1

Overview

- Goal: Learn to run basic (multilevel/hierarchical) regression in R.
  - Also focus on analysis of Intensive Longitudinal Data (ILD).
- Structure:
  - 3 days.
  - Alternating between lectures/discussion and practicals.

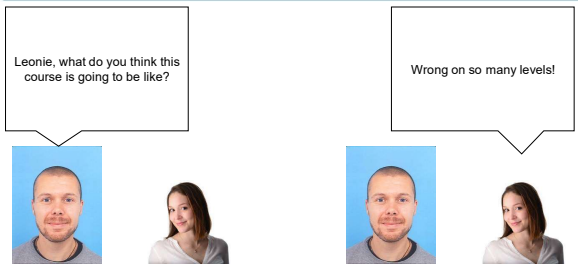
2

Topics

- Day 1:
  - (Multiple) Regression (the way I think about it).
  - How to approach a regression analysis….any analysis really.
  - Hierarchical data (what it is, what it implies, and what to do with it).
- Day 2 (common issues and special flavors of multilevel models).
- Day 3 (longitudinal and ILD data and whatever comes up during day 1 and 2).

3

The “Drill Sergeants”



4

Regression

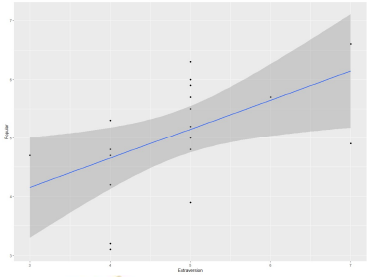
5

Regression

- What can you tell me about multiple regression?
  - Describe what it does.

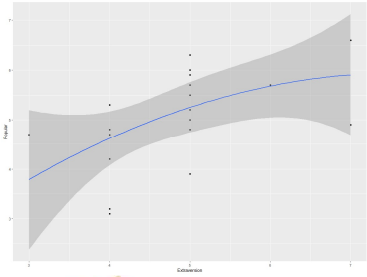
6

Regression



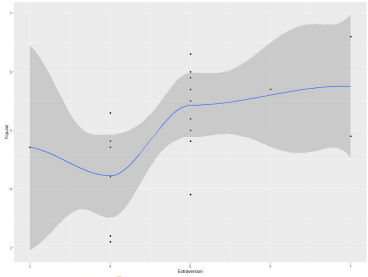
7

Regression



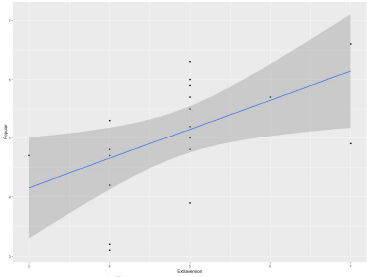
8

Regression



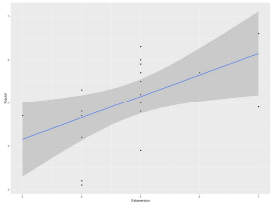
9

Regression → Curve Fitting!!



10

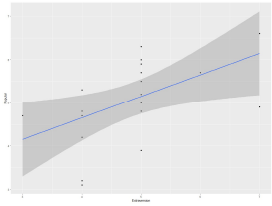
Curve Fitting



- Ok, so what are you saying with this model about your data?

11

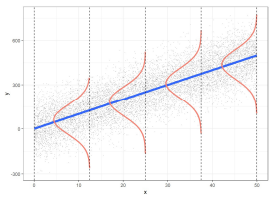
Curve Fitting



- Ok, so what are you saying with this model about your data?
  - **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 variance or, equivalently, the standard deviation of the responses is equal for all levels of X.

12

Curve Fitting

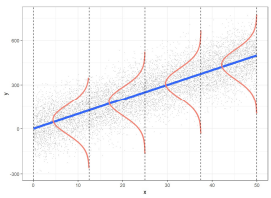


- Ok, so what are you saying with this model about your data?

- **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 variance or, equivalently, the standard deviation of the responses is equal for all levels of X.

