Methods 3: Multilevel Statistical Modeling and Machine Learning

Week 02: *Multilevel linear regression*

September 10, 2024

The course plan

Week 1: *Introduction*

Instructor sessions: *Setting up* R *and* Python *and recollection of the general linear model*

Week 2: *Multilevel linear regression*

Instructor sessions: *Modelling subject level effects – and how do they differ from group level effects?*

Week 3: *Link functions and fitting generalised linear multilevel models*

Instructor sessions: *What to do when the response variable is not continuous?*

Week 4: *Evaluating Generalised linear mixed models*

Instructor sessions: *How do we assess how models compare to one another?*

Week 5: *Explanation and Prediction*

Instructor sessions: *Code review*

Week 6: *Mid-way evaluation and Machine Learning Intro*

Instructor sessions: *Getting Python Running*

Week 7: *Linear regression revisited (machine learning)*

Instructor sessions: *How to constrain our models to make them more predictive*

Week 8: *Logistic regression revisited (machine learning)*

Instructor sessions: *Categorizing responses based on informed guesses*

Week 9: *Dimensionality Reduction, Principled Component Analysis (PCA)*

Instructor sessions: *What to do with very rich data?*

Week 10: *Outlook, unsupervised classification and neural networks*

Instructor sessions: *Data with no labels and networks*

Week 11: *Organising and preprocessing messy data*

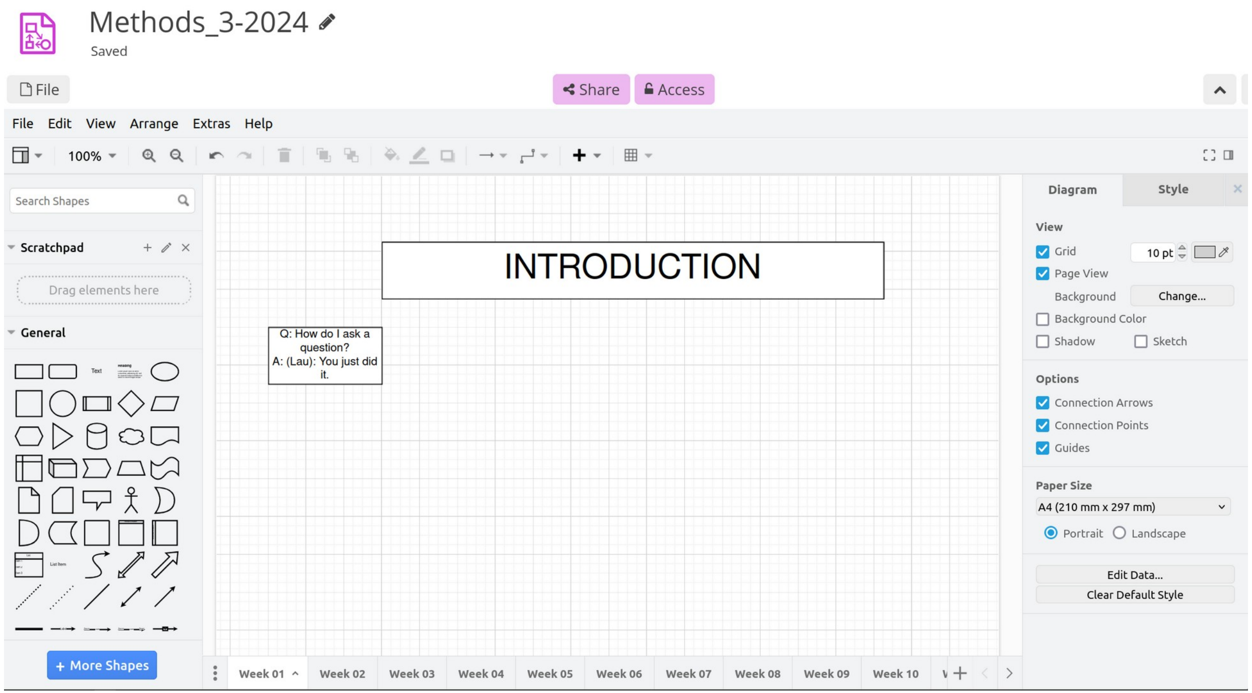
Instructor sessions: *Code review*

Week 12: Final evaluation and wrap-up of course

Instructor sessions: *Ask anything!*

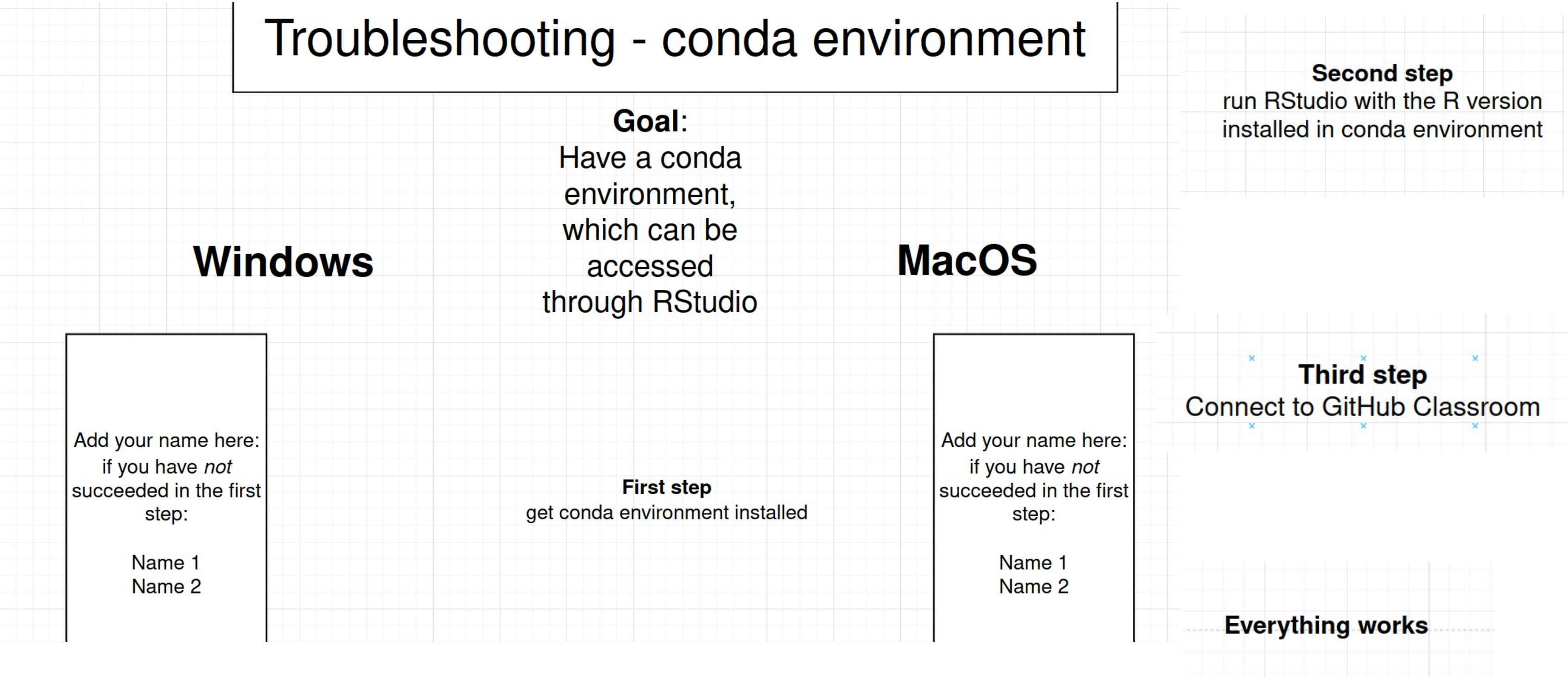
CryptPad

https://cryptpad.fr/diagram/#/2/diagram/edit/e2J7ywc5mVeuAmudDDsnkR7U/



Troubleshooting

CONDA ENVIRONMENTS



Learning goals and outline

Multilevel linear regression

1) Understanding the motivations for doing multilevel modelling
   - Using all the data
   - Respecting the data distribution
2) Understanding the basics of multilevel modelling
   - Different kinds of design matrices, X and Z
   - Variance-covariance matrices
   - Pooling: complete pooling, no pooling and partial pooling

Overall motivation

We want to use all the information in the data while respecting the data distributions the data is generated from

The four classical levels of variables

- Nominal
  - examples: true/false, correct/incorrect, female/male, dog/cat, apples/pear, also called *categorical*
  - they are ***names*** of categories, but it does not make sense to order them
- Ordinal
  - examples: senior/junior, adult/child 1/2/3
  - they are also ***names*** of categories, but there is an explicit or implicit ***ordering***, i.e. one is greater than another
- Interval
  - examples: the year 1984 AD; the temperature 100 °C
  - they ~~are~~ can be ***continuous***, and there is ***ordering***, e.g. 100 °C > 90 °C. And intervals can be compared, e.g. the interval from 80 °C to 100 °C is as long as the one from 40 °C to 60 °C
