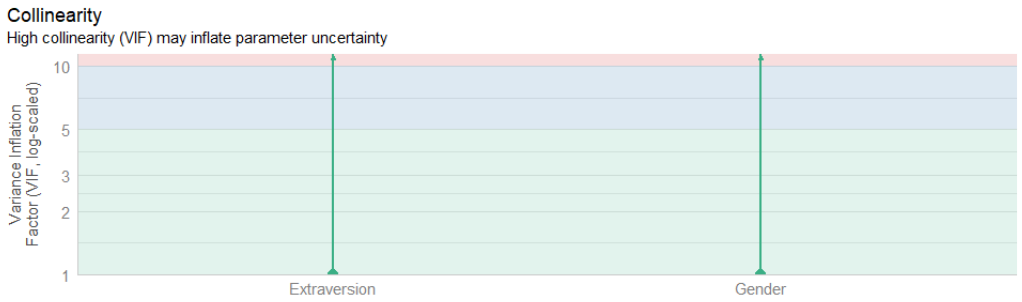
Visualization

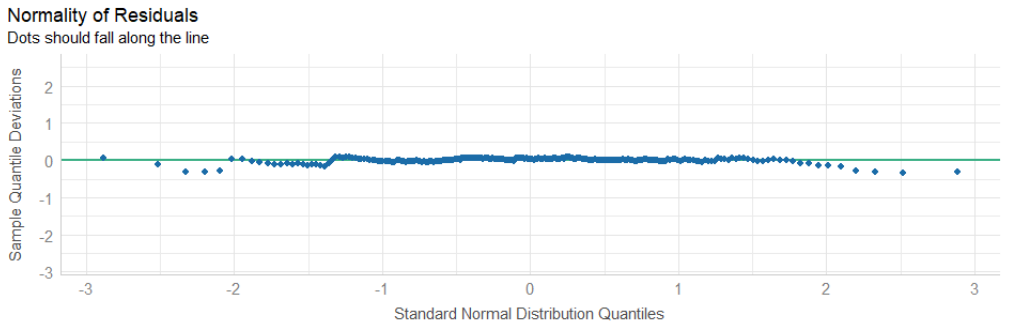








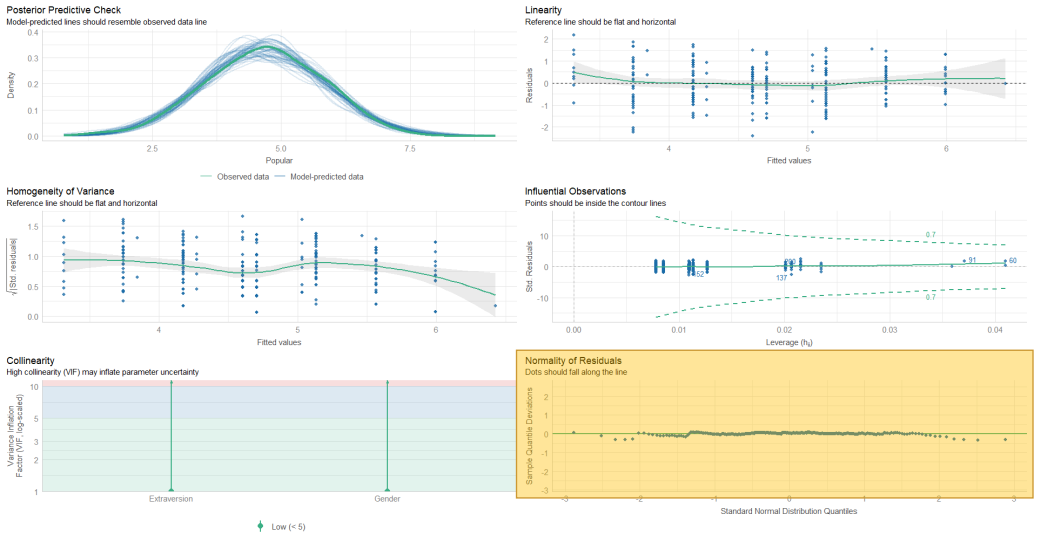




TILBURG UNIVERSITY

tesc

Visualization

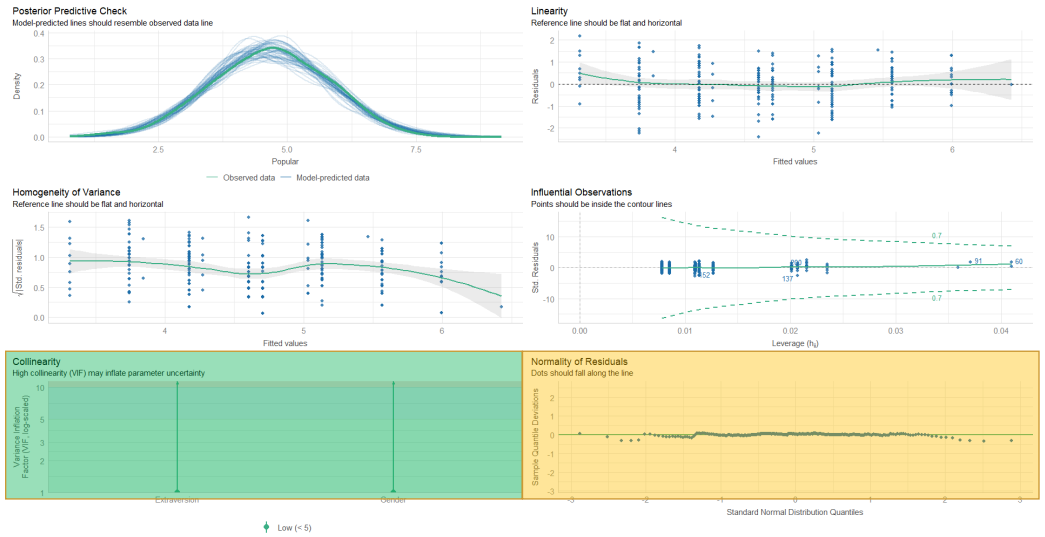


- Conclusions: Let's start with what we already know from exploring the data.
  - Normality is fine
  - No strong relation between predictors
  - Linearity is fine.

TILBURG UNIVERSITY

tesc

Visualization

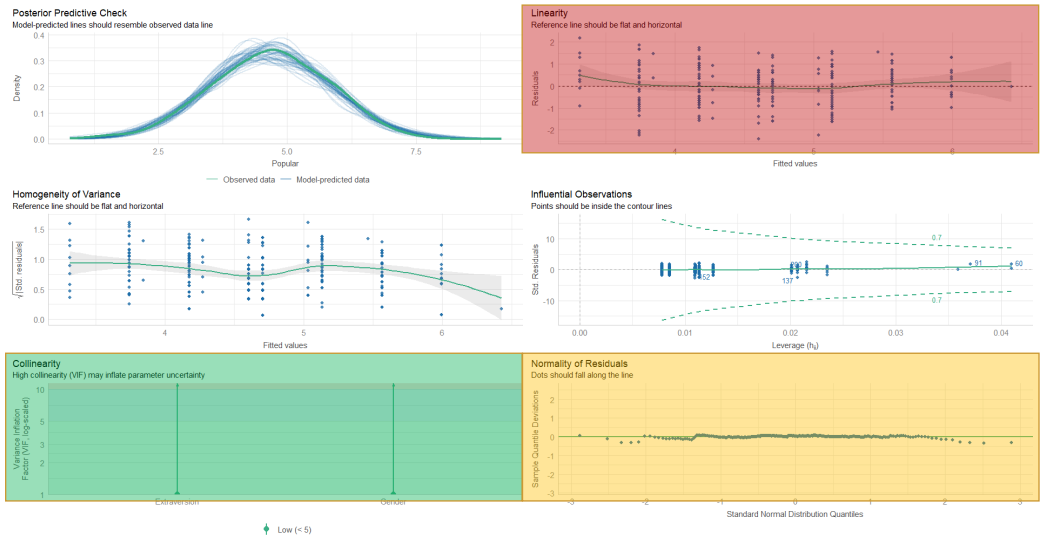


- Conclusions: Let's start with what we already know from exploring the data.
  - Normality is fine
  - No strong relation between predictors
  - Linearity is fine.

TILBURG UNIVERSITY

tesc

Visualization

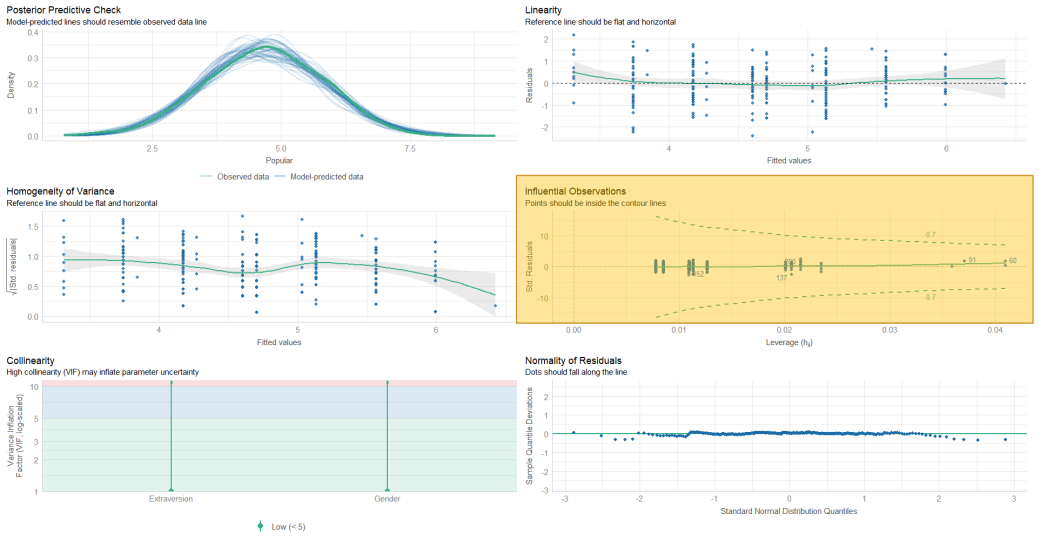


- Conclusions: Let's start with what we already know from exploring the data.
  - Normality is fine
  - No strong relation between predictors
  - Linearity is fine.

TILBURG UNIVERSITY

tesc

Visualization

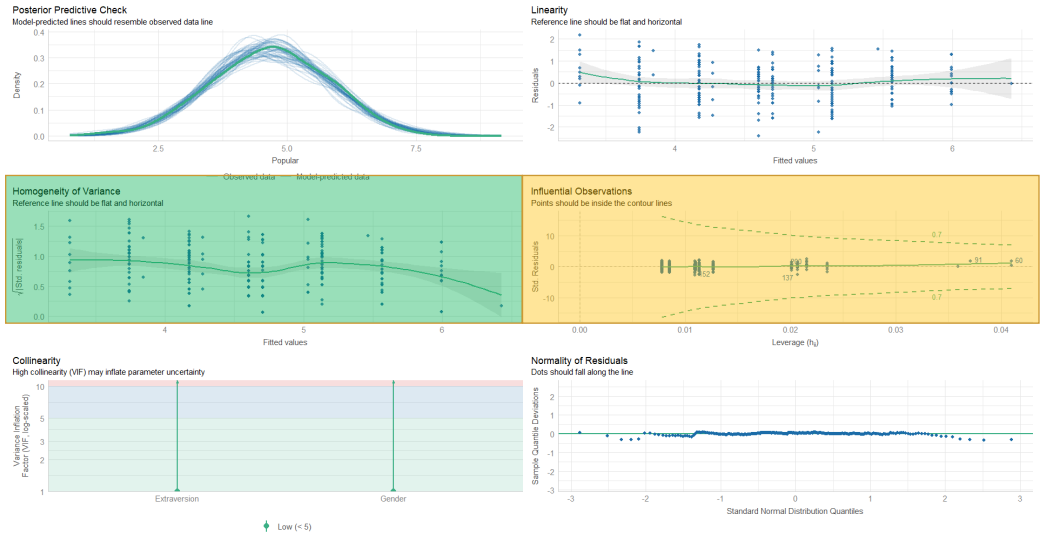


- New Conclusions:
  - No extreme (combinations of) scores.

TILBURG UNIVERSITY

tesc

Visualization

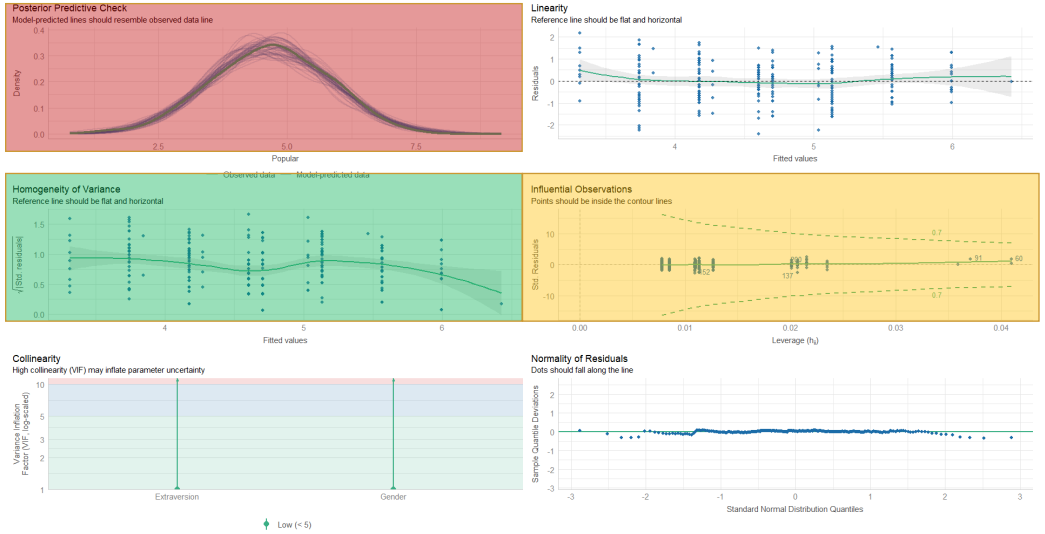


- New Conclusions:
  - No extreme (combinations of) scores.
  - Heterogeneity not ideal.

TILBURG UNIVERSITY

tesc

Visualization

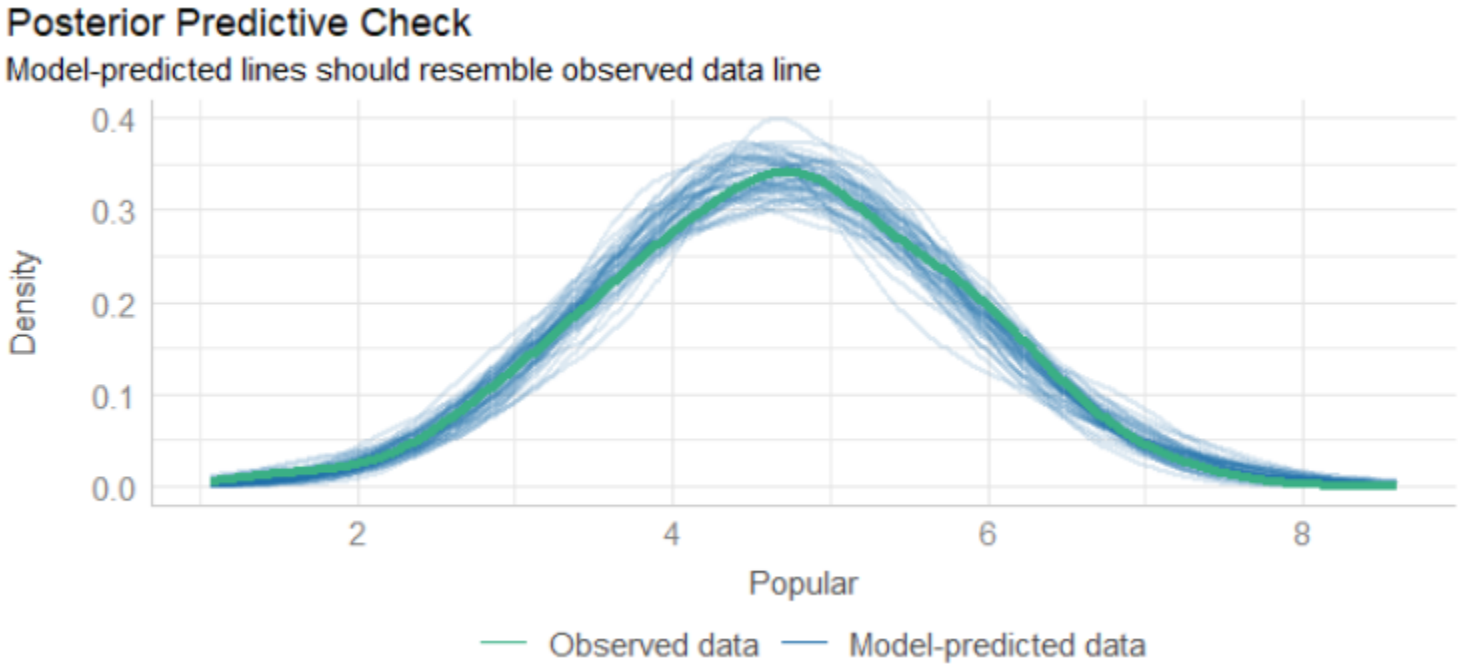


- New Conclusions:
  - No extreme (combinations of) scores.
  - Heterogeneity not ideal.
  - And…most importantly…our model seems to capture relevant aspects of our data!

TILBURG UNIVERSITY

tesc

Visualization



- This plot shows hypothetical data generated from your model (the blue curves) and your real data (the green curve)
- If the blue lines envelope the green on and look similar in shape, your model captures all important aspects of your data.

Tilburg University

t e s c

Visualization

- Final conclusions:
  - Heterogeneity is not ideal. We'll leave it for now but you could consider robust SEs or including heterogeneity in your model.
  - Extraversion seems more ordinal than continuous…but:
    - Relatively normally distributed (and more than 6 categories).
    - And inspection of model fit does not show any issues, so probably ok.