13

Curve Fitting

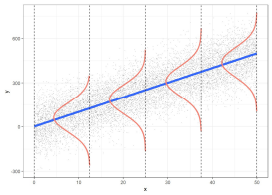


- **L:** There is a linear relationship between the mean response (Y) and the explanatory variable (X),
- **I:** The errors are independent—there's no connection between how far any two points lie from the regression line,
- **N:** The responses are normally distributed at each level of X, and
- **E:**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

What if one or more don't hold?

14

Curve Fitting



- **Take home:**

Regression ≠ confined to linear, normal, homoscedastic….or

independent

15

Visualize, Visualize, Visualize

- Visualization is like exercise; we know we should but typically don't.
- This is a bigger issue than we realize.
- We analyze data through modeling, but the result of that modeling is only useful if the model matches the data (or vice versa).
- You should spend several hours just visualizing and looking at your data!
  - And do more than just check scatterplots for linearity and outliers ;).

16

Current Practices

- Usually only provide descriptions of our data through tables with summary statistics (means, sd's).
- These summaries are only useful with symmetric distributions!
- Could provide info on median, mode, kurtosis, skew, but….who really get's that anyway?
  - Even more cognitively taxing when comparing multiple groups or multiple levels.
  - Two groups can have different means but the same mode, different modes but the same mean, or the same mean and standard deviation but a meaningful skew (Heino et al., 2019)

Tilburg University tesc

17

Example

Description:

Accelerometer data indicated, that girls were as active as boys (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

18

Example

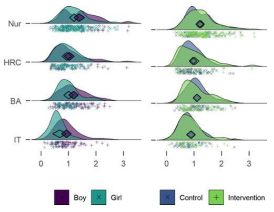
Gender	Arm	Nur	HRC	BA	IT
girl	control	M = 72.0; SD = 29.5; skewness = 0.9; kurtosis = 0.6; n = 104	M = 37.5; SD = 22.3; skewness = 0.8; kurtosis = 1.0; n = 90	M = 61.2; SD = 23.8; skewness = 0.8; kurtosis = 1.0; n = 53	M = 24.2; SD = 8.9; skewness = 0.4; kurtosis=−0.8; n = 14
girl	intervention	M = 71.8; SD = 28.4; skewness = 1.0; kurtosis = 1.6; n = 127	M = 52.0; SD = 23.6; skewness = 0.8; kurtosis=−0.2; n = 58	M = 56.7; SD = 22.8; skewness = 1.2; kurtosis = 0.7; n = 38	M = 36.1; SD = 22.1; skewness = 0.2; kurtosis=−0.8; n = 12
boy	control	M = 72.7; SD = 28.9; skewness = 0.3; kurtosis=−1.1; n = 21	M = 56.1; SD = 27.4; skewness = 1.5; kurtosis = 2.0; n = 55	M = 70.9; SD = 27.9; skewness = 0.7; kurtosis = 0.5; n = 72	M = 55.8; SD = 24.7; skewness = 1.1; kurtosis = 2.7; n = 82
boy	intervention	M = 89.6; SD = 35.1; skewness = 0.7; kurtosis = 0.2; n = 30	M = 71.1; SD = 43.9; skewness = 0.9; kurtosis = 0.5; n = 28	M = 72.8; SD = 31.0; skewness = 0.4; kurtosis=−0.8; n = 93	M = 54.1; SD = 25.1; skewness = 1.3; kurtosis = 1.5; n = 65

Table 2. Means, standard deviations and some distributional properties of a single variable in different educational tracks the participants were nested in. Nur = Practical nurse, HRC = Hotel, restaurant and catering studies, BA = Business and administration, IT = Business information technology

Tilburg University tesc

19

Example



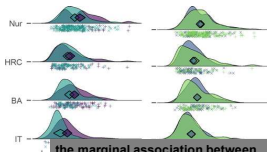
Notice:

- boys did more moderate-to-vigorous physical activity in every educational track.
- In spite of this, girls appeared more active when combining the educational tracks
- Reason: much more people in the practical nurse track, as well as those people being mostly girls.
- This is also known as Simpson's paradox, and is best investigated by visualizing data.