  - crucially, there is no real 0; the year before 1 AD is not characterised by absence of time; and 0 °C is not characterised by the absence of temperature This means that we cannot say that, say, 40 °C is twice as high a temperature as 20 °C
- Ratio
  - examples: the temperature 273 K, the reaction time of a subject
  - they ~~are~~ can be ***continuous***, there is ***ordering*** and there is a ***real 0***.
  - Thus we can say that 200 K is twice the temperature of 100 K, as 0 K *is* the absence of temperature; and we can say that subject 2, 400 ms, is twice as fast as subject 1, 200 ms, because 0 ms *is* the time when the event happened

SLIDE UPDATED 2024_10_03

Sub-optimal practices

TREATING VARIABLES AS ANOTHER LEVEL THAN THEY ARE

Journal of Sleep Research

Full Access

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

Gregory Belenky, Nancy J. Wesensten, David R. Thorne, Maria L. Thomas, Helen C. Sing, Daniel P. Redmond, Michael B. Russo, Thomas J. Balkin

First published: 21 February 2003 |
https://doi-org.ez.statsbiblioteket.dk/10.1046/j.1365-2869.2003.00337.x | Citations: 960

: Gregory Belenky, MD, Colonel, Medical Corps, U.S. Army, Division of Neuropsychiatry, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910-7500, USA. Tel.: +1-301-319-9085; fax: +1-301-319-9255; e-mail: gregory.belenky@na.amedd.army.mil

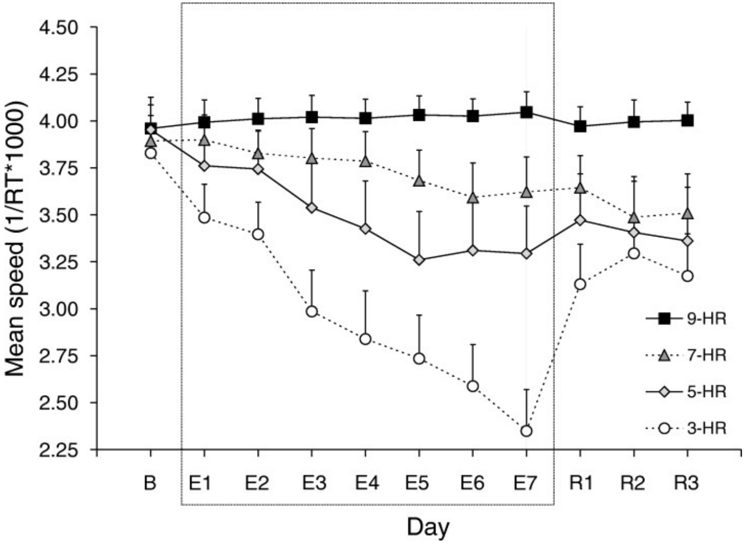
sleepstudy {lme4} R Documentation

Reaction times in a sleep deprivation study

Description

The average reaction time per day (in milliseconds) for subjects in a sleep deprivation study.

Days 0-1 were adaptation and training (T1/T2), day 2 was baseline (B); sleep deprivation started after day 2.



Q: judging from the figure: what kind of predictor is *Day* modelled as?
And should it have been modelled otherwise?

Sub-optimal practices
AGGREGATION – part 1

AMERICAN PSYCHOLOGICAL ASSOCIATION
Journal of Experimental Psychology: General

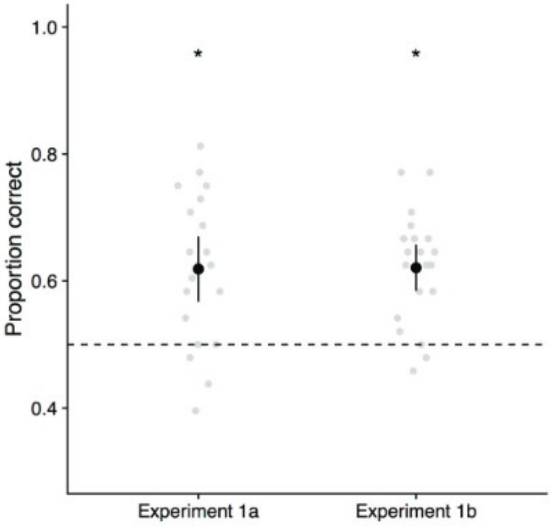
ISSN: 0096-3445
https://doi.org/10.1037/xge0001091

Effects of Statistical Learning in Passive and Active Contexts on Reproduction and Recognition of Auditory Sequences

Saloni Krishnan[1], Daniel Carey[2], Frederic Dick[3, 4], and Marcus T. Pearce[5, 6]
[1] Department of Psychology, Royal Holloway, University of London
[2] Novartis Pharmaceuticals, Dublin, Ireland
[3] Department of Psychological Sciences, Birkbeck, University of London
[4] Department of Experimental Psychology, University College London
[5] School of Electronic Engineering and Computer Science, Queen Mary University of London
[6] Department of Clinical Medicine, Aarhus University

Figure 3

Mean Proportions of Correct Responses for the Recognition Task in Experiments 1a (Familiarization Only) and 1b (Reproduction Only)



Note. Chance performance is at 0.5, shown by the horizontal line, and asterisks indicate performance differing significantly from chance. Error bars represent 95% confidence intervals around the mean.