Visualization

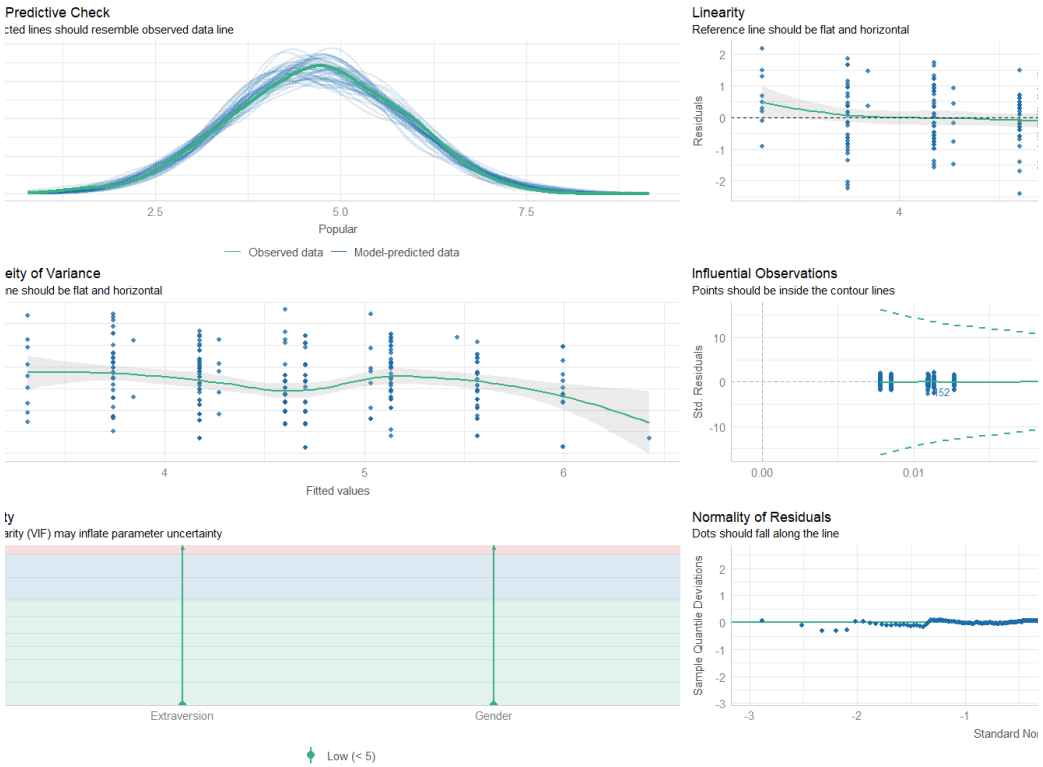
- Final conclusions:
  - Heterogeneity is not ideal. We'll leave it for now but you could consider robust SEs or including heterogeneity in your model.
  - Extraversion seems more ordinal than continuous…but:
    - Relatively normally distributed (and more than 6 categories).
    - And inspection of model fit does not show any issues, so probably ok…(but you can always compare both options!)

Visualization

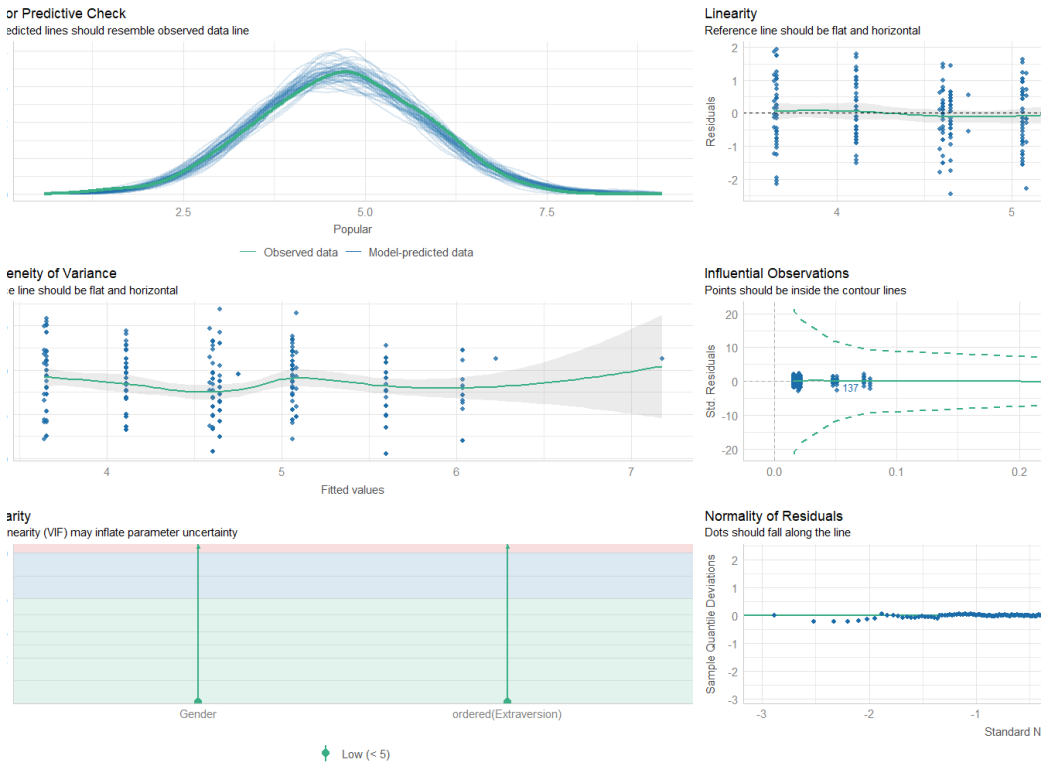
```r
# Extra: also check model assumptions when extraversion is entered as
# an ordinal variable
Regr1b <- lm(Popular ~ 1 + ordered(Extraversion) + Gender, data = Total)
check_model(Regr1b)
```

Visualization

Continuous



Ordinal



TILBURG UNIVERSITY

tesc

Visualization

- There are more formal tests to compare the two models, but visualizations shows that the ordinal options is:
  - Slightly better, but
  - No big difference.

Multilevel Data

Multilevel Data

- In previous videos we've looked at:
  - Regression.
  - How your model is "your story about the data".
  - How visualization can help make sure your model is capturing all relevant aspects of your data.
- Now we're going to apply that to so called "nested data"
  - Data in which (some) observations are grouped together.
- We'll us the same data we used in the video on visualization.

Multilevel Data

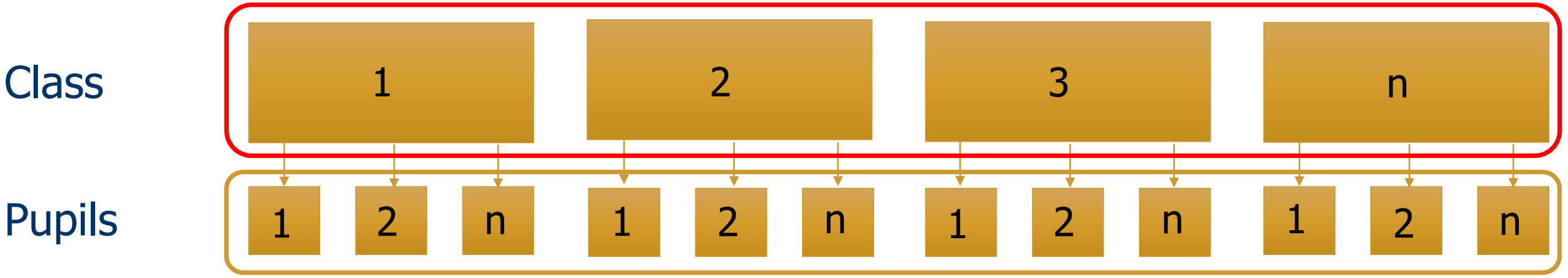
- Study on the popularity of high school students.
  - Total of 246 students from 12 different classes.
  - Determined how extraversion, gender, and teacher experience influenced a student's popularity.

- List of all the variables:
  - pupil: pupil identification variable, not needed in the analysis
  - class: class identification variable, the linking variable to define the 2 - level structure
  - student-level independent variables: extraversion (continuous; higher scores mean higher extraversion) and gender (dichotomous; 0=male, 1 =female)
  - class-level independent variables: teacher experience (in years)
  - outcome variable: popular (continuous outcome variable at the student-level, higher scores indicate higher popularity)

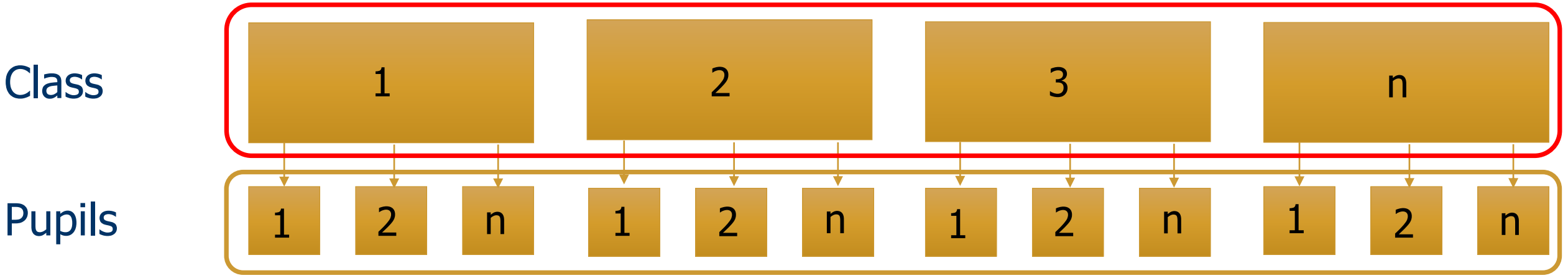
TILBURG UNIVERSITY

t e s c

Multilevel Data

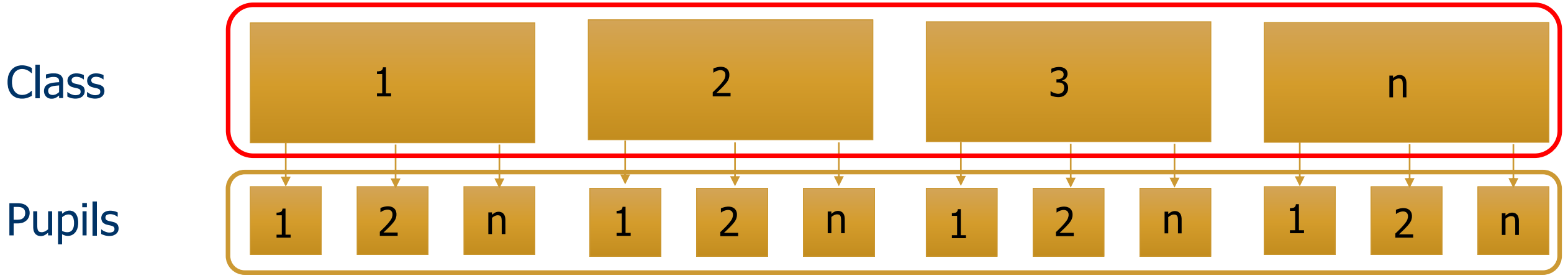


Multilevel Data



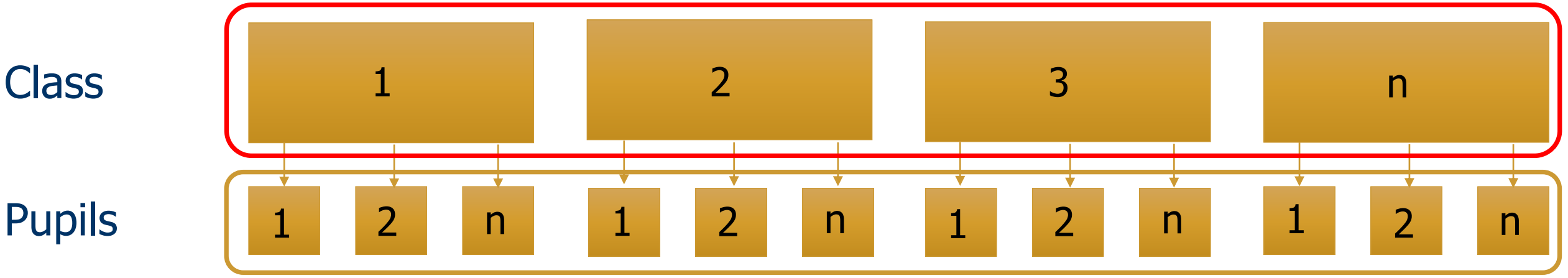
- What do we need to keep in mind for our model based on all this?

Multilevel Data



- Students from the same class not completely independent from each other.
  - Likely more alike in terms of SES and education level of their parents.
  - Influenced by the same (common) class factors
  - Etc.

Multilevel Data



- Scores within classes are not independent
  - Two students from the same class are more alike than two students from two different classes.

disaggvoorbeeld.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 8 of 8 Variables

	class	stud...	sex	SES	behav...	IQ	SA	groupsize	var	var	var	var	var
16	1	18	1	6	5	112	94	30					
17	1	19	1	2	3	91	91	30					
18	1	20	1	2	1	120	98	30					
19	1	21	2	2	5	117	109	30					
20	1	22	2	4	5	103	107	30					
21	1	23	1	4	3	119	111	30					
22	1	24	1	3	3	105	102	30					
23	1	25	2	2	4	97	112	30					
24	1	26	2	6	5	111	112	30					
25	1	27	1	6	4	117	120	30					
26	1	28	2	5	5	102	109	30					
27	1	29	1	6	4	101	96	30					
28	1	30	2	5	3	98	106	30					
29	2	31	1	5	4	70	106	24					
30	2	32	1	2	4	97	110	24					
31	2	33	1	2	2	103	89	24					
32	2	34	2	3	4	112	102	24					
33	2	35	1	2	3	95	103	24					
34	2	36	2	5	4	89	109	24					
35	2	37	1	2	4	100	106	24					
36	2	38	2	6	5	108	93	24					
37	2	39	1	4	4	102	94	24					
38	2	40	2	5	3	93	92	24					
39	2	41	1	1	1	76	99	24					
40	2	42	1	5	4	101	107	24					
41	2	43	1	2	2	103	108	24					
42	2	44	1	2	3	101	97	24					
43	2	45	2	4	4	116	96	24					
44	2	46	1	4	2	113	96	24					
45	2	47	1	5	4	116	86	24					

Data View Variable View

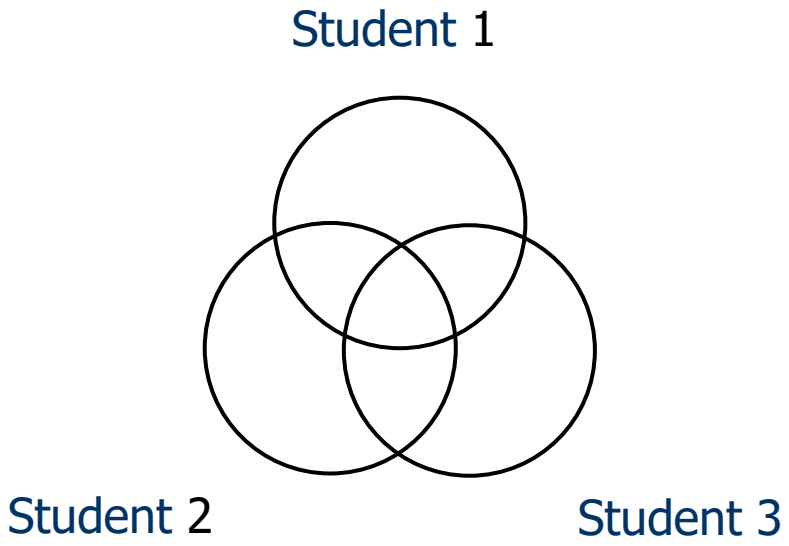
IBM SPSS Statistics Processor is ready Unicode:ON

TILBURG UNIVERSITY

Multilevel Data

- Ok, so data not completely independent.
- Independence is an assumption of most analyses, including regression, but so what?
- Helps to think in terms of information.
  - How much unique information does each student add to our sample?

Multilevel Data

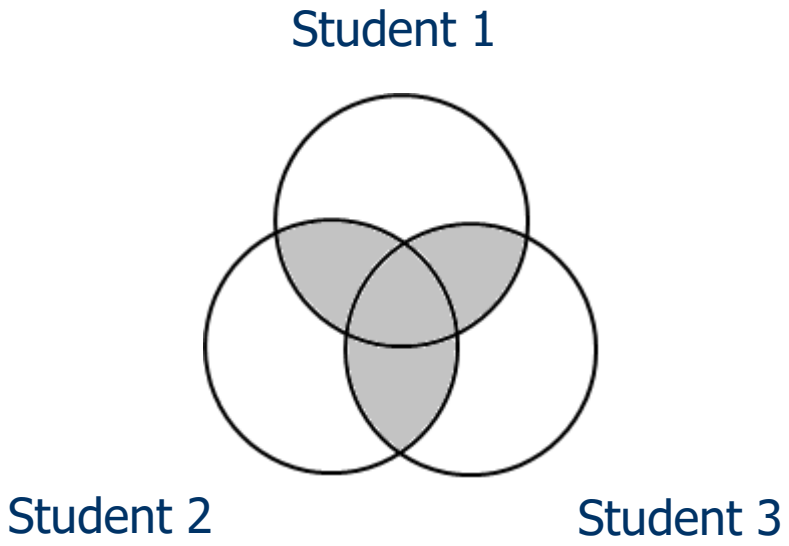


- The circle represents the information that 3 students (Student 1, 2 & 3) provide about popularity.
- Overlap indicates “similar information”.

TILBURG UNIVERSITY

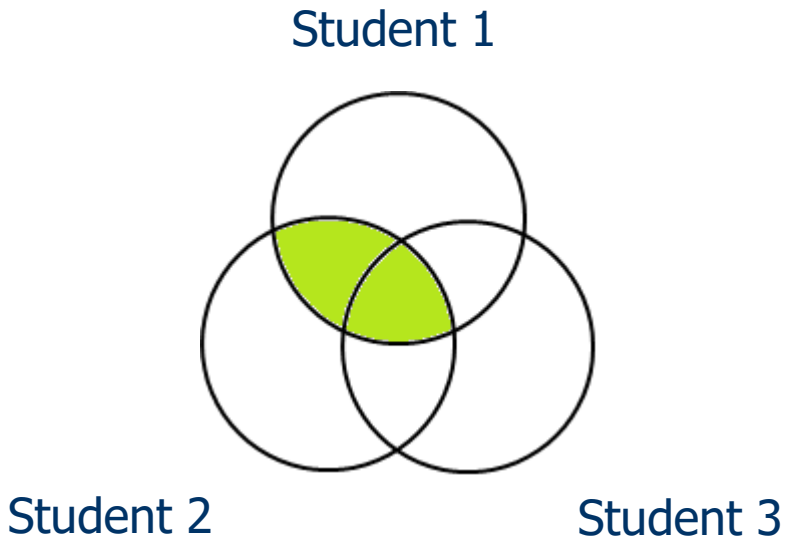
t e s c

Multilevel Data



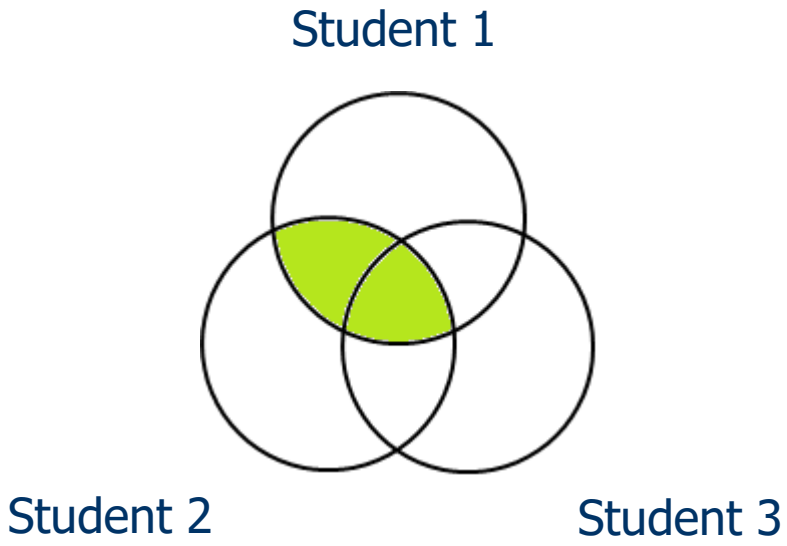
- The circle represents the information that 3 students (Student 1, 2 & 3) provide about popularity.
- Overlap indicates “similar information”.
  - So the grey parts are information that is provided by multiple students.