Tilburg University tesc

20

Example



the marginal association between two (categorical) variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.

Notice:

- boys did more moderate-to-vigorous physical activity in every educational track.
- In spite of this, girls appeared more active when combining the educational tracks
- Reason: much more people in the practical nurse track, as well as those people being mostly girls.
- This is also known as Simpson's paradox, and is best investigated by visualizing data.

Tilburg University tesc

21

Current Practices

- 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.

Tilburg University tesc

22

Practical Data Example

- 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 tesc

23

Getting to know your data.

Variable	n	mean	sd	min	max	range	se
Extraversion	20	4.85	0.99	3.0	7.0	4.0	0.22
Popular	20	5.08	0.95	3.1	6.6	3.5	0.21

Tilburg University tesc

24

Getting to know your data.

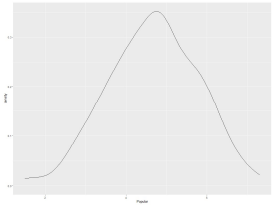
Variable	n	mean	sd	min	max	range	se
Extraversion	20	4.85	0.99	3.0	7.0	4.0	0.22
Popular	20	5.08	0.95	3.1	6.6	3.5	0.21

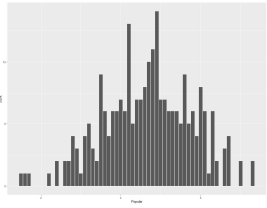
Variable	n	mean	sd	min	max	range	skew	kurtosis	se
Extraversion	20	4.85	0.99	3.0	7.0	4.0	0.59	0.11	0.22
Popular	20	5.08	0.95	3.1	6.6	3.5	-0.50	-0.58	0.21

Tilburg University tesc

25

Getting to know your data.

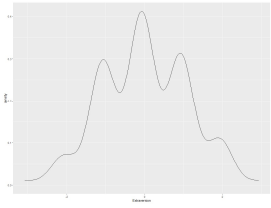


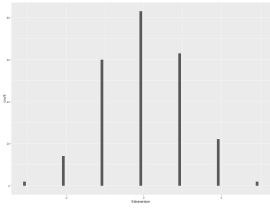


Tilburg University tesc

26

Getting to know your data.

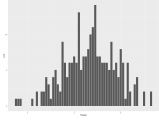


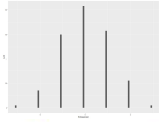


Tilburg University tesc

27

Getting to know your data.



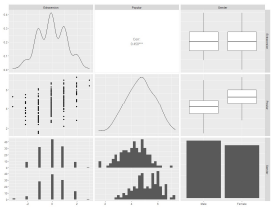


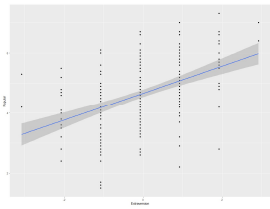
- Any issues?
- Can these pictures tell me anything I need to know for modeling?
- Did we actually test any of the **LINE** assumptions?

Tilburg University tesc

28

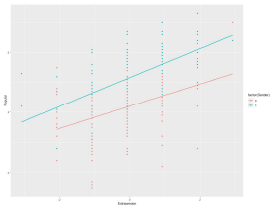
Getting to know your data.





29

Getting to know your data.



- Conclusions:
  - Popularity and Extraversion not-normal, but that is not necessarily a bad thing.
  - 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.

30

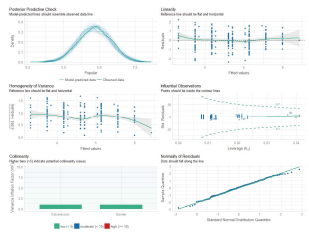
Intermediate conclusion

$$y = b_0 + b_1X_1 + b_2X_2 + e$$

$$Popular = b_0 + b_1Extraversion + b_2Gender + e$$

31

Getting to know your data.