Q's:

- What do each of the grey dots represent?
- What kind of response data is present at the single-trial level?
  - And how is the response data categorised here?
- Within a single-subject level framework, what would be the appropriate model to fit?
- Can confidence intervals extend beyond proportion correct 1 or 0 with the approach used here?

Sub-optimal practices

AGGREGATION – part 2


Cognitive Brain Research
Volume 7, Issue 4, March 1999, Pages 493-501

Research report

`Paradoxical' alpha synchronization in a memory task

W. Klimesch, M. Doppelmayr, J. Schwaiger, P. Auinger, Th. Winkler

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/S0926-6410(98)00056-1

Get rights and content

“Three factorial ANOVA's were calculated for hits and correct rejections [...]. For these [electrode] sites, data were averaged separately for the left and right hemisphere but only for those 1 s intervals of an epoch which represent the first 1000 ms after presentation onset of the memory set and frame.” section 2.4.5

“*The percentage of epochs that were excluded from data analysis (due to artifacts) for hits and correct rejections ranged from 28% to 42% in the four experimental conditions.*” section 2.4.2

Q: What may be a problem when data is aggregated like that?

General Linear Model

When all you have is a ~~hammer~~, everything looks like a ~~nail~~

Normally distributed variable (an interval or ratio variable)

W02_live_coding

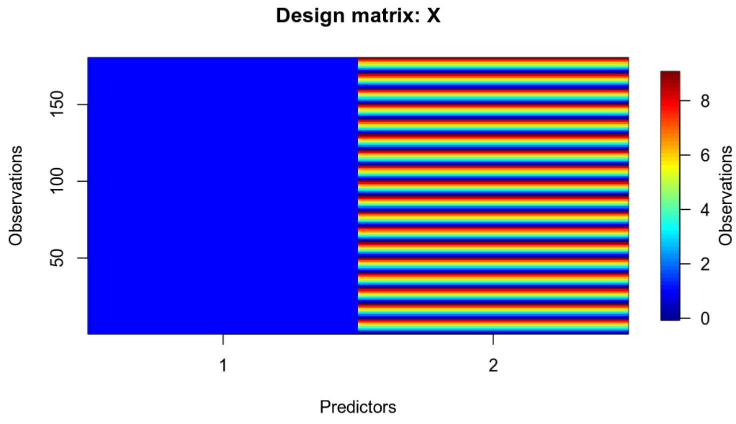
Understanding the basics of multilevel modelling

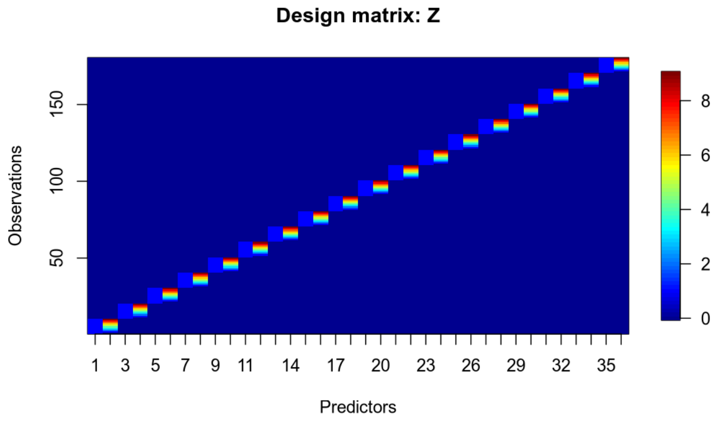
Design matrices

Level 1 **Level 2**

$$Y = \boxed{X\beta} + \boxed{Zb} + \epsilon$$

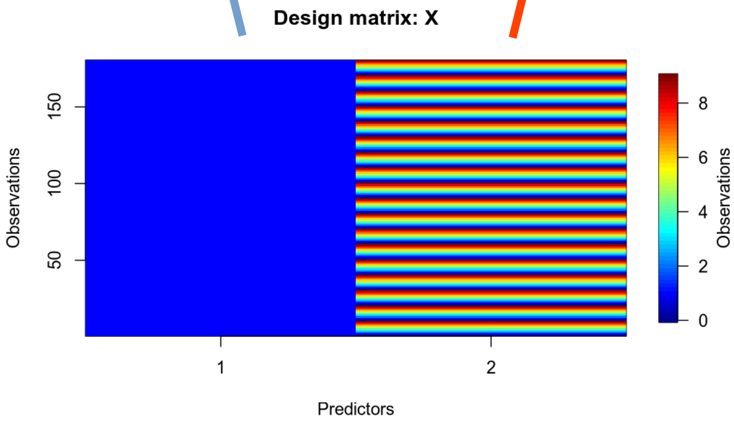
Fixed effects **Random effects**

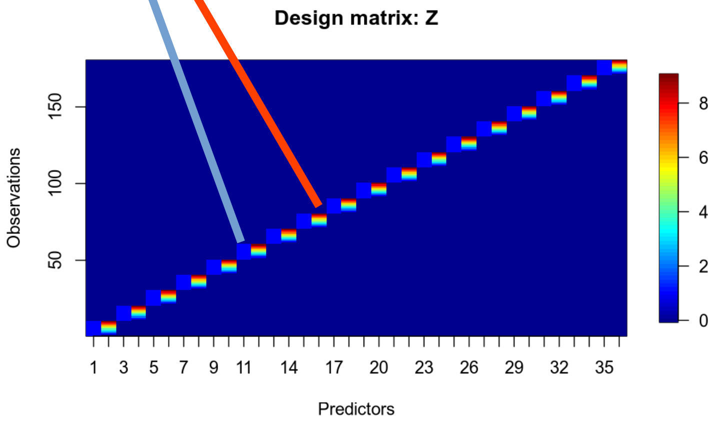




```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups   Name        Std.Dev. Corr
 Subject  (Intercept) 24.741
          Days         5.922   0.07
 Residual             25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

$$\hat{\sigma}; \epsilon \sim N\left(0, \sigma^2\right)$$





Variance-covariance matrices