Multilevel Data



- The circle represents the information that 3 students (Student 1, 2 & 3) provide about popularity.
- When we add Student 2 the student partially tells us what we already know.
  - The green part of the information provided by Student 2 we already "learned" from Student 1 (and 3).

Multilevel Data



- The circle represents the information that 3 students (Student 1, 2 & 3) provide about popularity.
- When we add Student 2 the student partially tells us what we already know.
  - The green part of the information provided by Student 2 we already "learned" from Student 1 (and 3).
- This implies that the observed number of students we have IS NOT the same as the number-of-students-worth of information we have.
  - The effective sample size is smaller than the observed sample size.

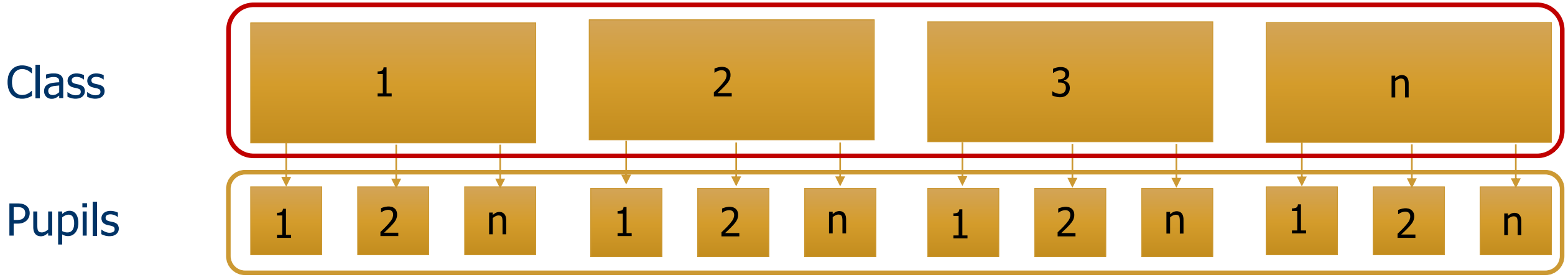
Tilburg University

tesc

Multilevel Data

- Ok, so data not completely independent.
- Independence is an assumption of most analyses, including regression, but so what?
- Helps to think in terms of information.
  - How much unique information does each student add to our sample?
- So we care about dependence because the observed sample size is not a good measure for how much information we have (and we need to account for that).

Multilevel Modeling



Variable:
Popularity
Extraversion
Gender

Teacher Experience

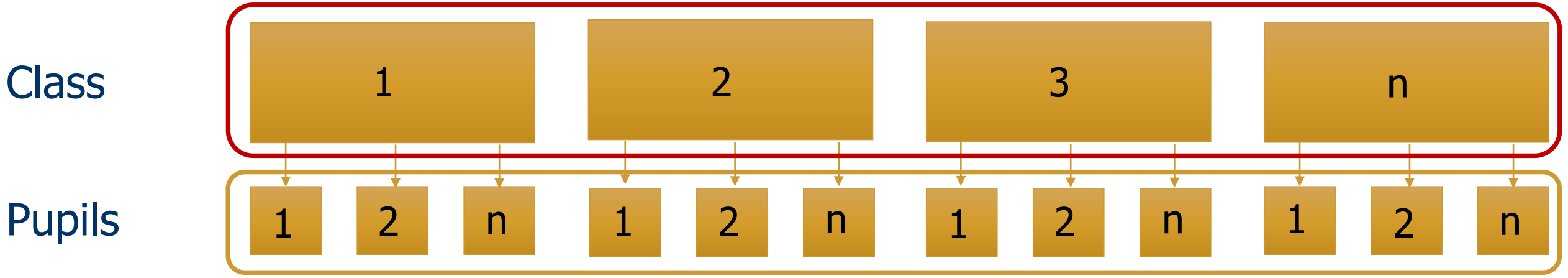
Level:
Pupil (level 1)
Pupil (level 1)
Pupil (level 1)

Class (level 2)

Tilburg University

t e s c

Multilevel Data



- We have two "units of interest"
  - Pupils
  - Class

- We want to test the effect of both pupil- and Class characteristics (and their interaction) on Popularity (also a student characteristic).

Tilburg University

t e s c

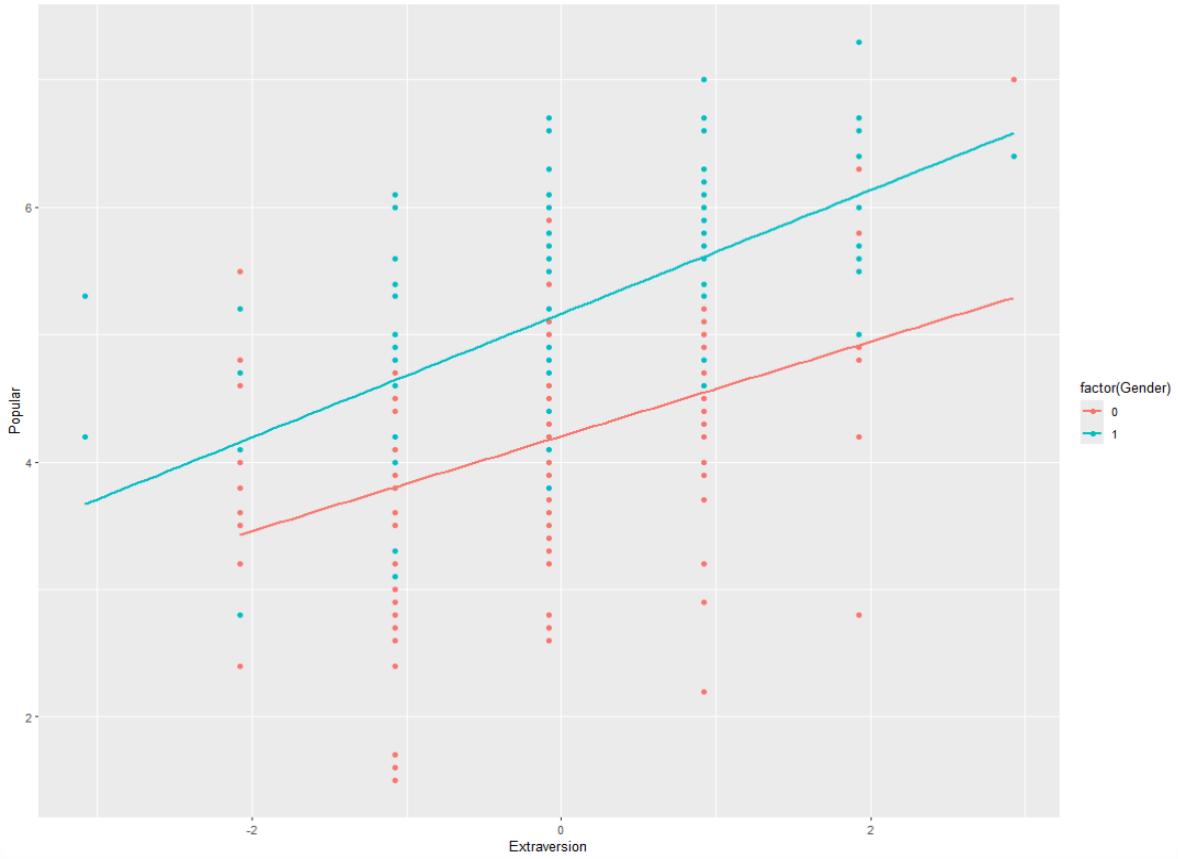
Multilevel Data

- So, our model needs to:
  - Use the correct sample size (the effective one, NOT the observed one)
  - Combine two units of interest (here, pupils and classes) into one analysis
  - Meet other regression assumptions (e.g., normality, linearity).

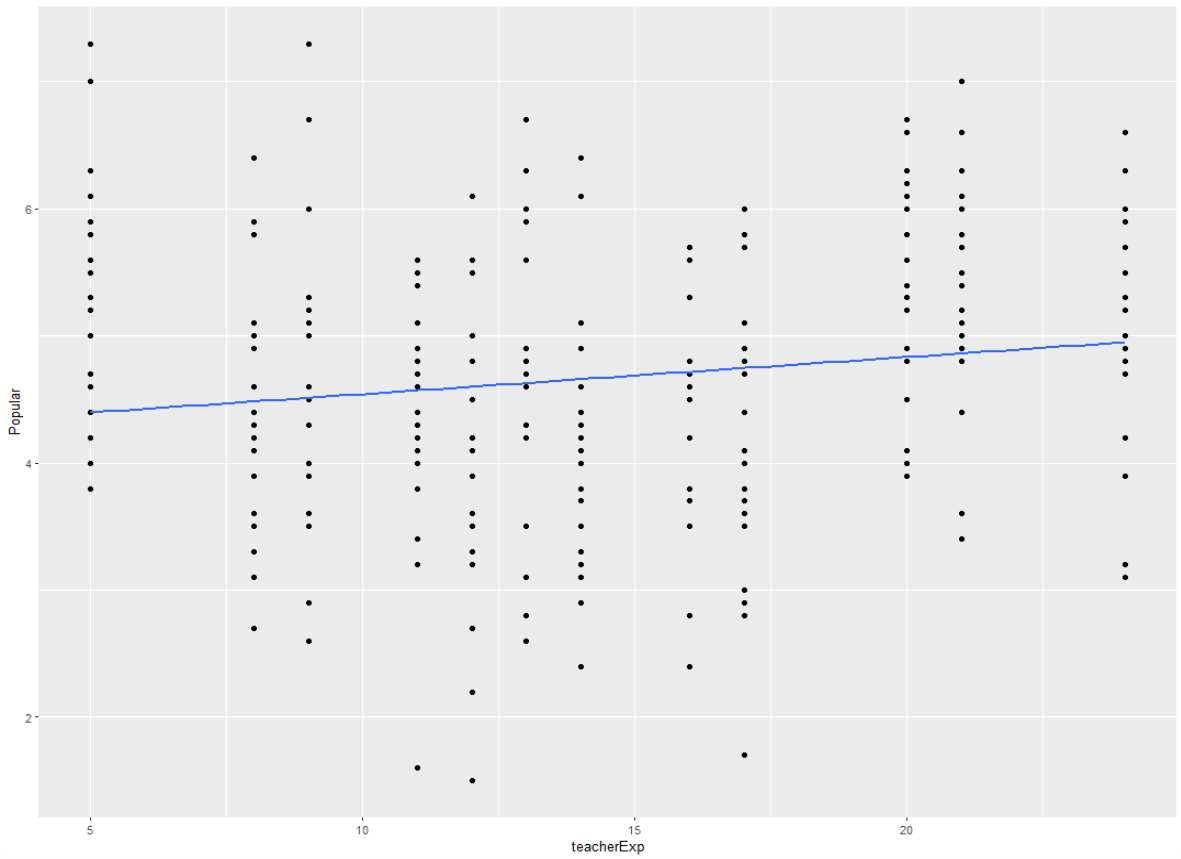
- This is basically done by running several regression analyses and tying these together in a clever way.

Multilevel Data

Level 1



Level 2

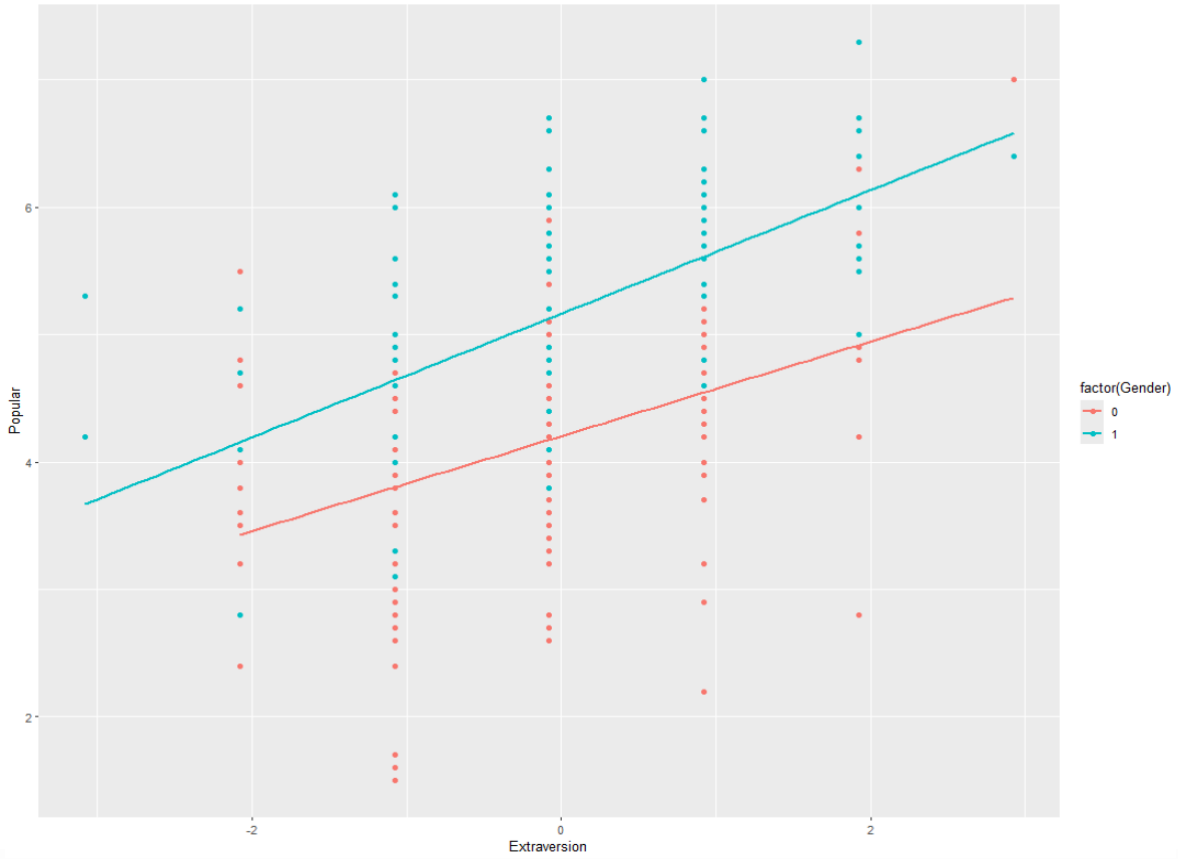


TILBURG UNIVERSITY

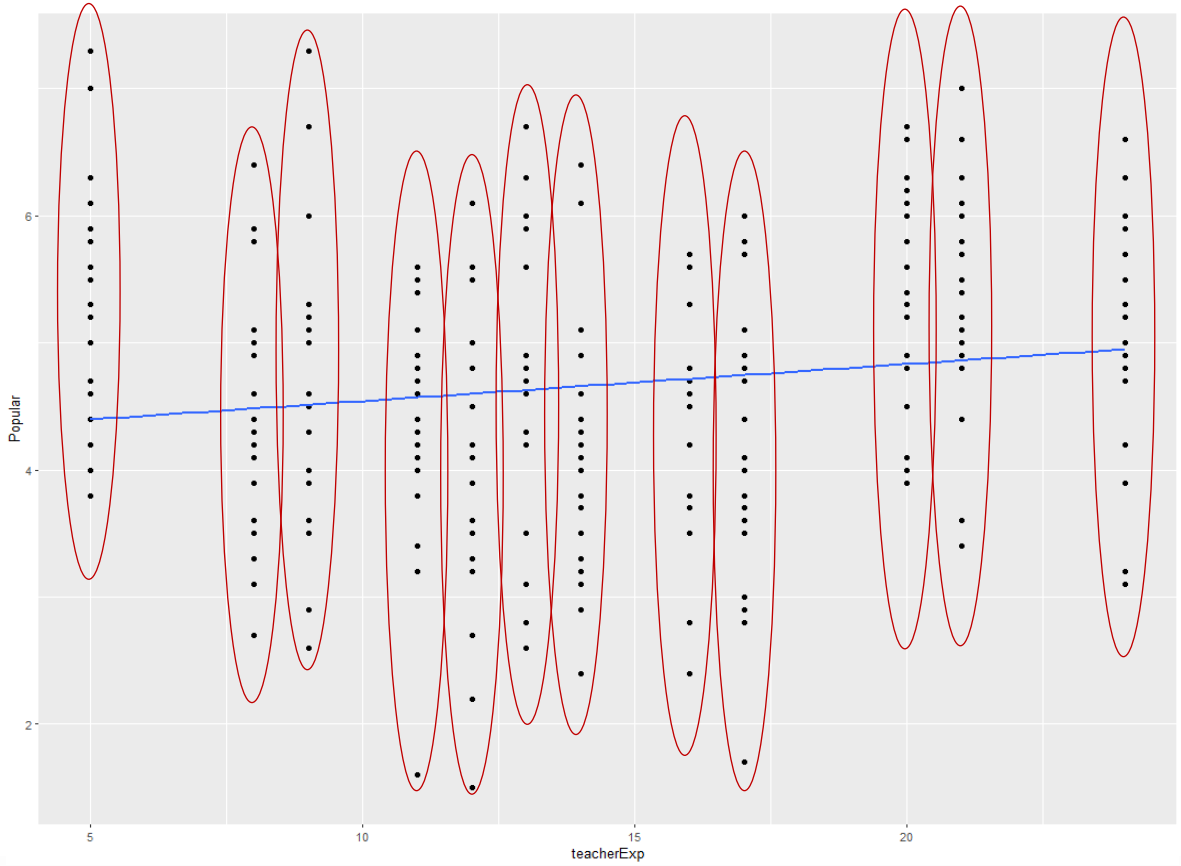
t e s c

Multilevel Data

Level 1



Level 2



TILBURG UNIVERSITY

tesc

Multilevel Data

	Pupil	Class	Extraversion	Gender	teacherExp	Popular
1	1	1	-0.07723577	1	24	6.3
2	2	1	1.92276423	0	24	4.9
3	3	1	-1.07723577	1	24	5.3
4	4	1	-2.07723577	1	24	4.7
5	5	1	-0.07723577	1	24	6.0
6	6	1	-1.07723577	0	24	4.7
7	7	1	-0.07723577	0	24	5.9
8	8	1	-1.07723577	0	24	4.2
9	9	1	-0.07723577	0	24	5.2
10	10	1	-0.07723577	0	24	3.9
11	11	1	-0.07723577	1	24	5.7
12	12	1	-0.07723577	1	24	4.8
13	13	1	-0.07723577	0	24	5.0
14	14	1	-0.07723577	1	24	5.5
15	15	1	-0.07723577	1	24	6.0
16	16	1	0.92276423	1	24	5.7
17	17	1	-1.07723577	0	24	3.2
18	18	1	-1.07723577	0	24	3.1
19	19	1	1.92276423	1	24	6.6
20	20	1	-1.07723577	0	24	4.8
21	1	2	2.92276423	1	14	6.4
22	2	2	-1.07723577	0	14	2.4
23	3	2	0.92276423	0	14	3.7
24	4	2	-0.07723577	1	14	4.4
25	5	2	-0.07723577	1	14	4.3