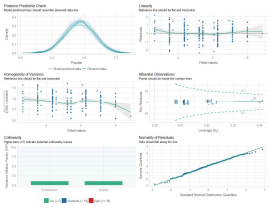


- Conclusions:
  - Homogeneity of Variance not ideal, but all else ok.
  - Now what?

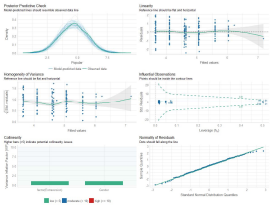
32

Getting to know your data.

Continuous



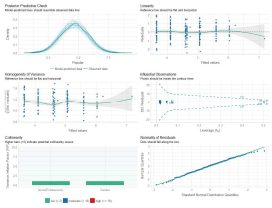
Categorical



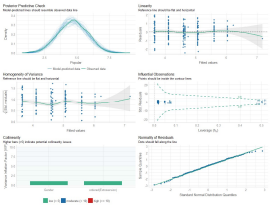
33

Getting to know your data.

Categorical



Ordinal



34

Robustness

- If you really worry about something in your data not matching your model….change the model.
- Heteroscedasticity? → Sandwich estimator, MELSM
- Non-normality → Choose different distribution
- Outliers → Choose t-distributed residuals
- Non-linearity → GAMM/GP/Polynomials
- Dependence → Model the dependence structure (e.g. with multilevel)

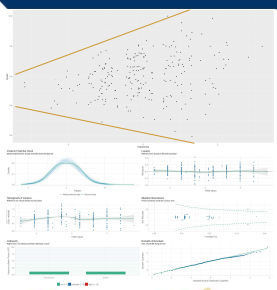
35

Robustness

- Never forget…you don't have to model all aspect of your sample data!
- Only what is relevant!

36

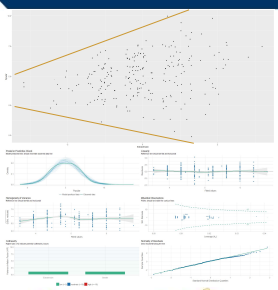
Robustness



- Residual Var = exp(Extraversion)
- True model:
  - Intercept = 4.5
  - $B_{extraversion}$ = .4
  - B_{gender} = 1
- Analyze with:
  - Normal analysis
  - Robust SE
  - Bayesian

37

Robustness



- Bayesian:

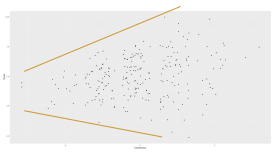
$$Popular = b_0 + b_1 Extraversion + b_1 Gender + e$$

$$e \sim N(0, sd_{res})$$

$$sd_{res} = b_{0\_res} + b_{1\_res} Extraversion$$

38

Robustness



Normal	Est.	s.e.	p
Intercept	4.347	.140	2^{-16}
Extraversion	.386	.091	3.3^{-5}
Gender	1.290	.202	8.6^{-10}

Robust	Est.	s.e.	p
Intercept	4.347	.150	2^{-16}
Extraversion	.386	.088	1.9^{-5}
Gender	1.290	.203	1.0^{-5}

Bayes	Est.	s.e.	p
Intercept	4.42	.14	
Extraversion	.42	.08	
Gender	1.16	.18	

39

Robustness

- Never forget…you don't have to model all aspect of your sample data!
- Only what is relevant!
- Somethings don't matter…but
  - Nobody completely knows which things don't ☹.
  - Subject matter knowledge is suuuuuuuper important.

40

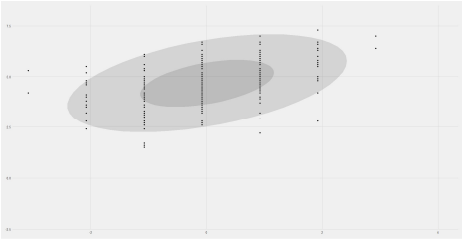
Robustness

- Heteroscedasticity? → Sandwich estimator, MELSM
- Non-normality → Choose different distribution
- Outliers → Choose t-distributed residuals
- Non-linearity → GAMM/GP/Polynomials
- Dependence → Model the dependence structure (e.g. with multilevel)

THESE ARE REASONS TO USE BAYESIAN STATISTICS/THE BRMS PACKAGE.