```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups   Name        Std.Dev. Corr
 Subject  (Intercept) 24.741
          Days         5.922   0.07
 Residual             25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

```
$Subject
    (Intercept)        Days
308   2.2585509   9.1989758
309 -40.3987381  -8.6196806
310 -38.9604090  -5.4488565
330  23.6906196  -4.8143503
331  22.2603126  -3.0699116
332   9.0395679  -0.2721770
333  16.8405086  -0.2236361
334  -7.2326151   1.0745816
335  -0.3336684 -10.7521652
337  34.8904868   8.6282652
349 -25.2102286   1.1734322
350 -13.0700342   6.6142178
351   4.5778642  -3.0152621
352  20.8636782   3.5360011
369   3.2754656   0.8722149
370 -25.6129993   4.8224850
371   0.8070461  -0.9881562
372  12.3145921   1.2840221
with conditional variances for "Subject"
```

Q: What are the means of these columns?

$$B \sim N(0, \Sigma)$$

$$\hat{\Sigma} = \begin{matrix} 24.741^2 & 9.61 \\ 9.61 & 5.922^2 \end{matrix}$$

$$Corr = \frac{\hat{\Sigma}_{(1,2)}}{\sqrt{\left(\hat{\Sigma}_{(1,1)} \hat{\Sigma}_{(2,2)}\right)}} = 0.07$$

Variance-covariance matrix

$$B \sim N(0, \Sigma) \quad (3)$$

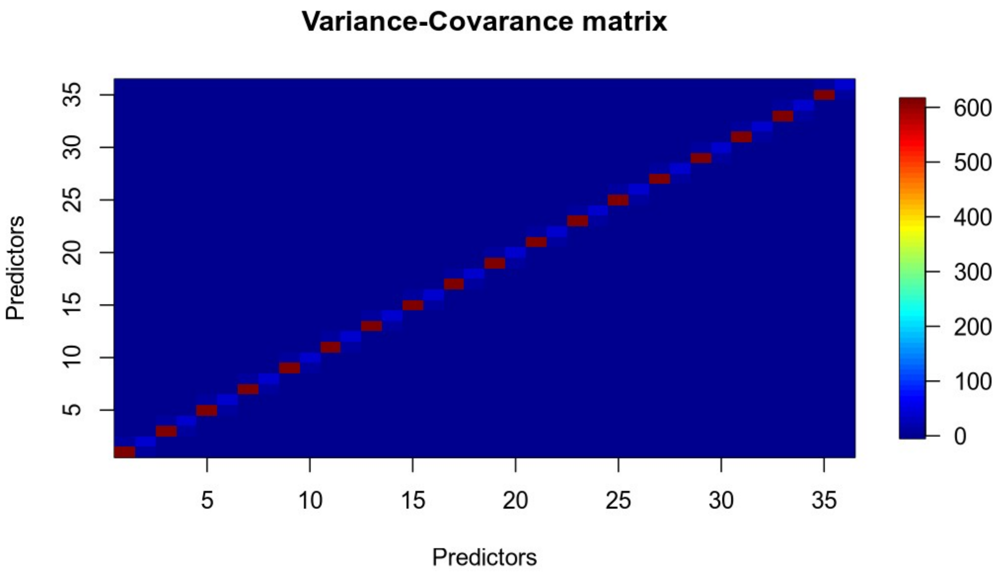
$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^T \quad (4)$$

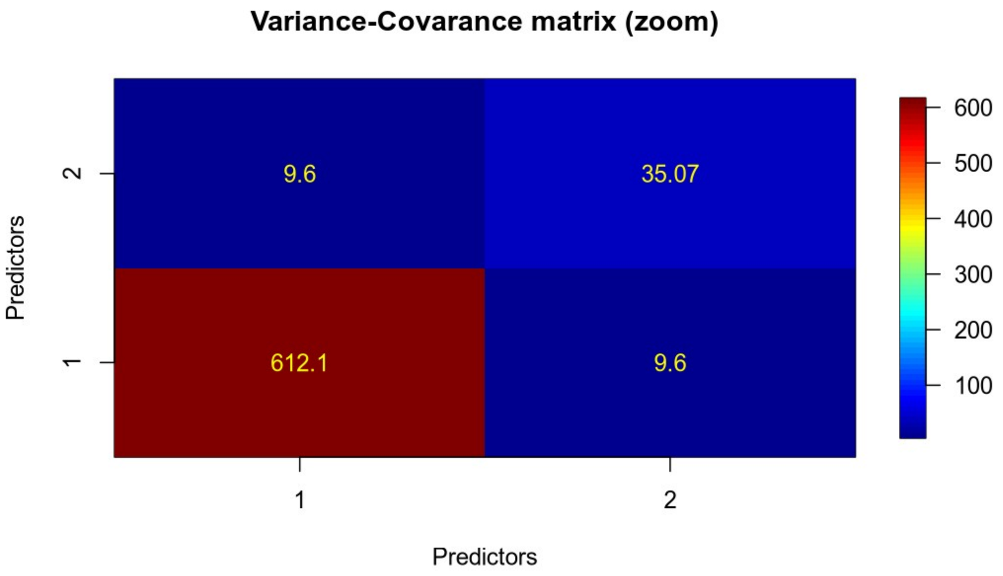
Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67:1–48. https://doi.org/10.18637/jss.v067.i01

$\hat{\theta} =$

```
Subject.(Intercept) Subject.Days.(Intercept)              Subject.Days
         0.96674177               0.01516906                0.23090995
```

$\hat{\Sigma}_{\theta}$





Q: What would the plot look like if there was no covariance between predictors?

$$corr = \frac{9.6}{\sqrt{(612.1 * 35.07)}} = 0.07$$

Why do we care about covariance?

Too much covariance between predictors, and your model doesn’t converge ...

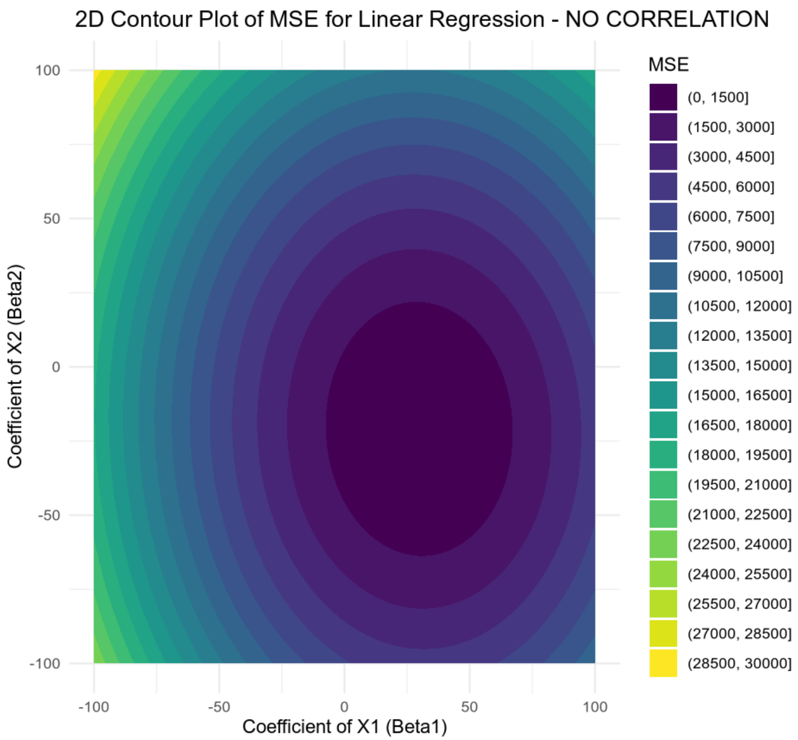
... and even in you get convergence, the estimated coefficients may be sensitive to small changes in input