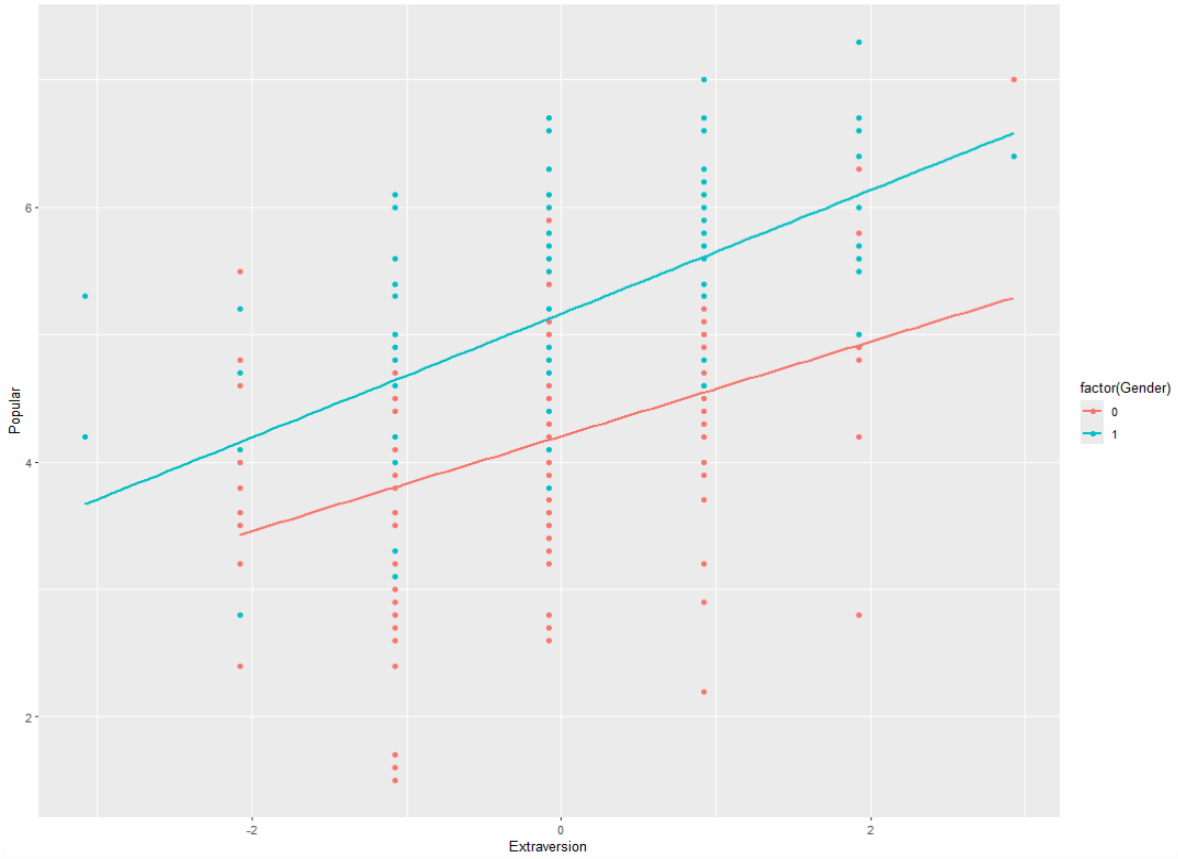
- We can't just predict popularity scores because we overestimate the number of observations on the Teacher Experience variable!
  - Observed the score 24 once! Not 20 times.
- What we actually do is predict average class popularity using Teacher Experience.

TILBURG UNIVERSITY

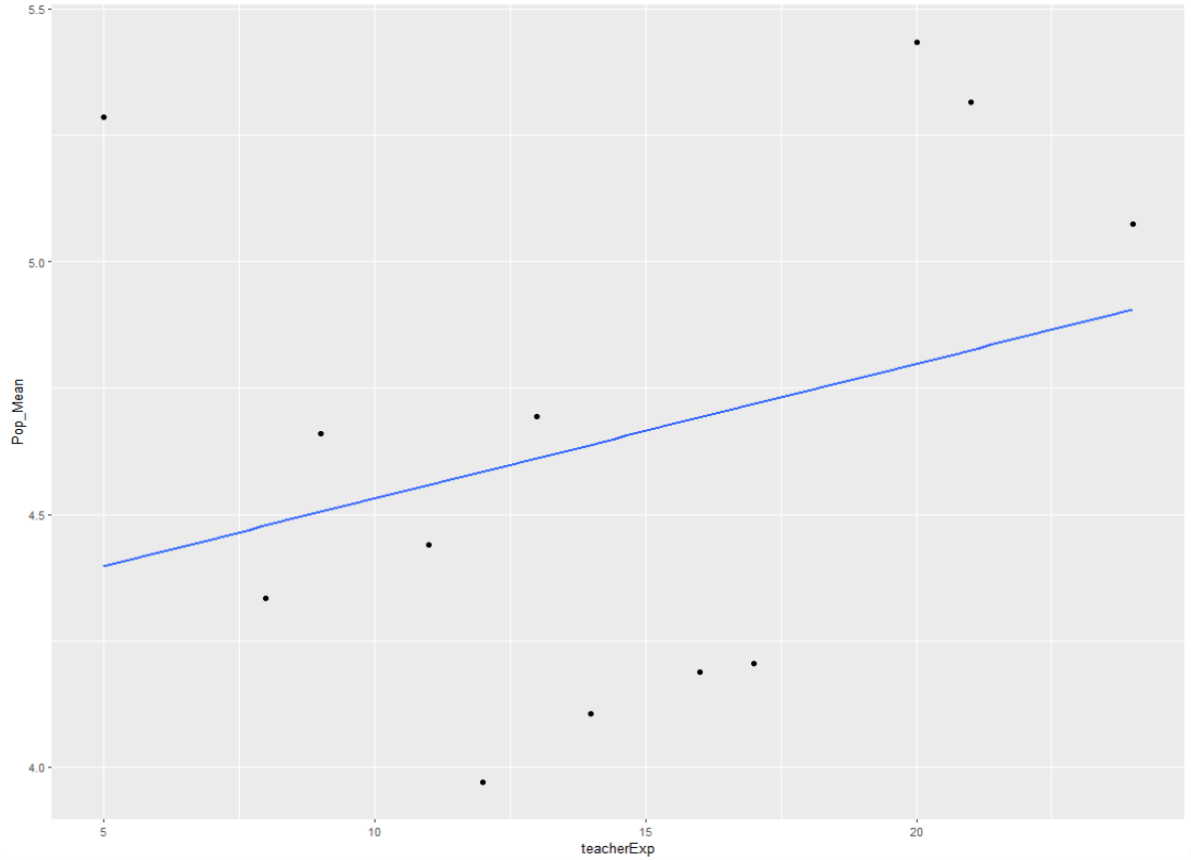
t e s c

Multilevel Data

Level 1



Level 2

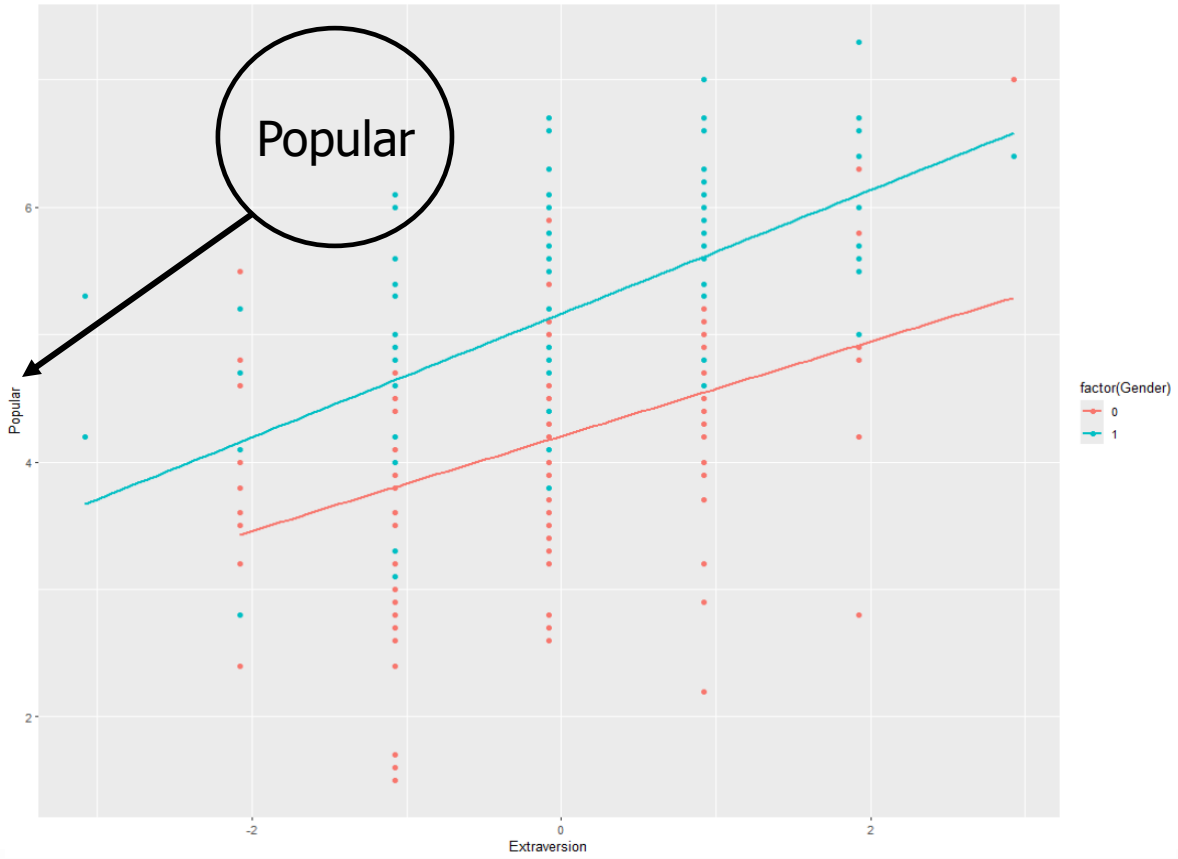


TILBURG UNIVERSITY

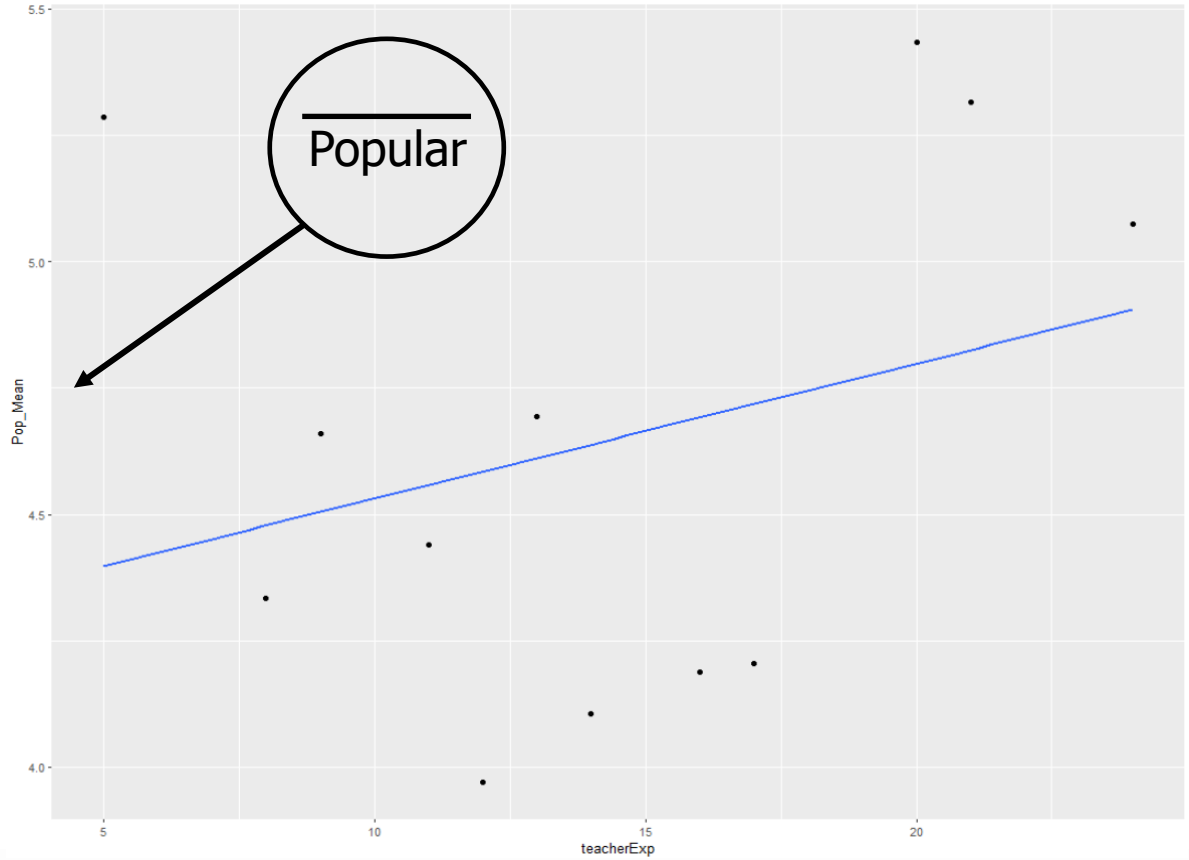
tesc

Multilevel Data

Level 1



Level 2



TILBURG UNIVERSITY

tesc

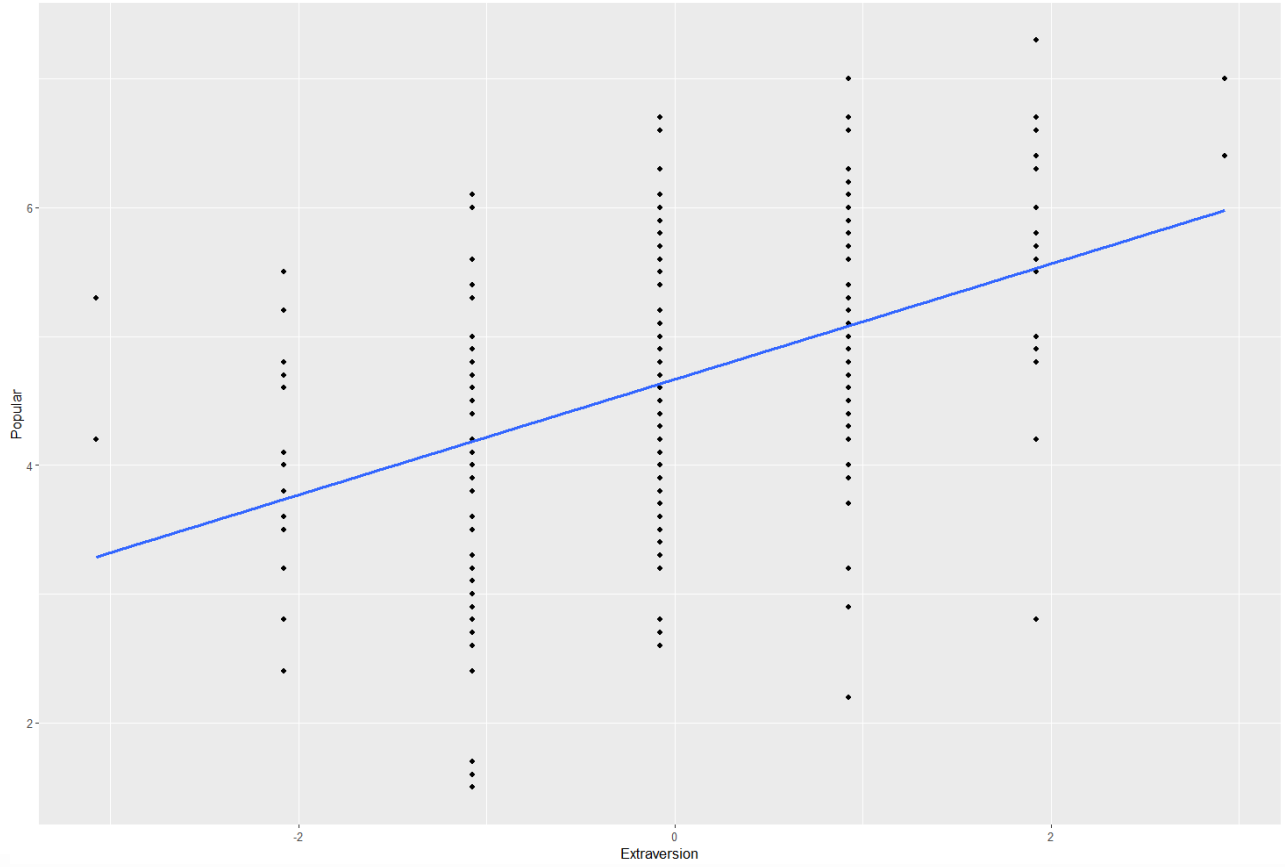
Multilevel Data

- So we have to combine two regression analyses
  - One on the individual Popularity scores.
  - One on the class averages of the Popularity scores.

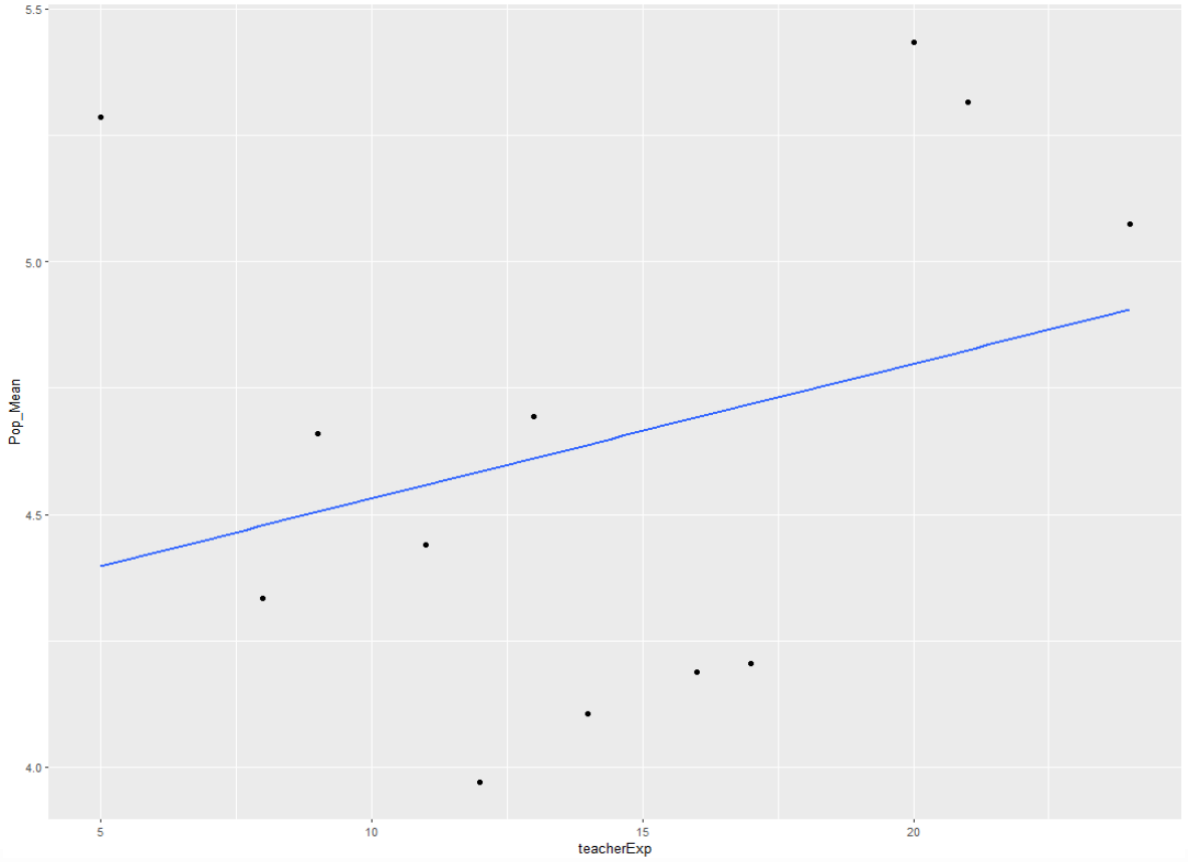
- How can we do that? How can we relate the means (the DV on level 2) to the regression analysis on level 1?