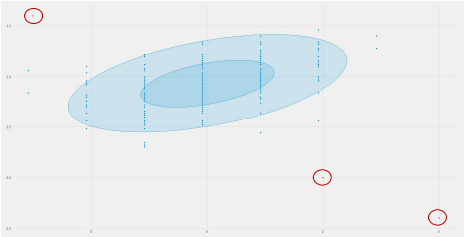
41

Example 2: Outliers



42

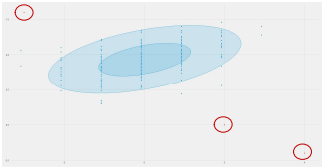
Example 2: Outliers



43

Example 2: Outliers

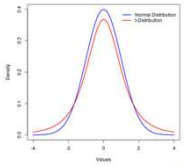
- The problem is that observations are further away than normal distribution "allows".

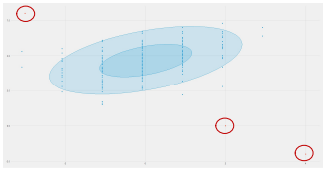


44

Example 2: Outliers

- The problem is that observations are further away from normal distribution “allows”.





45

Example 2: Outliers

Original	Est.	s.e.	p
Intercept	4.206	.077	2^{-16}
Extraversion	.430	.050	1.0^{-15}
Gender	0.961	.111	6.6^{-16}

Normal	Est.	s.e.	p
Intercept	4.160	.097	2^{-16}
Extraversion	.255	.061	3.8^{-5}
Gender	.964	.140	5.2^{-11}

Robust	Est.	s.e.	p
Intercept	4.160	.091	2^{-16}
Extraversion	.255	.125	.043
Gender	.964	.140	4.9^{-11}

Bayes (t-dist)	Est.	s.e.	p
Intercept	4.22	.08	
Extraversion	.42	.05	
Gender	.96	.12	

46

Conclusion 1

- Data analysis is all about modeling!
- Your model needs to represent the underlying structure in your data (as close as possible)
- So, think and look(!!) at your data…forget what different ‘types’ of analyses you know at first.
- So basically forget the structure of every stats course ever and forget SPSS 😉

47

Conclusion 1 - Addendum

- Never forget…you don’t have to model all aspect of your sample data!
- Only what is relevant!
- Somethings don’t matter…but
  - Nobody completely knows which things don’t ☹.
  - Subject matter knowledge is suuuuuuuper important.

48

Practical 1

49

Multilevel Analysis

50

~~Multilevel~~ Hierarchical Analyses

51

Recap

- Analyzing your data is all about modeling.
- If your data is linear, model it as such!
  - If it isn't…don't.

52

Recap

- 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

53

Recap

- 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

54

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(!!).

55

Example

- Let's start with an example,
- I have 58 pupils from three different classes
- Interested whether being more extraverted makes you more popular, and if there is a gender difference
- List of all the variables:
  - *pupil*: pupil identification variable
  - *class*: class identification variable
  - student-level independent variables: *extraversion* (continuous; higher scores mean higher extraversion) and *gender* (dichotomous; 0=male, 1 =female)
  - *popular:* continuous outcome variable at the student-level (higher scores indicate higher popularity).

56

Example

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

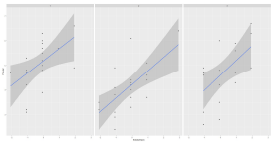
Pupil	Class	Extraversion	Gender	Popular
1	1	-0,08	1	6,30
2	1	1,92	0	4,90
3	1	-1,08	1	5,30
4	1	-2,08	1	4,70
5	1	-0,08	1	6,00
6	1	-1,08	0	4,70
7	1	-0,08	0	5,90
8	1	-1,08	0	4,20
9	1	-0,08	0	5,20
10	1	-0,08	0	3,90

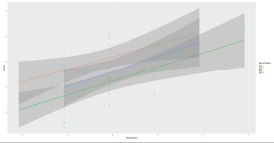
Tilburg University tesc

57

Some Options: Option 1

- Analyze all classes separately



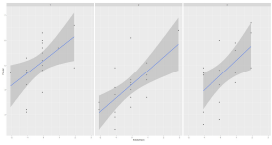


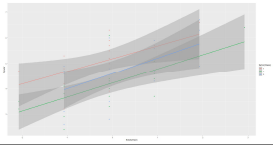
Tilburg University tesc

58

Some Options : Option 1

- Analyze all classes separately
- Benefits?
- Disadvantages?