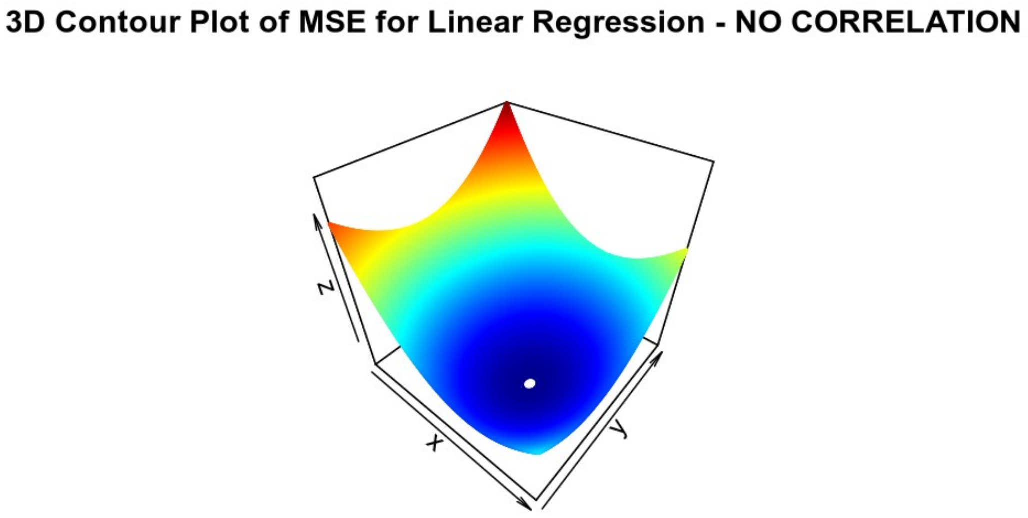
```r
get.SIGMA <- function(model) sigma(model)^2 * (getME(model, 'Lambda') %*% getME(model, 'Lambdat'))

for(seed in c(1, 7))
{
    print(paste('Random seed is:', seed))
    set.seed(seed)
    sleepstudy$Days2 <- 2 * sleepstudy$Days + rnorm(length(sleepstudy$Days))
    model <- lmer(Reaction ~ Days + Days2 + (Days + Days2 | Subject), data=sleepstudy)
    print(paste('Estimated rank is:', rankMatrix(get.SIGMA(model))[1]))
    print('\n')
}
```

```
[1] "Random seed is: 1"
boundary (singular) fit: see help('isSingular')
[1] "Estimated rank is: 36"
[1] "\n"
[1] "Random seed is: 7"
Advarsel i checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.0433554 (tol = 0.002, component 1)
[1] "Estimated rank is: 54"
[1] "\n"
```

True function: $y = 31x_1 - 21x_2$

$x_1 = rnorm(100); x_2 = rnorm(100)$

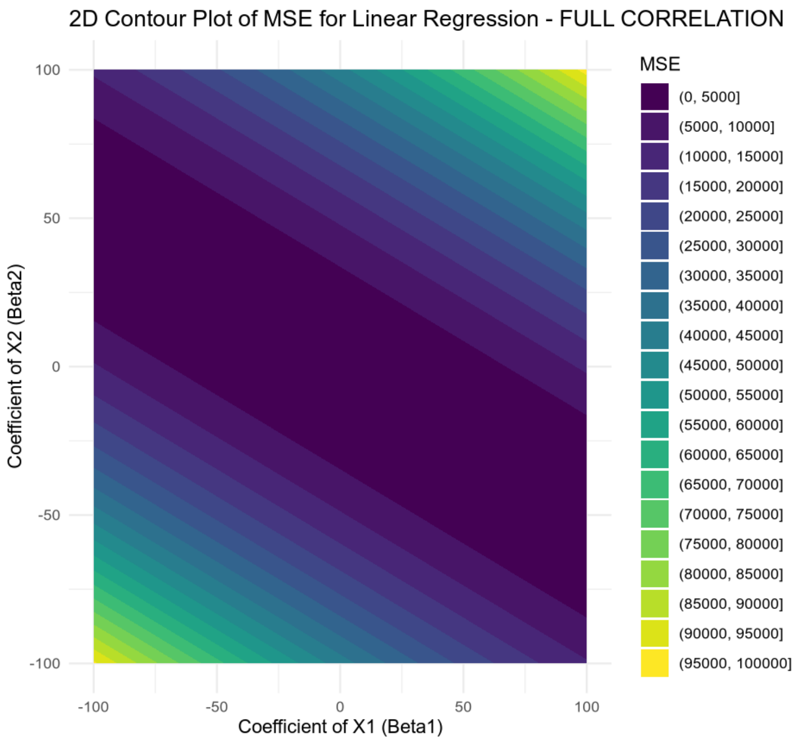


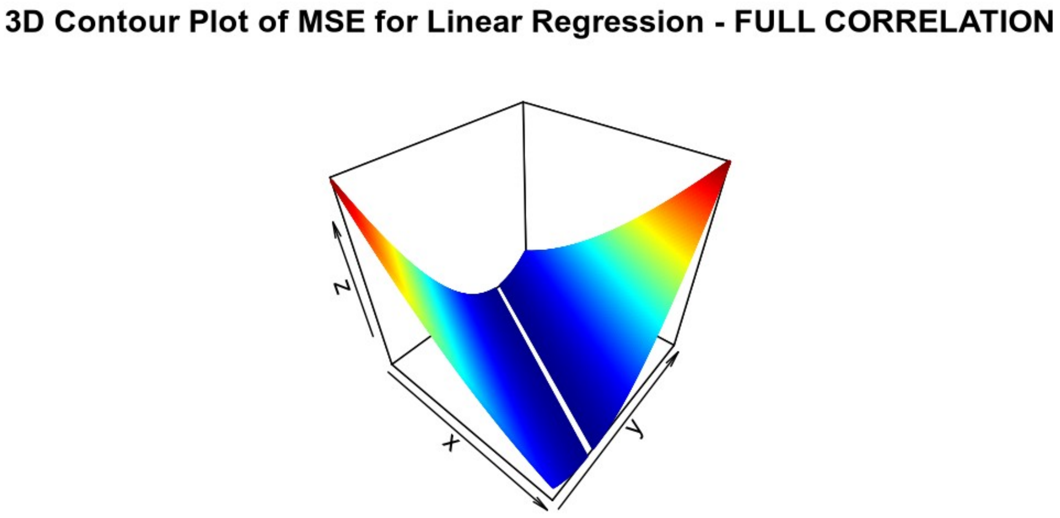


Full rank

True function: $y = 31x_1 - 21x_2$

$x_1 = rnorm(100);\ x_2 = 2x_1$





Singular

Hint for the future

With “big data”, the variance-covariance matrices will matter a lot (machine learning)