Multilevel Data

Level 1



Level 2

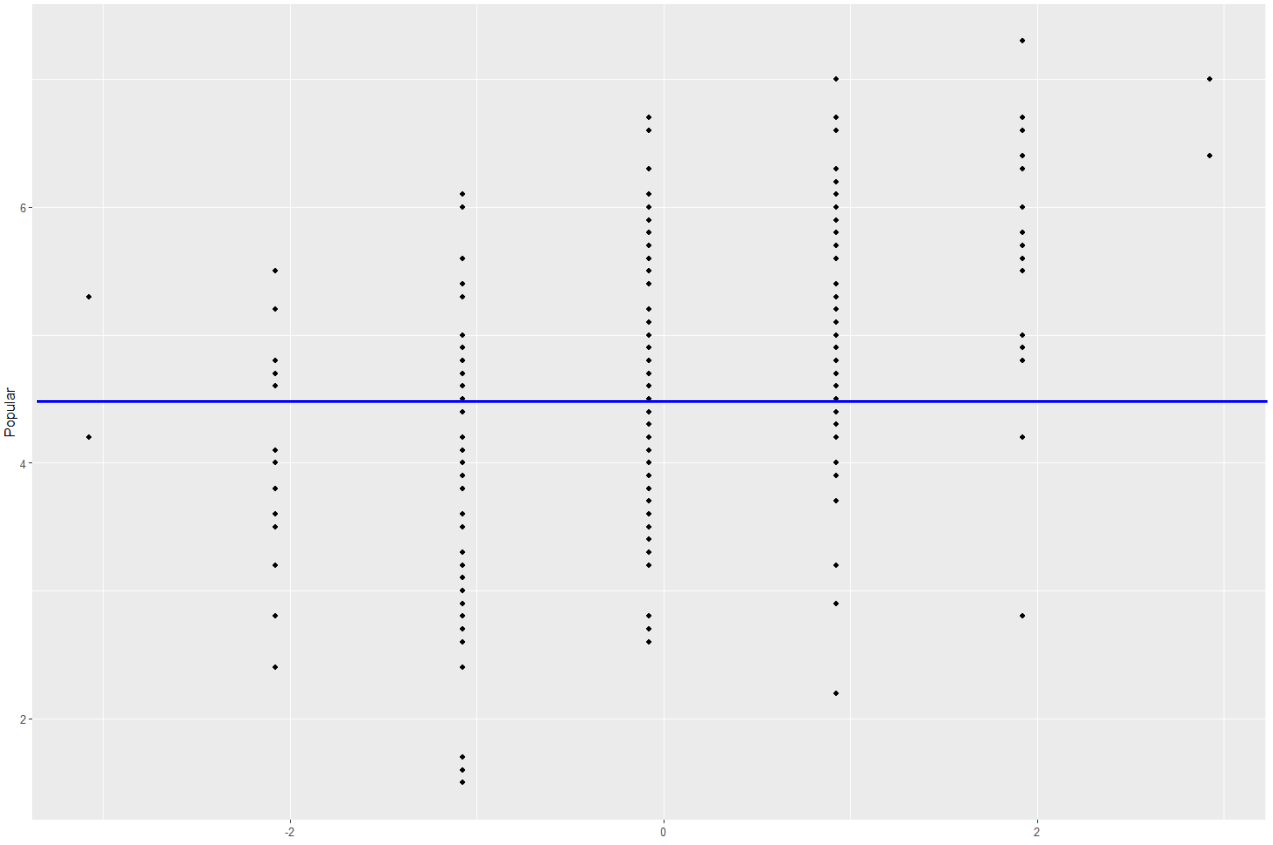


TILBURG UNIVERSITY

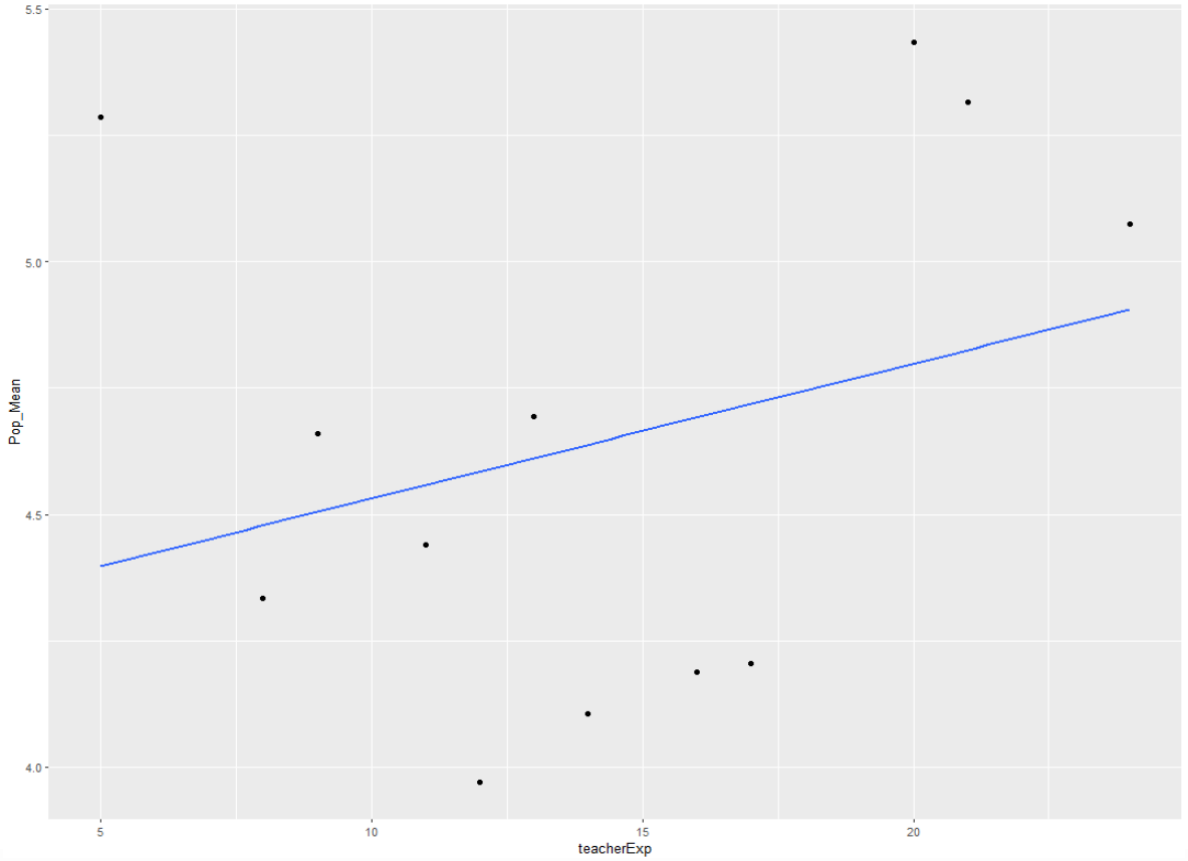
tesc

Multilevel Data

Level 1



Level 2

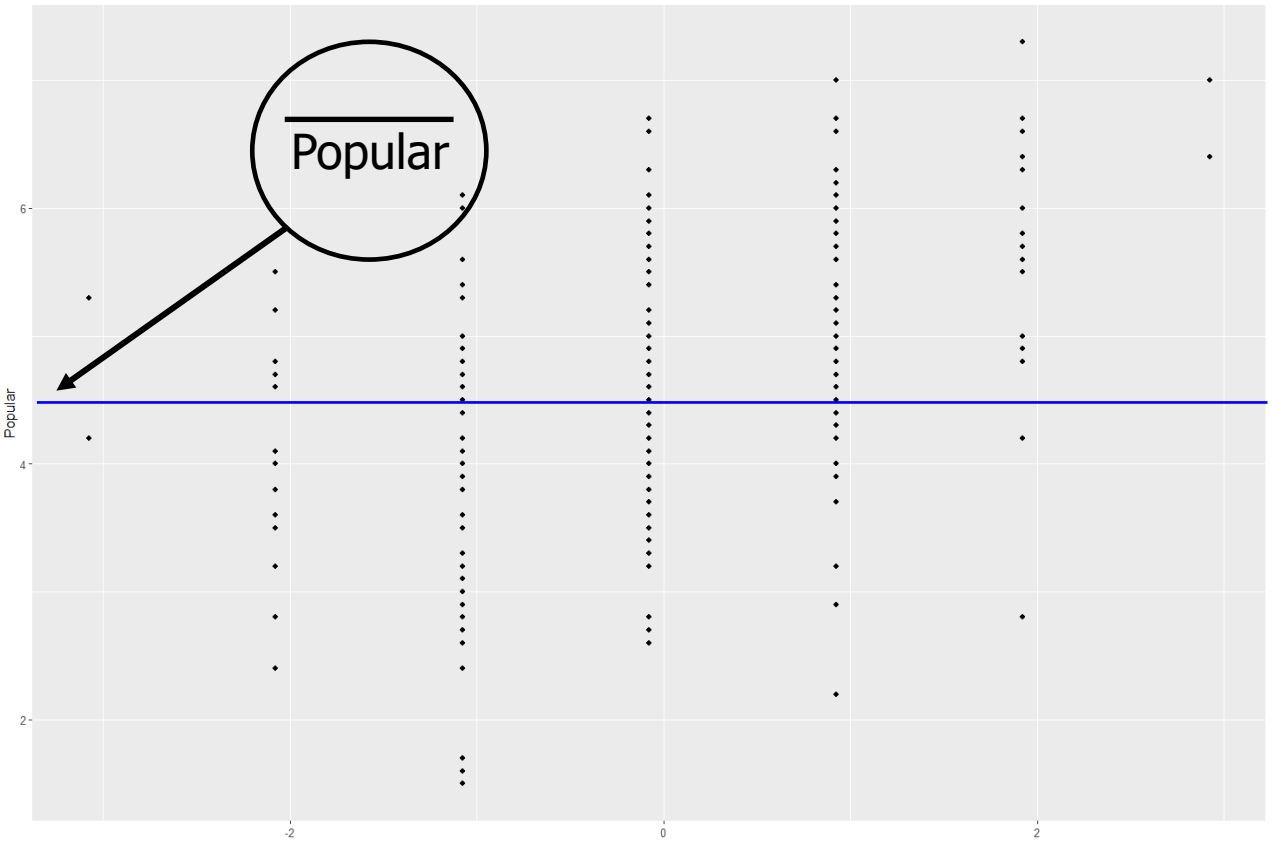


TILBURG UNIVERSITY

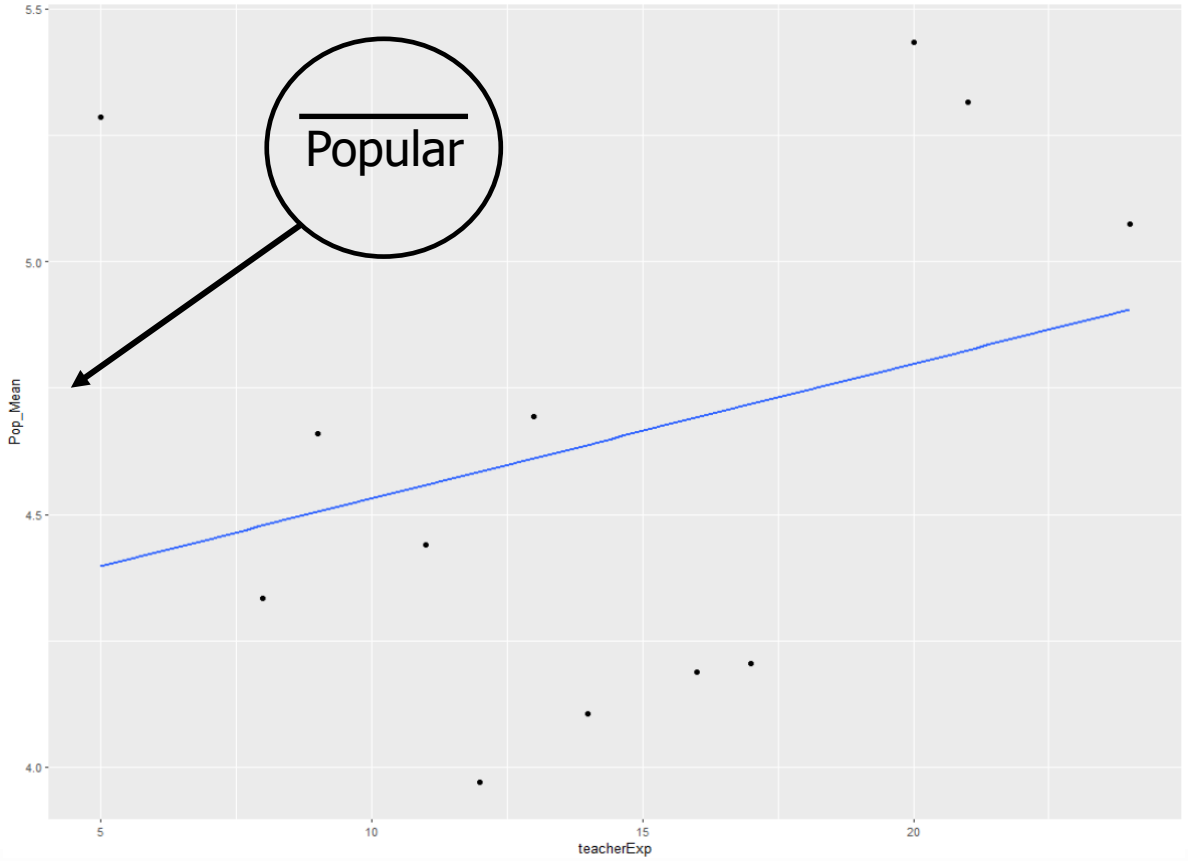
tesc

Multilevel Data

Level 1



Level 2



TILBURG UNIVERSITY

tesc

Multilevel Data

- So we connect the two regression equations through the intercept of the level one equation.

- The intercept of the regression analysis on level 1 is:
  - a parameter on level 1
  - the dependent variable on level 2!

- Later we'll see that we can turn other parameters from the level one equation into outcomes on level 2 as well (e.g. regression coefficients).

Multilevel Data

- Doing this properly requires some “estimation tricks”.
  - For example, to account for the effect that the number of observations per level 2 units will likely differ, meaning we have more certainty about the (e.g.) intercepts of some level 2 units.

- In the video on Partial Pooling, a general intuition will be provided for these “tricks”.

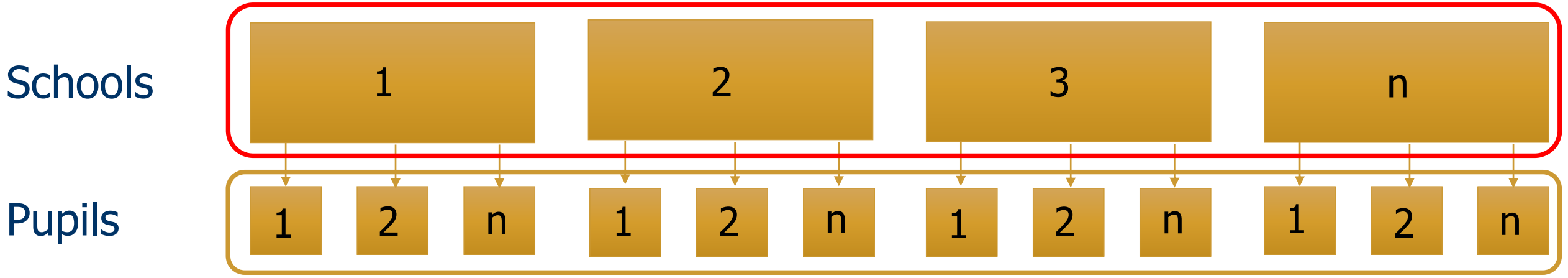
Modeling Multilevel Data

Modeling Multilevel Data

- Study on the popularity of highschool students.
  - Total of 246 students from 12 different classes.
  - Determined how extraversion, gender, and teacher experience influenced a student's popularity.

- List of all the variables:
  - pupil: pupil identification variable, not needed in the analysis
  - class: class identification variable, the linking variable to define the 2 - level structure
  - student-level independent variables: extraversion (continuous; higher scores mean higher extraversion) and gender (dichotomous; 0=male, 1 =female)
  - class-level independent variables: teacher experience (in years)
  - outcome variable: popular (continuous outcome variable at the student-level, higher scores indicate higher popularity)

Modeling Multilevel Data



Variable:	**Level:**
Popularity	Pupil (level 1)
Extraversion	Pupil (level 1)
Gender	Pupil (level 1)
Teacher Experience	School (level 2)

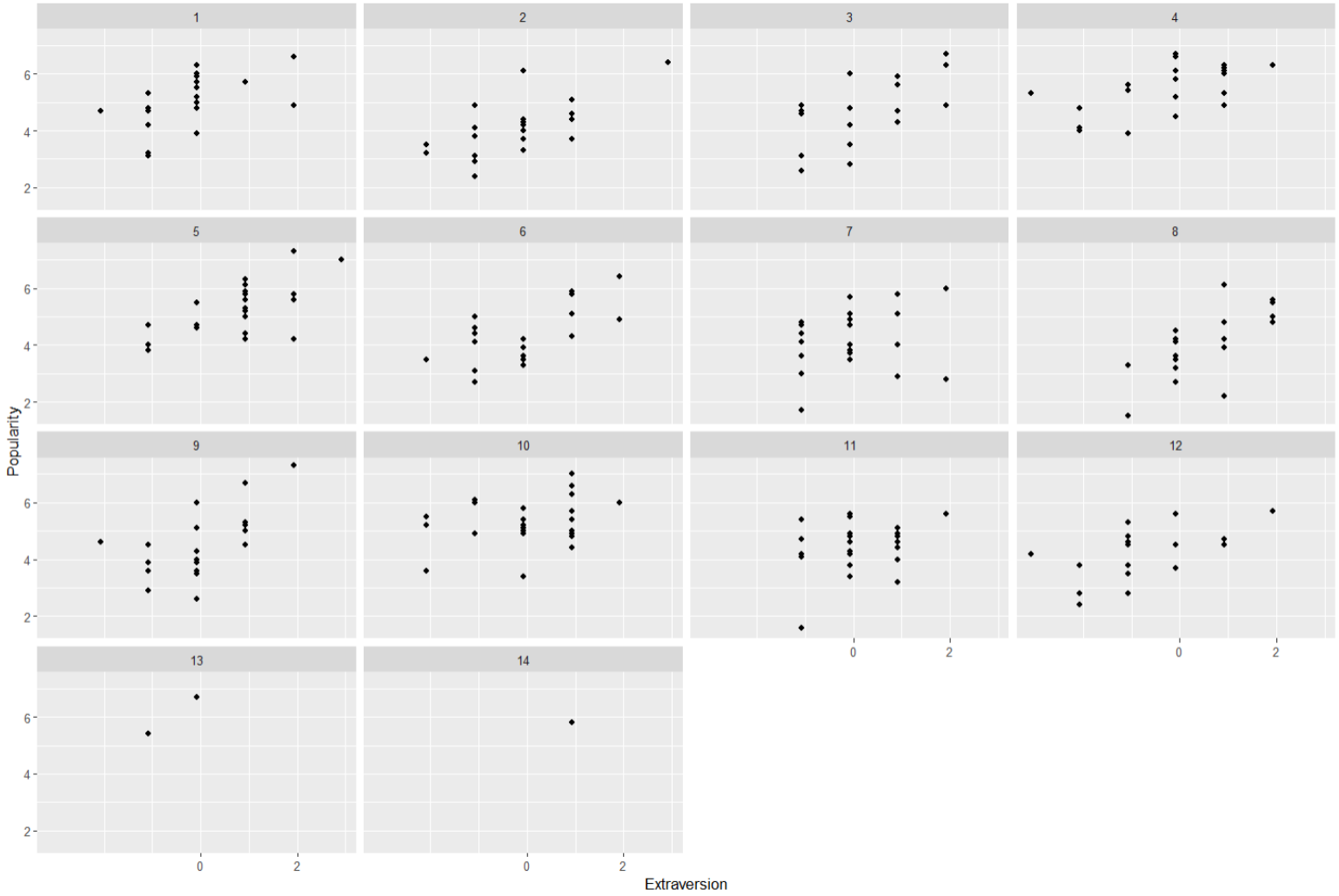
Tilburg University

t e s c

Modeling Multilevel Data

- Let's look at how we could analyze these data.
- Will focus on just Popularity and Extraversion for now.