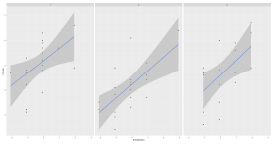


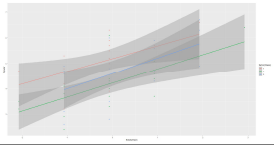
Tilburg University tesc

59

Some Options : Option 1

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





Tilburg University tesc

60

Richard McElreath

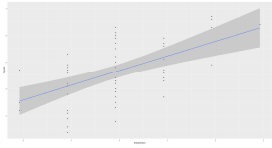
- *Many statistical models also have anterograde amnesia. As the models move from one cluster—individual, group, location—in the data to another, estimating parameters for each cluster, they forget everything about the previous clusters. They behave this way, because the assumptions force them to. Any 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*

Tilburg University tesc

61

Some Options: Option 2

- Ok…so…analyze together then?

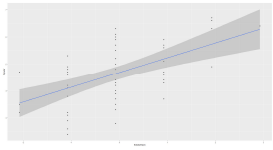


Tilburg University tesc

62

Some Options: Option 2

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

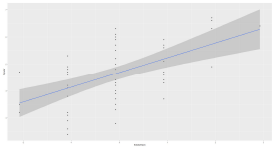


Tilburg University tesc

63

Some Options: Option 2

- Ok…so…analyze together then?
- What is the main issue in doing so?
- Hint: How much data do we have?



Tilburg University tesc

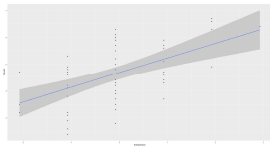
64

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?



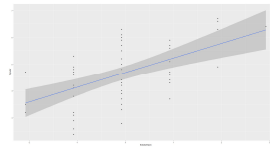
65

Some Options: Option 2

- Adjust se's

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

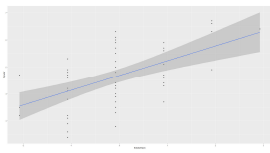
- Fortunately, smart ways to do this.
- Cluster robust s.e's



66

Some Options: Option 2

- Advantages?
- Disadvantages?



67

Some Options: Option 2

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

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

- 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

68

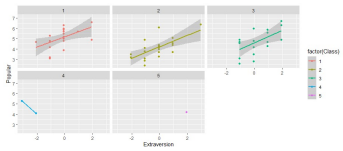
Some Options: Option 3 ---- Pool(ing) Party

- Who ever heard of partial-pooling?
- I'll add two classes to our dataset to illustrate it better
  - One class with two pupils
  - One class with one pupil

Tilburg University tesc

69

Pool(ing) Party

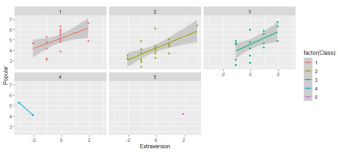


- What would be a problem if I analyzed everyone as N=1?

Tilburg University tesc

70

Pool(ing) Party

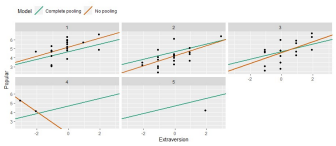


- Each panel shows an independently estimated regression line.
- This approach of fitting a separate line for each class is sometimes called the no pooling model
- Information from different classes is NOT combined or pooled together.

Tilburg University tesc

71

Pool(ing) Party

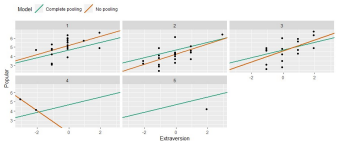


- Could also combine information from all classes (like we saw earlier).
- Fit a single line for the combined data set, unaware that the data came from different participants.