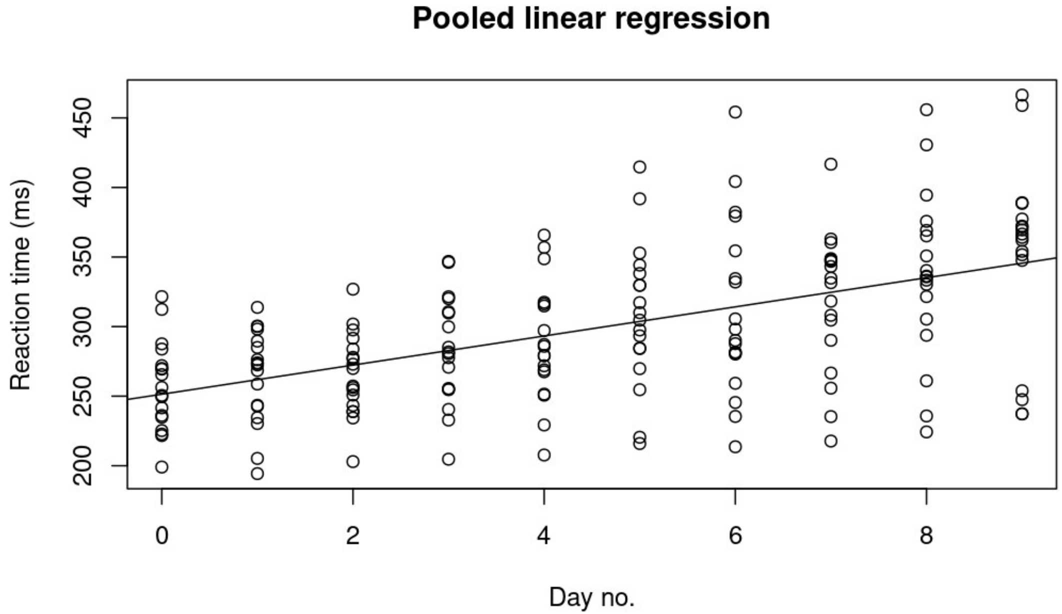
Pooling

Overview – pooling

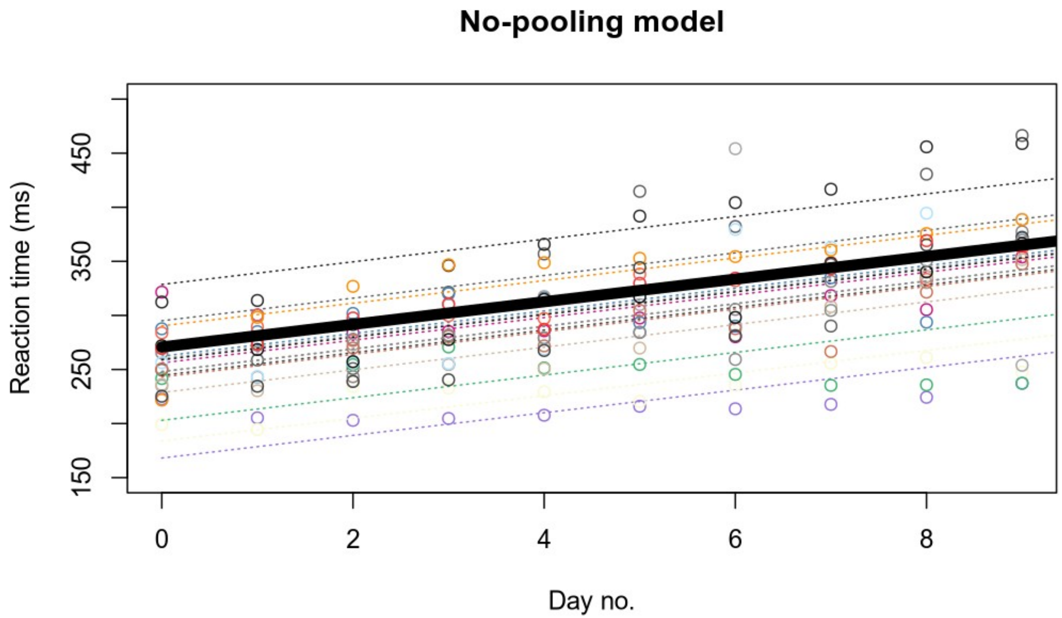
- Complete pooling
  - Ignores the categorical predictor, e.g. *Subject*, *altogether*
  - *lm(Reaction ~ Days)*
- No pooling
  - Overfits the categorical predictor, e.g. *Subject*, i.e. overstates the variation among *Subjects*
  - *lm(Reaction ~ Days * Subject – 1); models slopes and intercepts for each subject*
  - *lm(Reaction ~ Days + Subject – 1); models intercepts for each subject*
- Partial pooling
  - A compromise between the two extremes above. If a group, e.g. *Subject*, has few observations (high variance), it will be shrunk towards the overall mean. If *Subject* has many observations (low variance), it will be shrunk less towards the overall mean
  - *lmer(Reaction ~ Days + (Days | Subject) # models slopes and intercepts for each subject*
  - *lmer(Reaction ~ Days + (1 | Subject) # models intercepts for each subject*