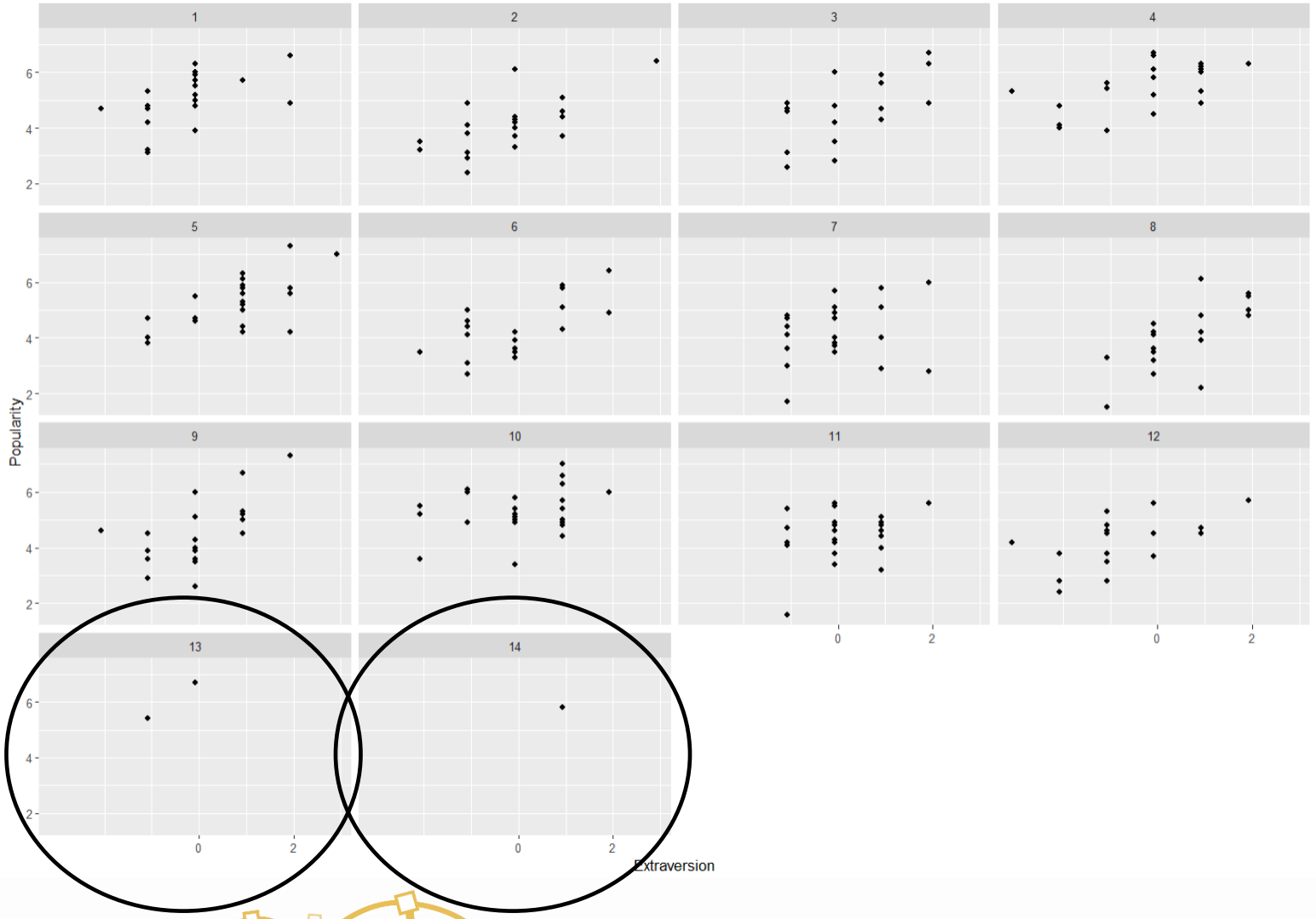
Modeling Multilevel Data



TILBURG UNIVERSITY

tesc

Modeling Multilevel Data



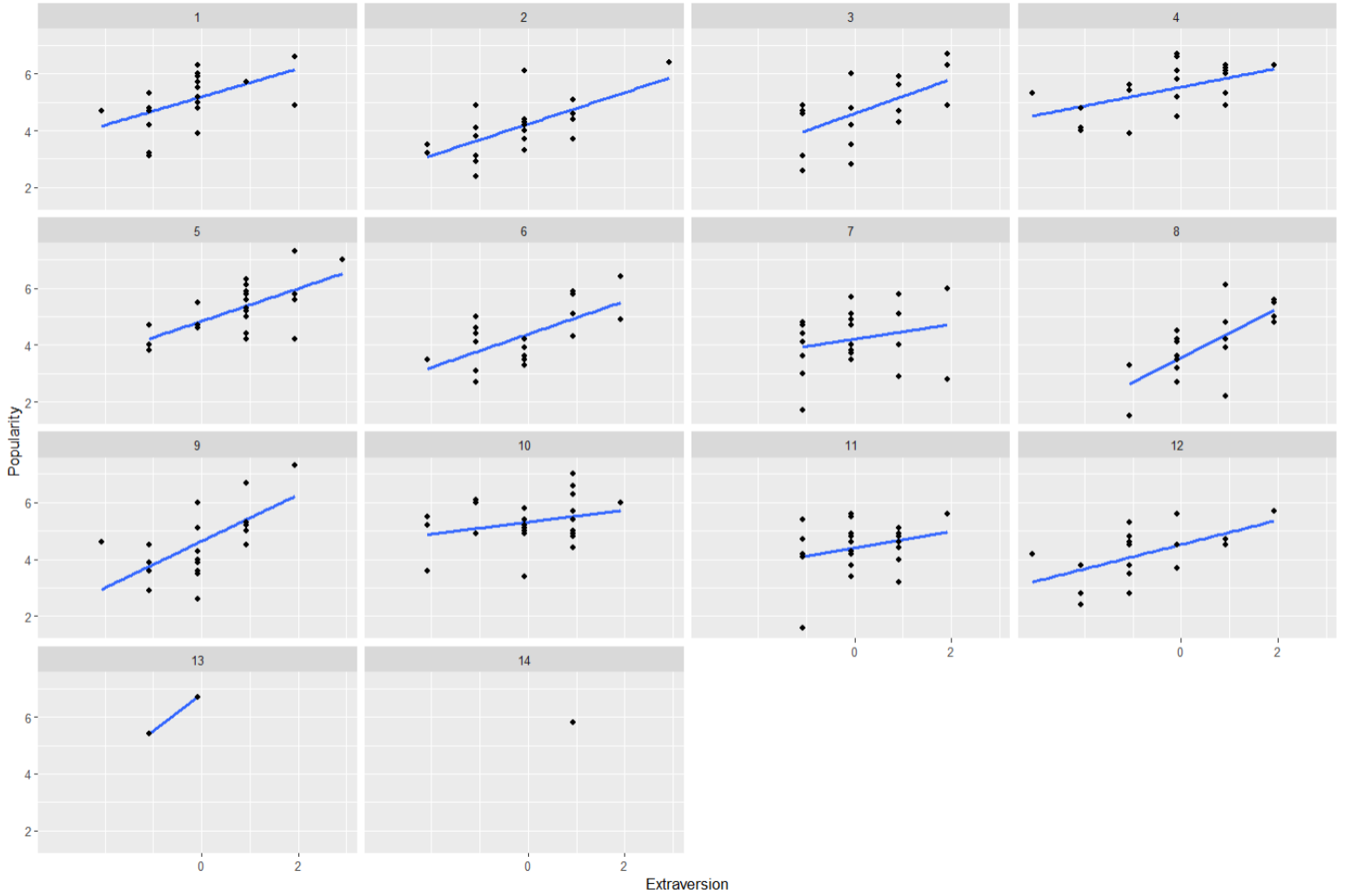
TILBURG UNIVERSITY

tesc

Some Options: No Pooling

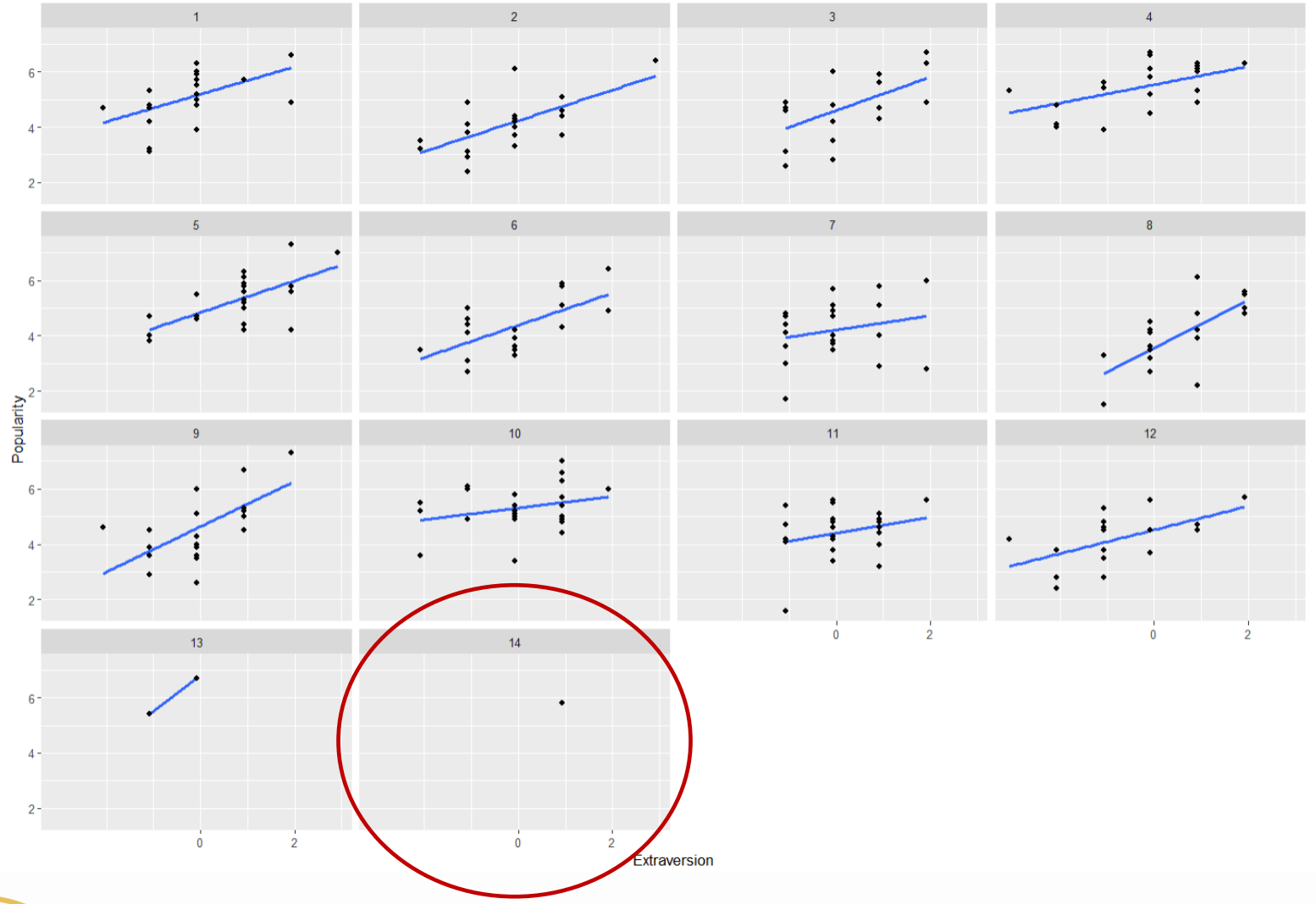
No Pooling

- Analyze all classes separately
- Benefits?
- Disadvantages?



No Pooling

- Analyze all classes separately
- Benefits?
- Disadvantages?



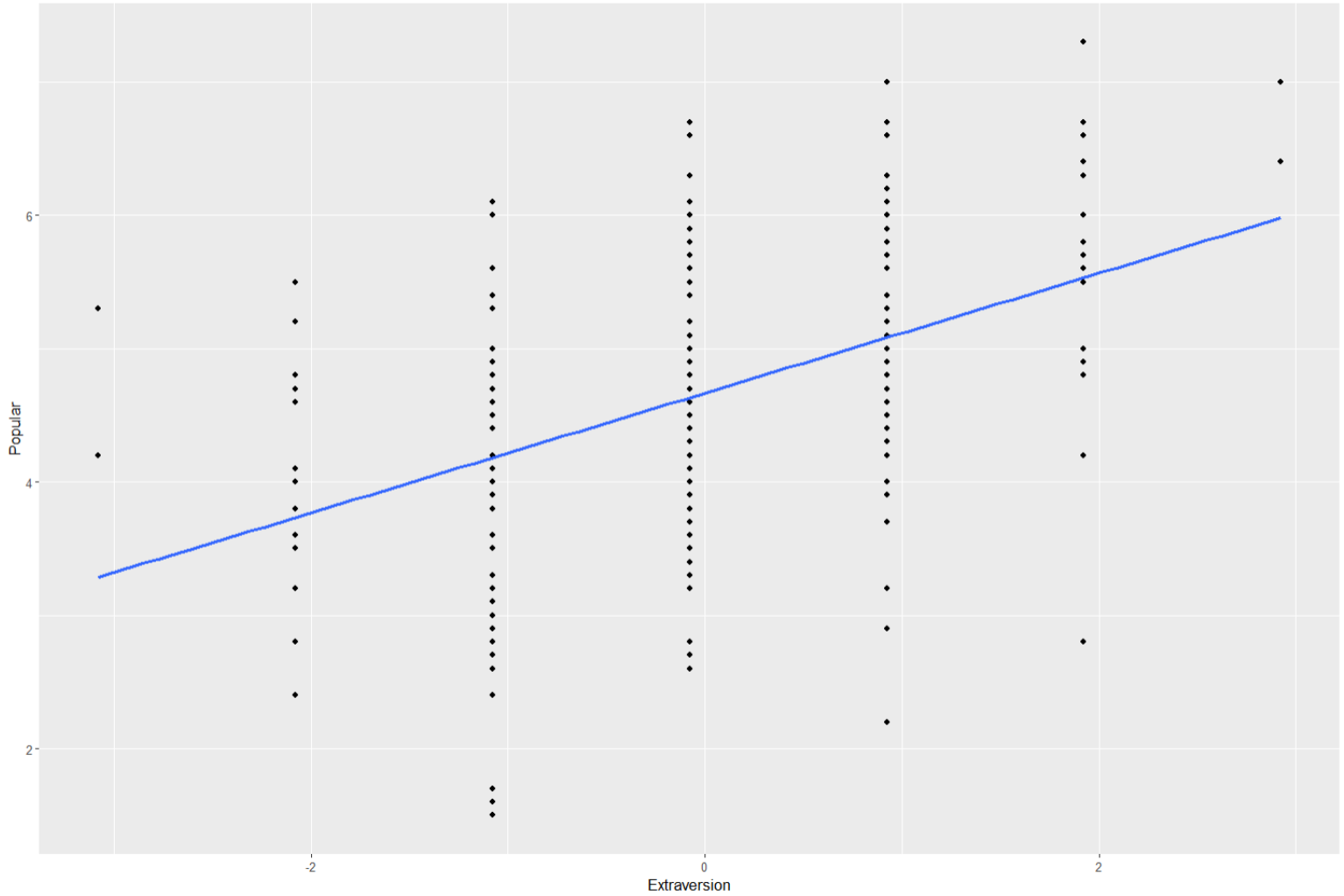
No Pooling

- "Many statistical models also have anterograde amnesia. As the models move from one cluster [...] 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" (Richard McElreath)

Some Options: Complete Pooling

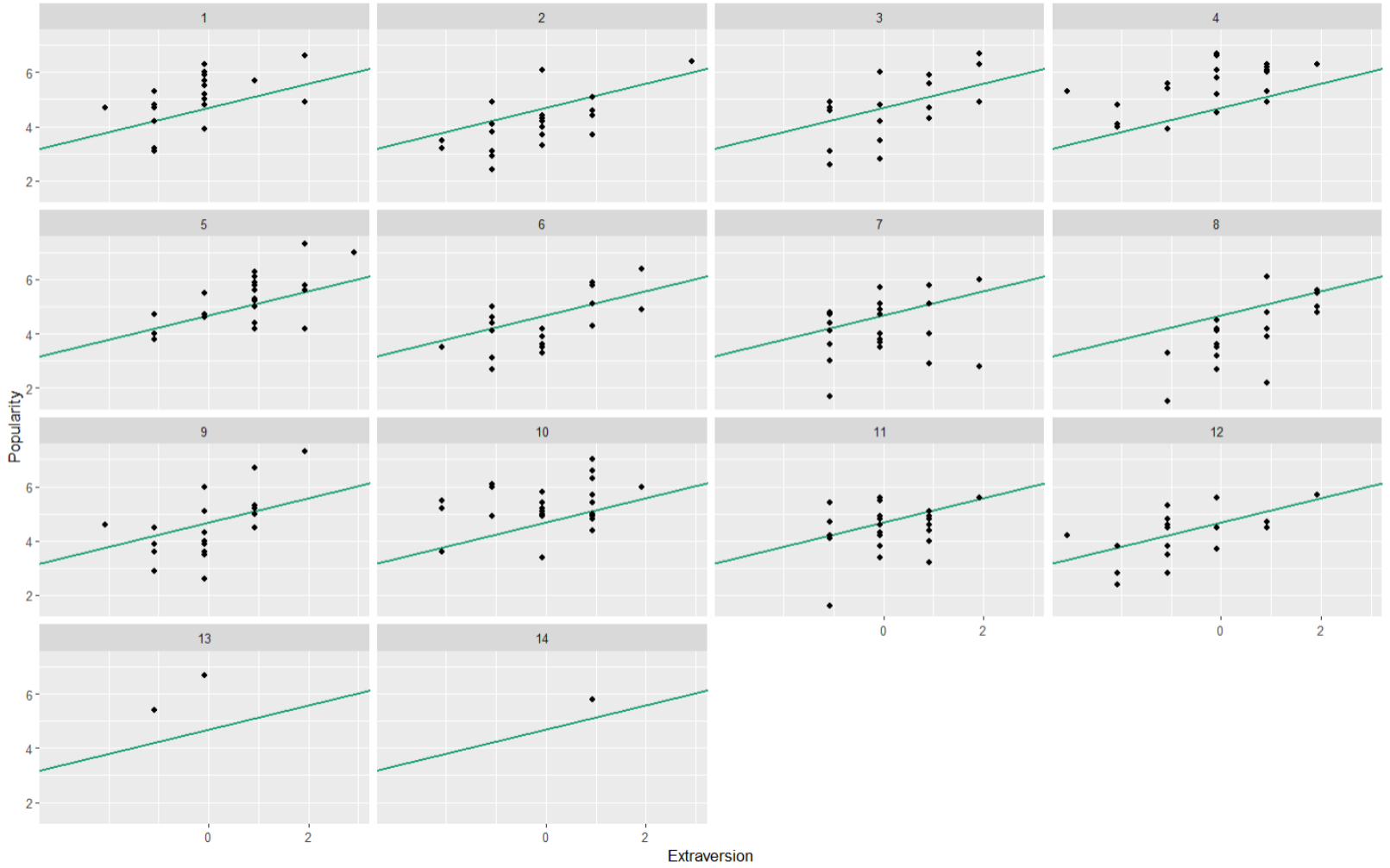
Complete Pooling

- Ok…so…analyze together then?