Tilburg University tesc

72

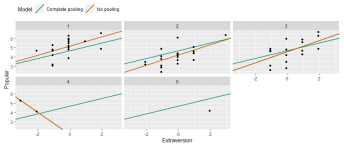
Pool(ing) Party



- Notice anything?
- Advantages?
- Disadvantages?

73

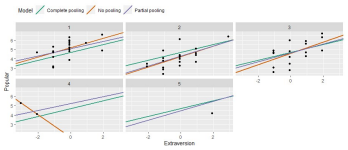
Pool(ing) Party



- 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.

74

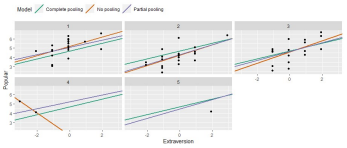
Pool(ing) Party



- 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.

75

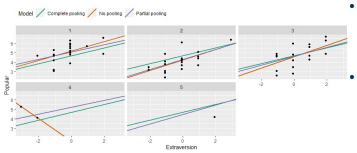
Pool(ing) Party



- Most of the time, the no pooling and partial pooling lines are the same.
- When the two differ, it's because the partial pooling model's line is pulled slightly towards the complete-pooling line.

76

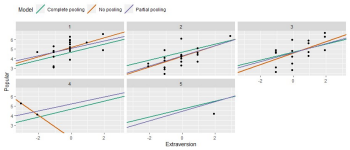
Pool(ing) Party



- Amount of pull toward no-pooling line depends on amount of data.
- Also depends on how extreme a person is.

77

Pool(ing) Party



- Advantages?
- Disadvantages?

78

Partial Pooling = Multilevel

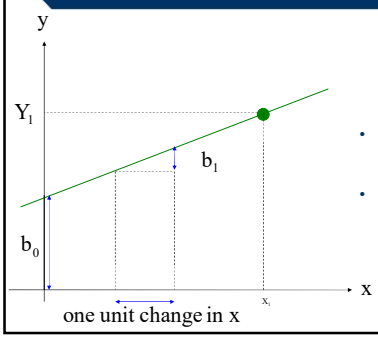
- So that's the beauty of Multilevel:
  - We get information about individual units (like in separate analyses).
  - But, we don't have "amnesia".
- And can statistically compare for differences between units.
  - Is extraversion equally important in all classes?
  - Does experience of the teacher also play a role?

79

Hierarchical Analyses: The Equations

80

Multiple Regression

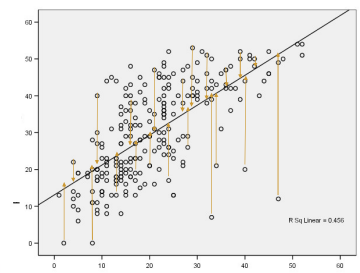


$$\hat{Y}_i = b_0 + b_1 X_i$$

- Intercept (b_0): predicted value for Y when (all) X = 0
- Slope(b_1): The amount of predicted change in Y when X changes by 1

81

Multiple Regression



Minimize the sum of squares of the errors of prediction:

$$\sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2$$

82

Example

Levels:

Schools: 1, 2, 3, n

Pupils: 1 2 n | 1 2 n | 1 2 n | 1 2 n

- Interested in the relation between Popularity (Pop) and pupils Extraversion
- In addition, we want to take "years of teaching experience" into account.
  - we suspect that students with a more experienced teacher will be seen as more equal despite differences in extraversion.