SLIDE UPDATED 2024_10_03

Complete pooling

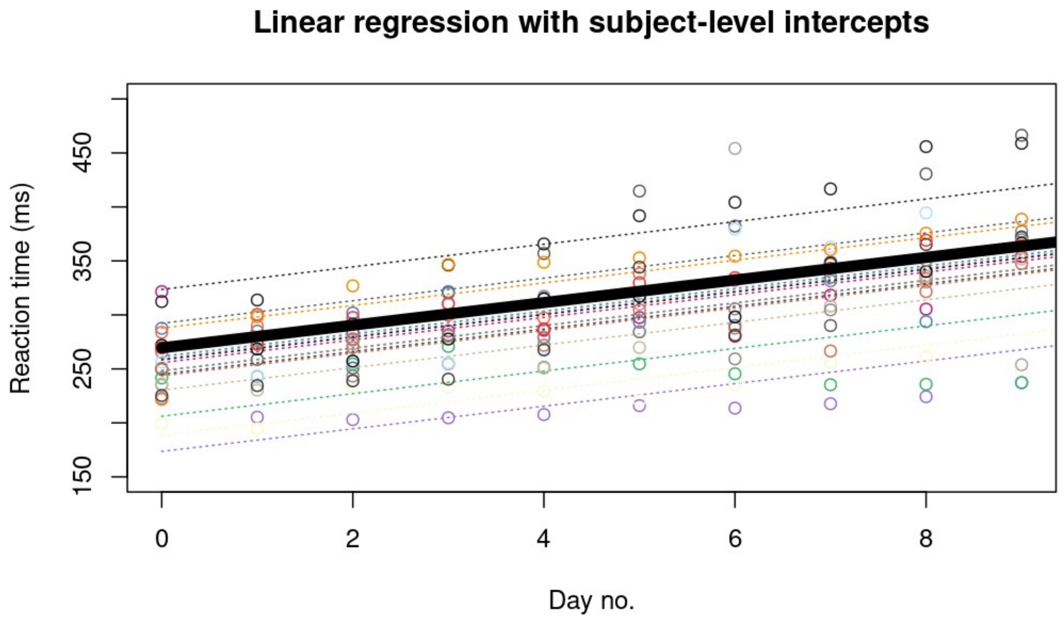


No pooling



No-pooling model
Reaction time (ms)
150
250
350
450
0
2
4
6
8
Day no.

Partial pooling



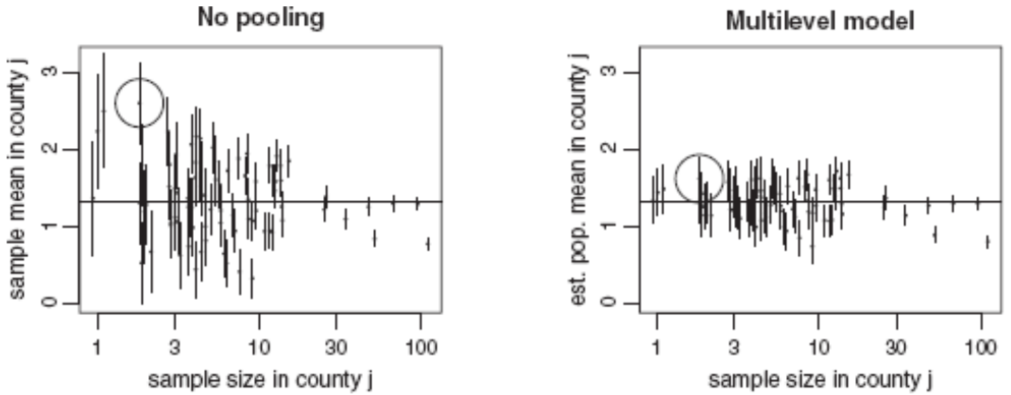


Figure 12.1 *Estimates ± standard errors for the average log radon levels in Minnesota counties plotted versus the (jittered) number of observations in the county: (a) no-pooling analysis, (b) multilevel (partial pooling) analysis, in both cases with no house-level or county-level predictors. The counties with fewer measurements have more variable estimates and larger higher standard errors. The horizontal line in each plot represents an estimate of the average radon level across all counties. The left plot illustrates a problem with the no-pooling analysis: it systematically causes us to think that certain counties are more extreme, just because they have smaller sample sizes.*

Gelman A, Hill J (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press

How is shrinking done?

HOW IS BIAS ADDED?

Penalised least squares:

$$r^2(\theta, \beta, u) = \rho^2(\theta, \beta, u) + \|u\|^2 \qquad (14)$$

$\rho^2(\theta, \beta, u)$ is the (weighted) residual sum of squares

$$\mu_{Y|U=u} = X\beta + Z\Lambda_\theta u$$

(compare with: $\mu_Y = X\beta$ from classical regression)

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67:1–48. https://doi.org/10.18637/jss.v067.i01

Another way to look at shrinking

OR HOW TO ADD BIAS

$$\alpha_j = \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \left(\bar{y}_j - \beta \bar{x}_j\right) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \mu_\alpha \qquad (12.4)$$

α_j: partially pooled response
$\left(\bar{y}_j - \beta \bar{x}_j\right)$: subject estimate of mean
μ_α: group estimate of mean
σ_y^2: within-group variance
σ_α^2: between-group variance

$$\frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} = 1$$

$\frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$: proportion assigned to subject estimate

$\frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$: proportion assigned to group estimate

Q's:
What happens to the estimated α_j,
when respectively n_j, σ_y, or σ_α:

1) increases?
2) decreases?
3) is 0?
4) goes towards infinity?

Gelman A, Hill J (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press

```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 | Subject)
   Data: sleepstudy
REML criterion at convergence: 1786.465
Random effects:
 Groups   Name        Std.Dev.
 Subject  (Intercept) 37.12
 Residual             30.99
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

$\hat{\sigma}_{\alpha}$

$\hat{\sigma}_{y}$

$$j = 308$$

$$n_{308} = length(y_{308}) = 10$$

$$\bar{y}_{308} = mean(y_{308})$$

$$\bar{x}_{308} = mean(x_{308})$$