Complete Pooling

- Ok…so…analyze together then?

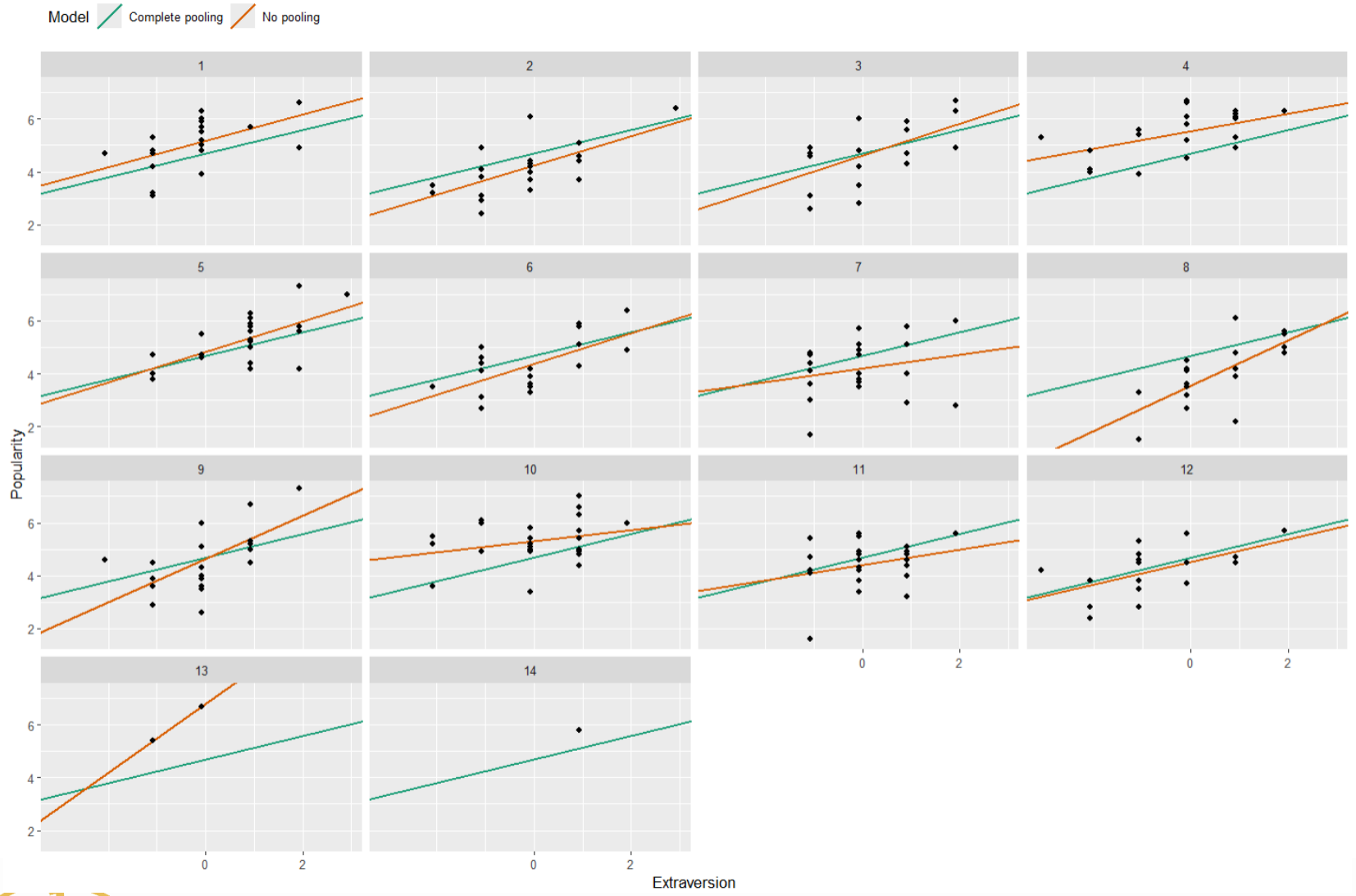


TILBURG UNIVERSITY

tesc

Complete Pooling

- Ok…so…analyze together then?

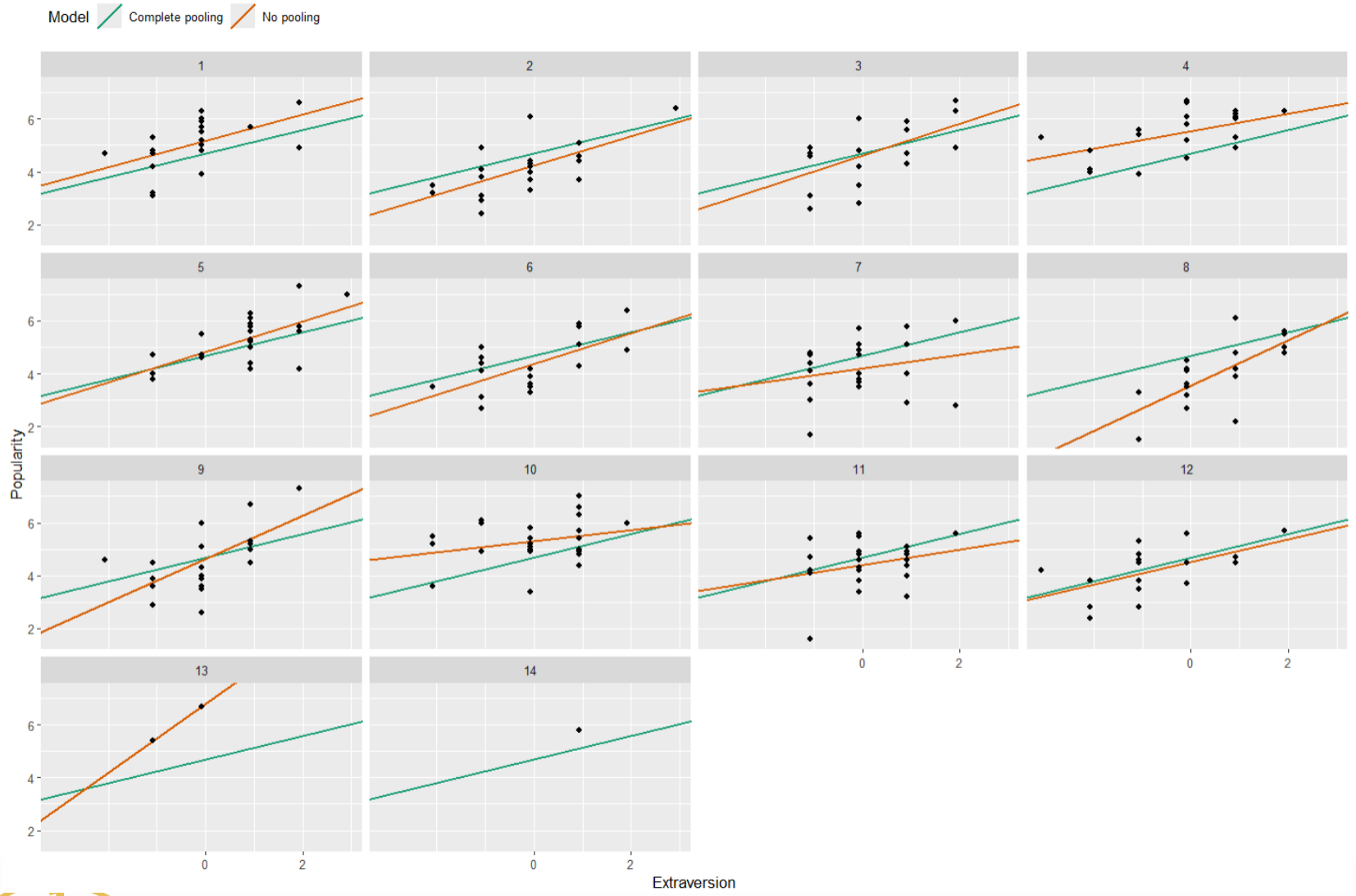


Tilburg University

tesc

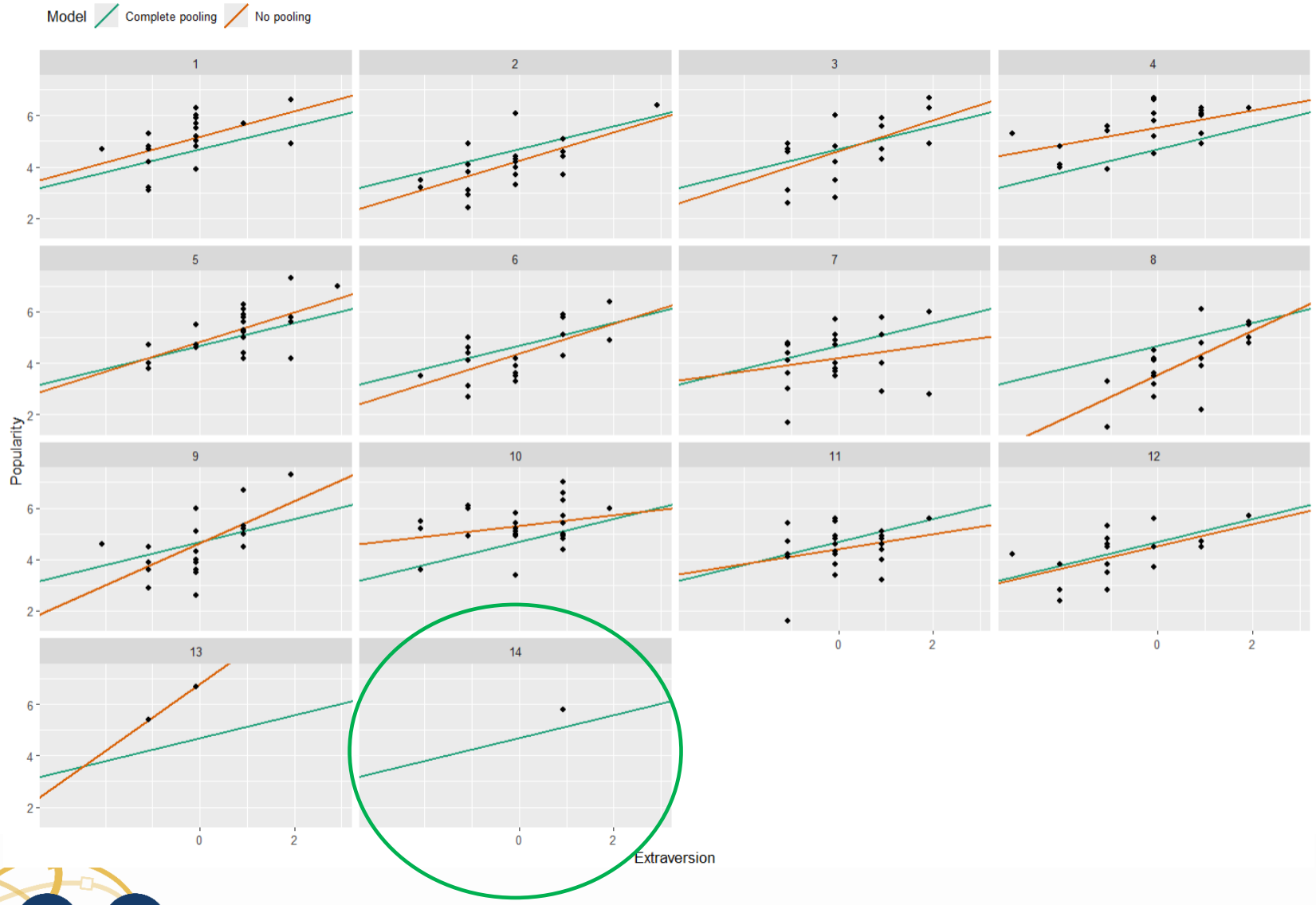
Complete Pooling

- Ok…so…analyze together then?
- Advantages?
- Disadvantages?



Complete Pooling

- Ok…so…analyze together then?
- Advantages?
- Disadvantages?

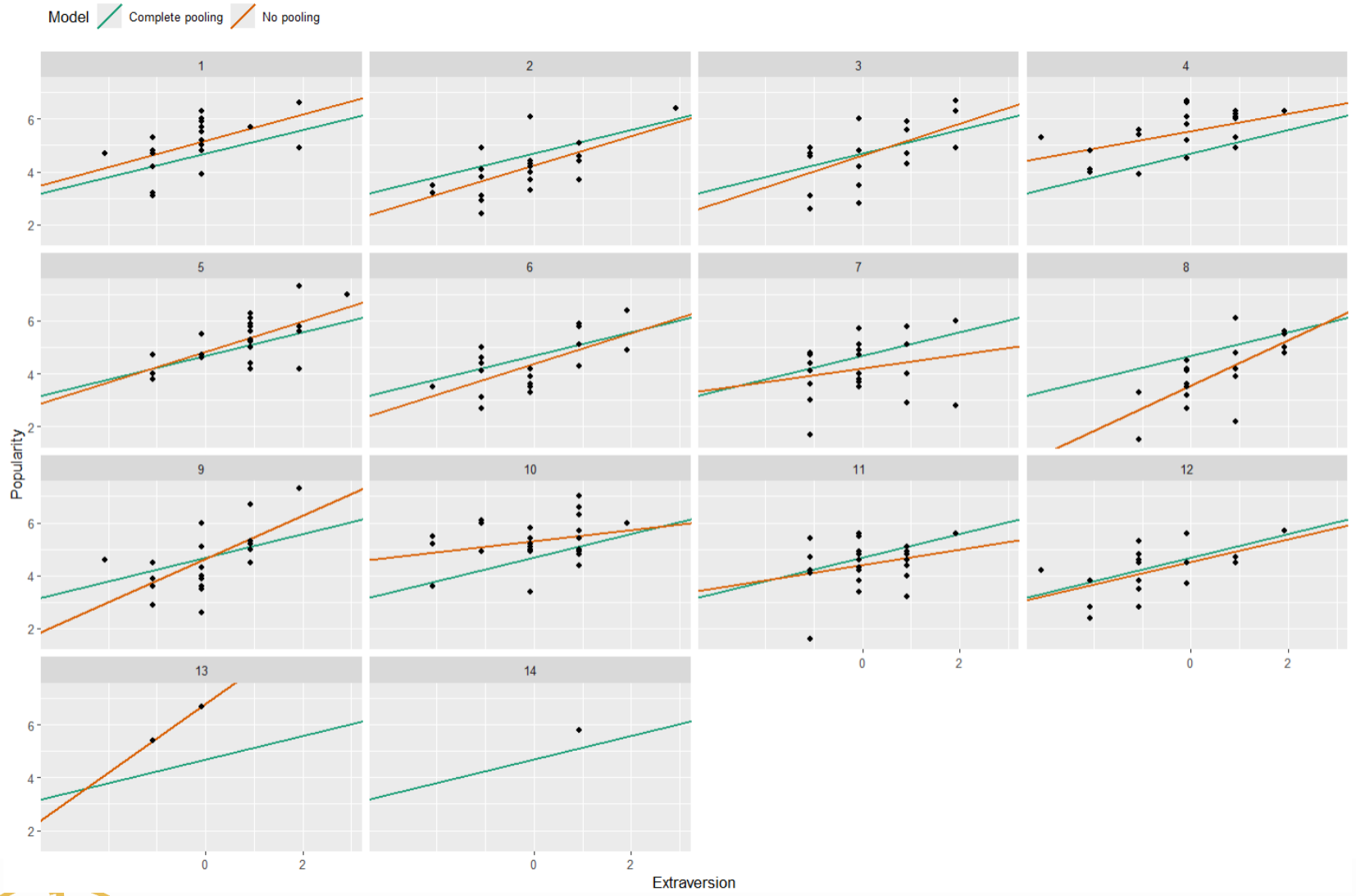


TILBURG UNIVERSITY

tesc

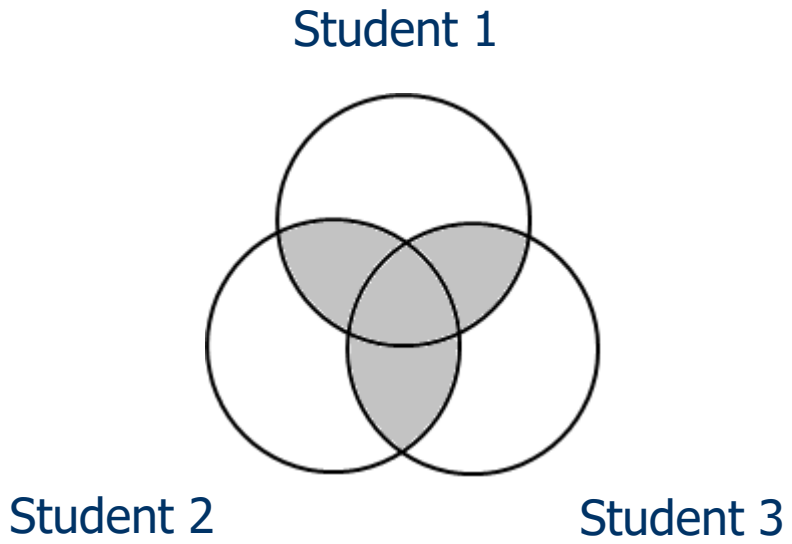
Complete Pooling

- Ok…so…analyze together then?
- Advantages?
- Disadvantages?