83

Example

Levels:

Schools: 1, 2, 3, n

Pupils: 1 2 n | 1 2 n | 1 2 n | 1 2 n

Variable:	**Level:**
Popularity	Pupil (level 1)
Extraversion	Pupil (level 1)

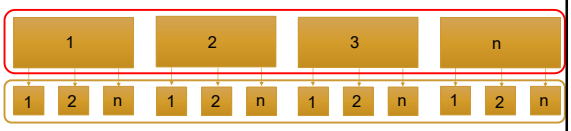
84

Example

Levels:

Schools

Pupils



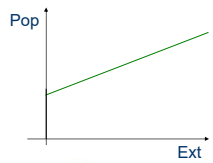
Variable:	**Level:**
Popularity	Pupil (level 1)
Extraversion	Pupil (level 1)
Teacher Experience	School (level 2)

Tilburg University tesc

85

Multilevel Regression

Level 1: *Extraversion* (*Ext*) as predictor of *Popularity* (*Pop*)



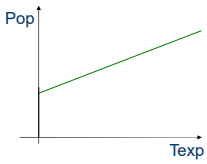
$$Pop_{ij} = b_0 + b_1 Ext_{ij} + e_{ij}$$

Tilburg University tesc

86

Multilevel Regression

Level 2: *Teacher Experience* (*Texp*) as predictor of *Popularity* (*Pop*)



$$\cancel{Pop_{ij} = b_{0\,class} + b_{1\,class}\, Texp_j + uj}$$

$$\cancel{\overline{Pop_j} = b_{0\,class} + b_{1\,class}\, Texp_j + u_j}$$

Tilburg University tesc

87

Multilevel Regression

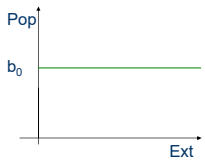
- Level 1: $Pop_{ij} = b_0 + b_1 Ext_{ij} + e_{ij}$
- Level 2: $\overline{Pop_j} = b_{0\,class} + b_{1clas}\, Texp_j + u_j$

Tilburg University tesc

88

Multilevel Regression

- Level 1:



$$Pop_{ij} = b_0 + e_{ij}$$

$$b_0 = \overline{Pop}$$

89

Multilevel Regression

Level 1: $Pop_{ij} = \boxed{b_0} + b_1 Ext_{ij} + e_{ij}$

Level 2: $\overline{Pop_j} = b_{0\,class} + b_{1class} Texp_j + u_j$

90

Multilevel Regression

Level 1: $Pop_{ij} = \boxed{b_{0j}} + b_1 Ext_{ij} + e_{ij}$

Level 2: $b_{0j} = b_{0\,class} + b_{1class} Texp_j + u_j$

91

Multilevel Regression

Level 1: $Pop_{ij} = b_{0j} + b_1 Ext_{ij} + e_{ij}$

Level 2: $b_{0j} = b_{0\,class} + b_{1\,class} Texp_j + u_j$

Combined:

$$Pop_{ij} = b_{0j} + b_1 Ext_{ij} + ei_j$$
$$= \boxed{b_{0\,class}} + \boxed{b_1} Ext_{ij} + b_{1cla}\ \ Texp_j \boxed{+ e_{ij} + u_j}$$

92

Multilevel Regression

Level 1: $Pop_{ij} = b_{0j} + b_1 Ext_{ij} + e_{ij}$

Level 2: $b_{0j} = \gamma_{00} + \gamma_{01} Texp_j + u_{0j}$

Combined:

$$Pop_{ij} = b_{0j} + b_1 Ext_{ij} + eij$$
$$= \gamma_{00} + b_1 Ext_{ij} + \gamma_{01} Texp_j + e_{ij} + u_{0j}$$

Tilburg University tesc

93

Multilevel Regression

Level 1: $Pop_{ij} = b_{0j} + b_{1j} Ext_{ij} + e_{ij}$

Level 2: $b_{0j} = \gamma_{00} + \gamma_{01} Texp_j + u_{0j}$

$b_{1j} = \gamma_{01} + u_{1j}$

Combined:

$$Pop_{ij} = b_{0j} + b_{1j} Ext_{ij} + eij$$
$$= \gamma_{00} + \gamma_{01} Ext_{ij} + \gamma_{01} Texp_j + e_{ij} + u_{0j} + u_{1j} Ext_{ij}$$

Tilburg University tesc

94

Multilevel Regression

- Can you relate these equations back to partial pooling?

Tilburg University tesc

95

Practical 2

Tilburg University tesc

96