```
Call:
lm(formula = Reaction ~ Days,
data = sleepstudy)
Coefficients:
(Intercept)         Days
     251.41        10.47
```

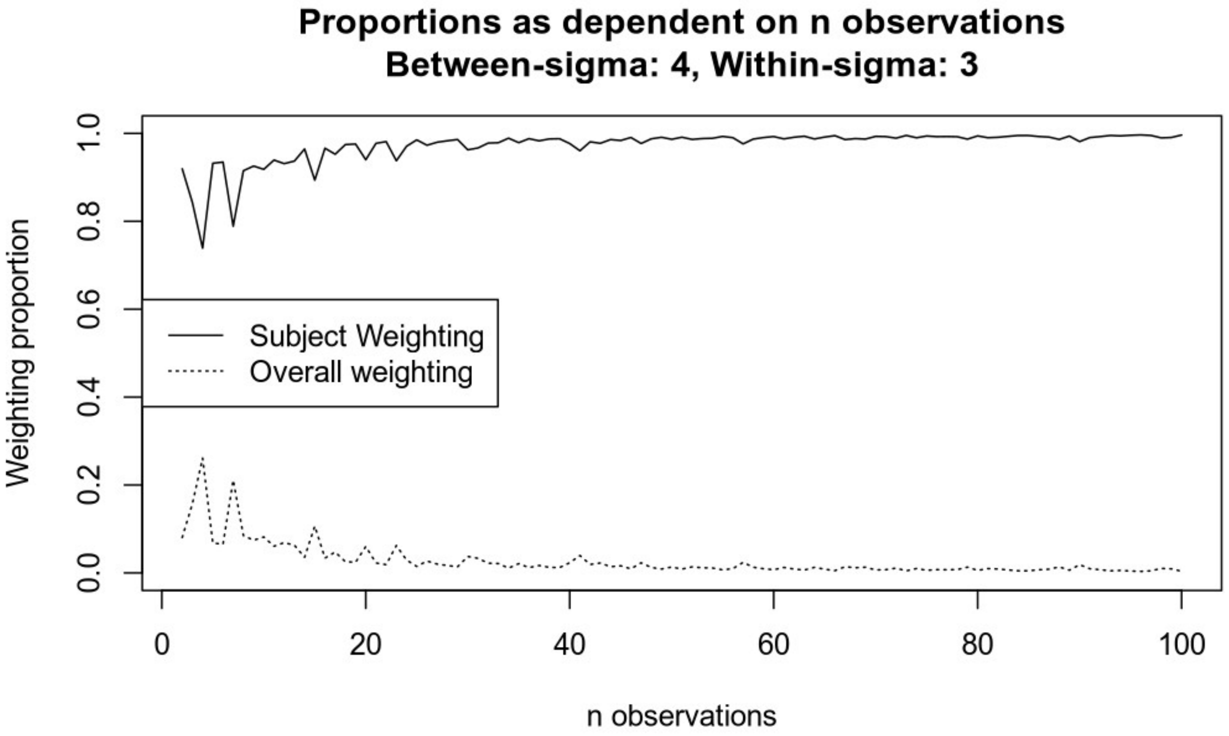
$\hat{\mu}_{\alpha}$

```
Call:
lm(formula = Reaction ~ Days + Subject - 1,
data = sleepstudy)
Coefficients:
      Days  Subject308  Subject309  ...
     10.47      295.03      168.13  ...
```

$\hat{\beta}$

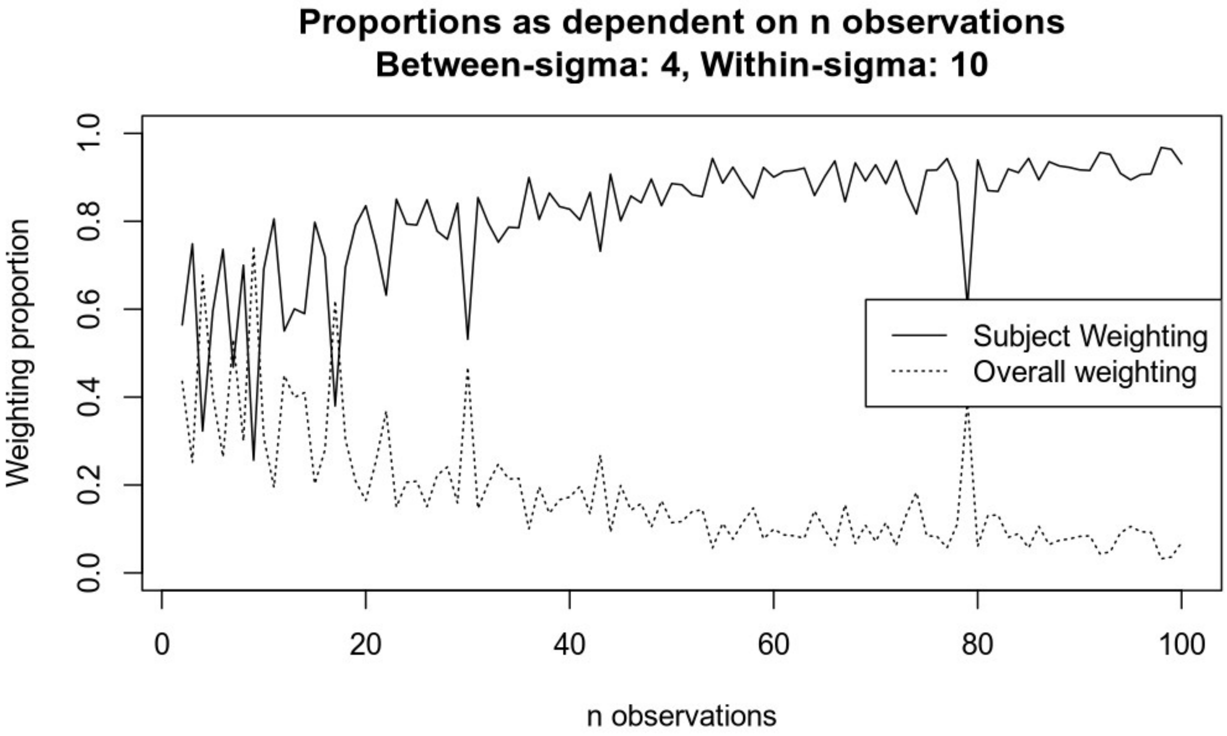
When $\sigma_\alpha > \sigma_y$

100 SIMULATIONS



When $\sigma_\alpha < \sigma_y$

100 SIMULATIONS



The course plan

Week 1: *Introduction*

Instructor sessions: *Setting up* R *and* Python *and recollection of the general linear model*

Week 2: *Multilevel linear regression*

Instructor sessions: *Modelling subject level effects – and how do they differ from group level effects?*

Week 3: *Link functions and fitting generalised linear multilevel models*

Instructor sessions: *What to do when the response variable is not continuous?*

Week 4: *Evaluating Generalised linear mixed models*

Instructor sessions: *How do we assess how models compare to one another?*

Week 5: *Explanation and Prediction*

Instructor sessions: *Code review*

Week 6: *Mid-way evaluation and Machine Learning Intro*

Instructor sessions: *Getting Python Running*

Week 7: *Linear regression revisited (machine learning)*

Instructor sessions: *How to constrain our models to make them more predictive*

Week 8: *Logistic regression revisited (machine learning)*

Instructor sessions: *Categorizing responses based on informed guesses*

Week 9: *Dimensionality Reduction, Principled Component Analysis (PCA)*

Instructor sessions: *What to do with very rich data?*

Week 10: *Outlook, unsupervised classification and neural networks*

Instructor sessions: *Data with no labels and networks*

Week 11: *Organising and preprocessing messy data*

Instructor sessions: *Code review*

Week 12: Final evaluation and wrap-up of course

Instructor sessions: *Ask anything!*

Next time –
Link functions and fitting generalised linear multilevel models

- Generalising general models
- Link functions
  - transforming from one scale to another
  - inverse functions
- Maximum likelihood estimation

Reading questions

- Chapter 5, Gelman & Hill
  - Why do we want to transform $X_i\beta$ using $logit^{-1}$?
  - What is an odds ratio?
  - How are logistic regression coefficients interpreted?
- Sections 6.1 and 6.2 Gelman & Hill
  - What is the parameter, θ, of the Poisson distribution?
  - Why may overdispersion be an issue?