Complete Pooling

- Hint: How much data do we have?

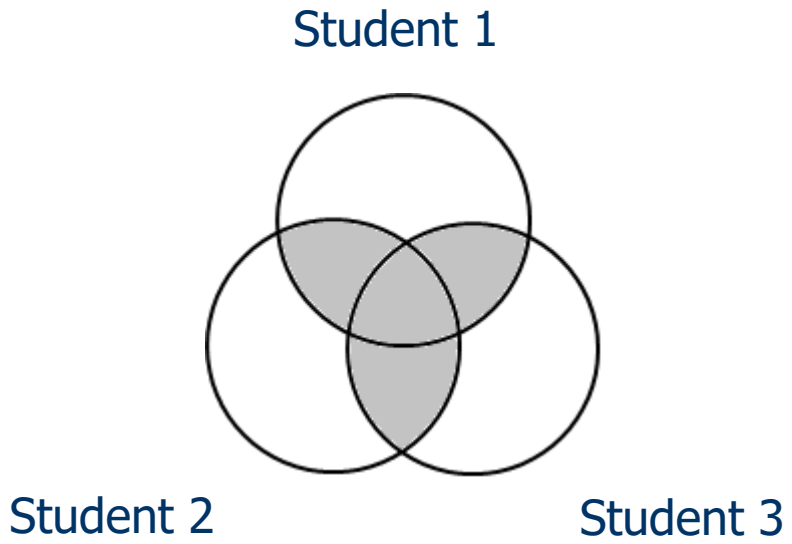


Complete Pooling

- Hint: How much data do we have?

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

- We have less data than data points.

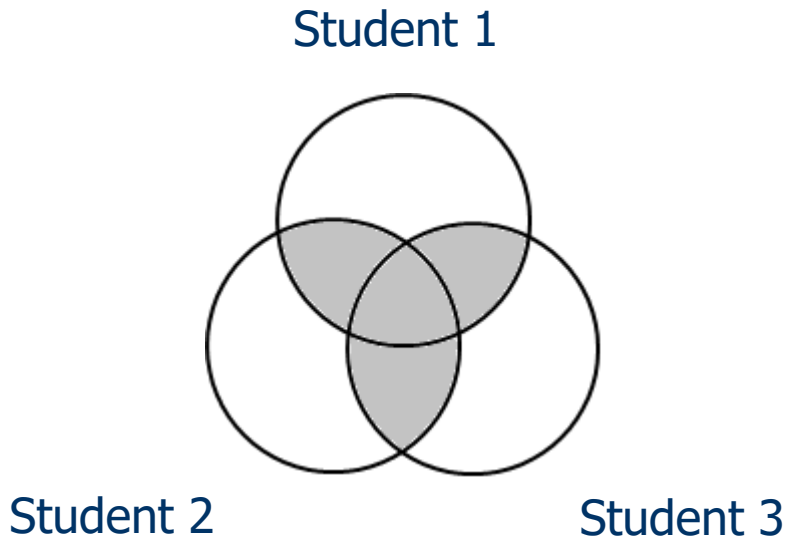


Complete Pooling

- Could adjust SEs

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

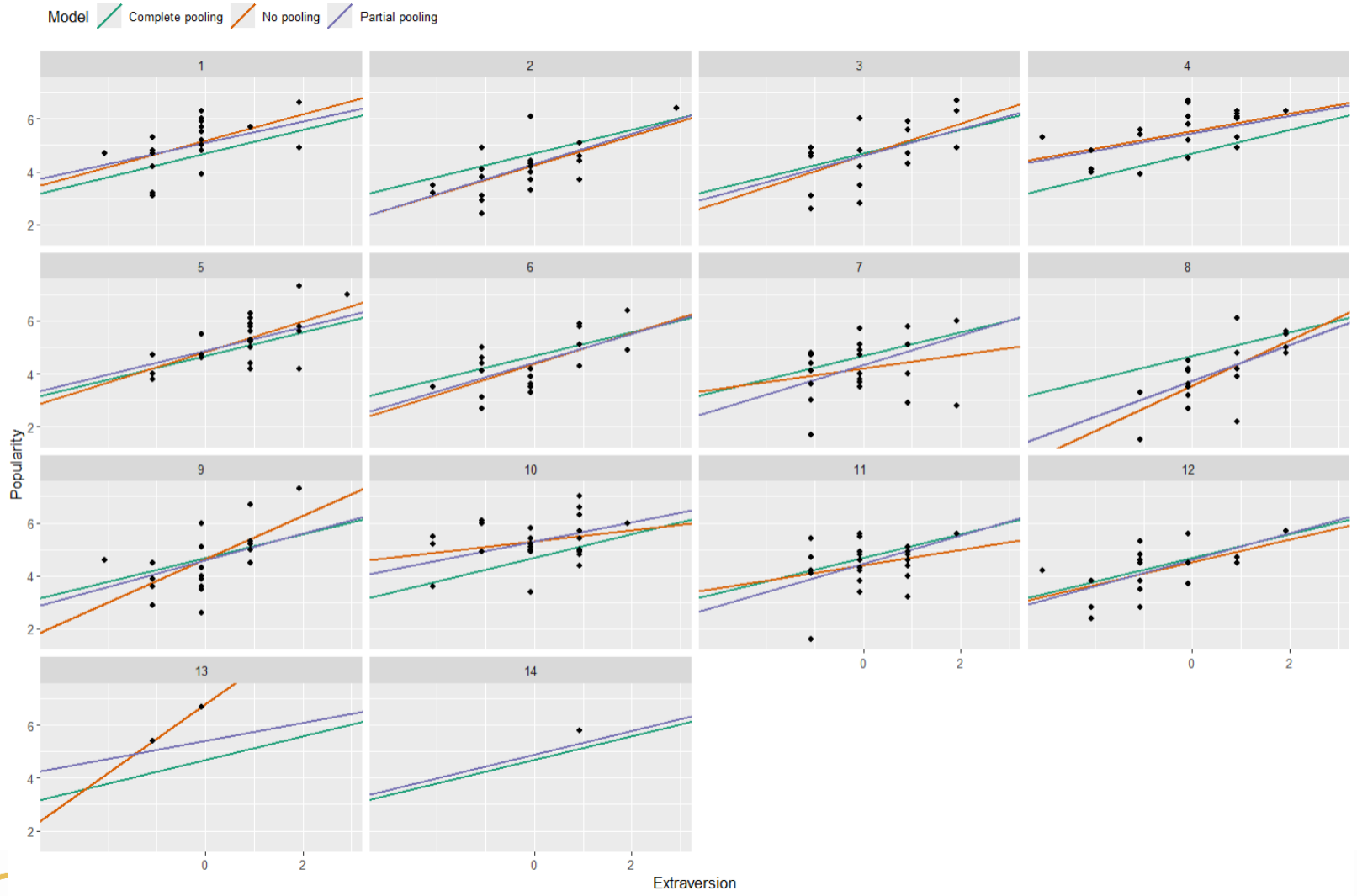
- Fortunately, there are smart ways to do this:
  - Cluster robust SEs



Some Options: Partial Pooling

Partial Pooling

- Pool information from all the lines together to improve our estimates of each individual line.
- After seeing the 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 classes!!

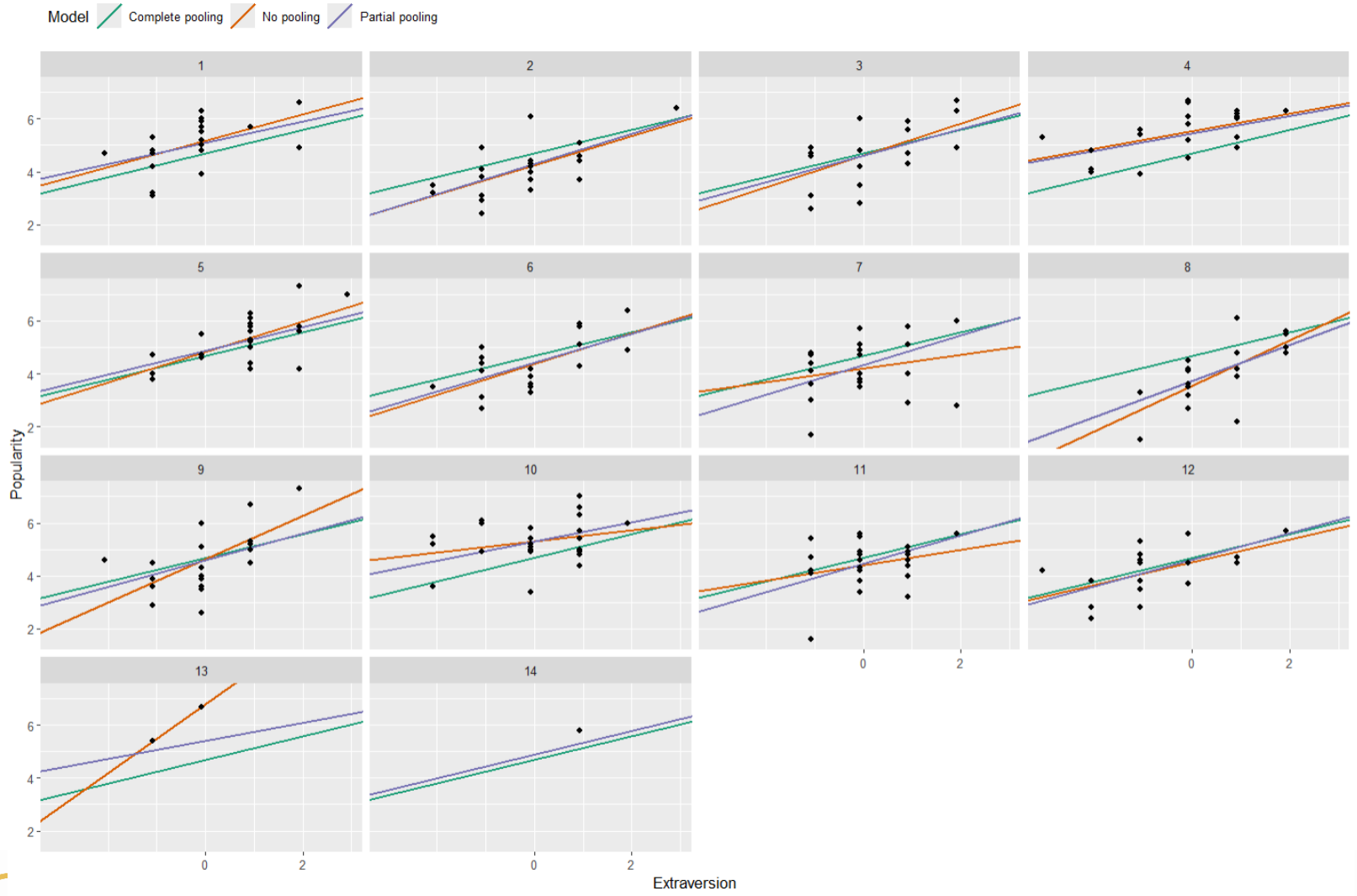


TILBURG UNIVERSITY

tesc

Partial Pooling

- 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 school is.

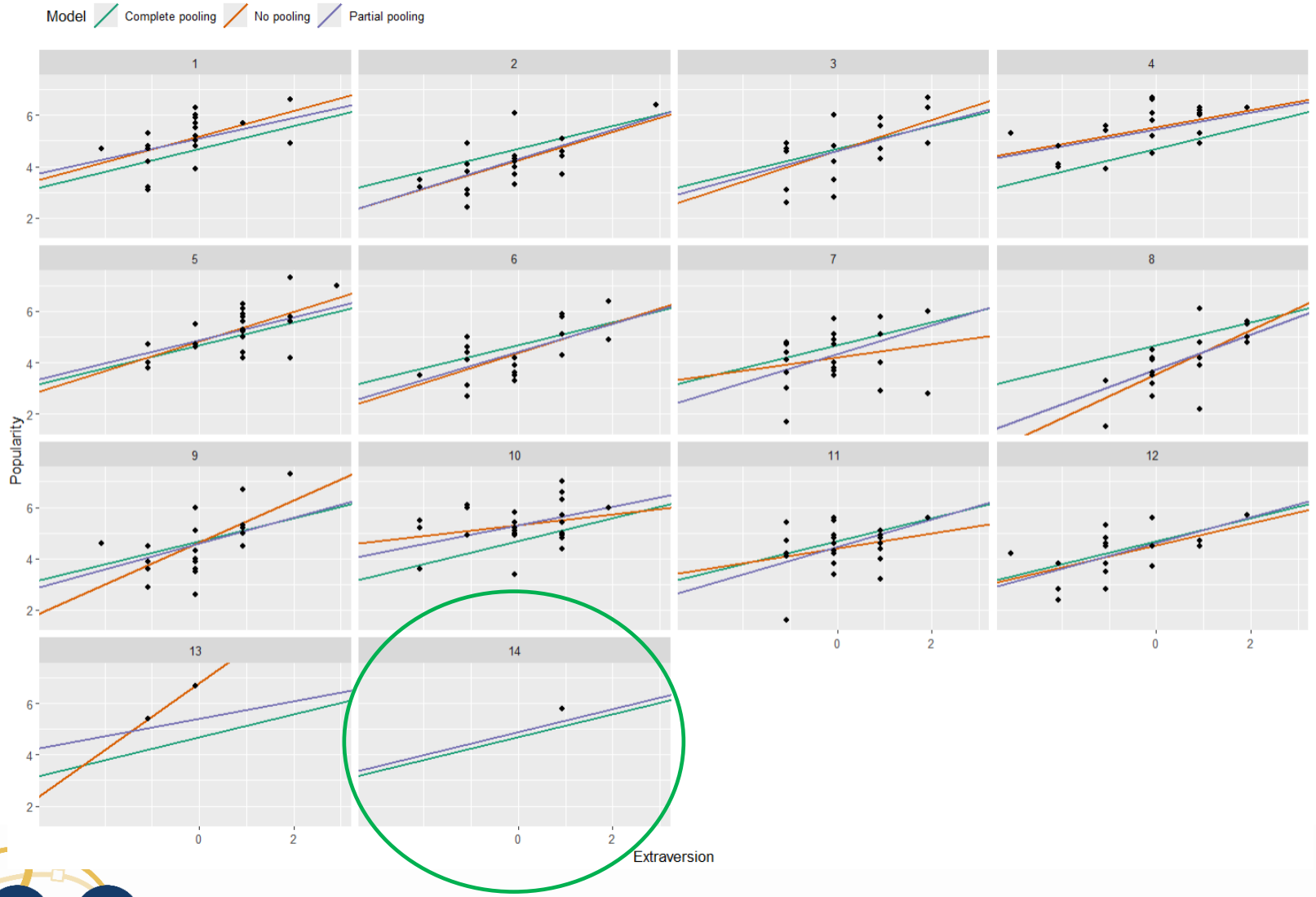


Tilburg University

tesc

Partial Pooling

- 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.

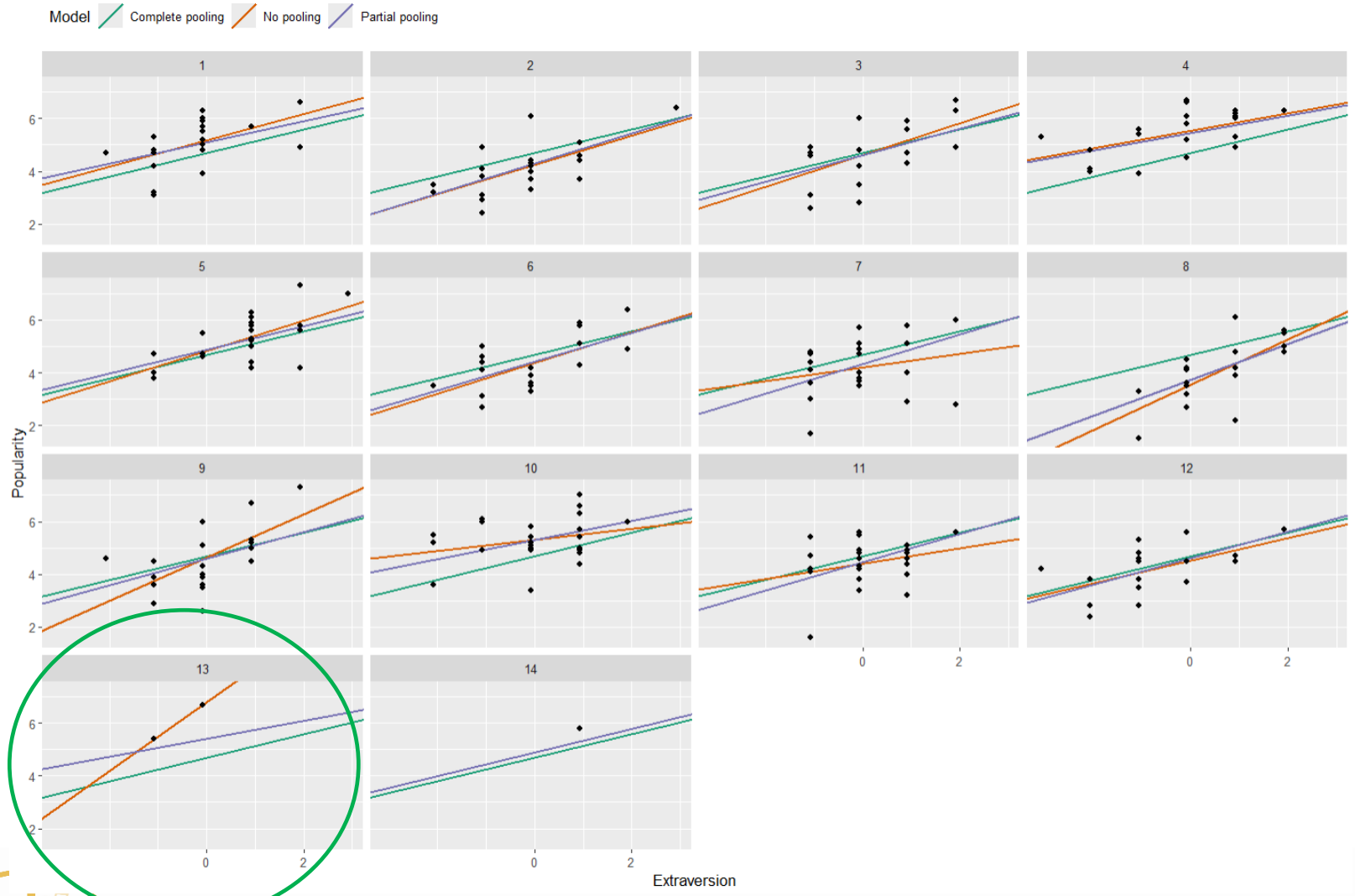


TILBURG UNIVERSITY

tesc

Partial Pooling

- 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.



TILBURG UNIVERSITY

tesc

Partial Pooling

- 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 the effect of Extraversion on Popularity the same in all schools?
  - Does Teacher Experience play a role?