Methods 3: Multilevel Statistical Modeling and Machine Learning

Week 2: *Linear mixed effects models*
September 21, 2021

by: Lau Møller Andersen

Follow up from last time:

Martin’s point: if population variance (the assumption of the *z*-test) is known, population mean is also known

~~If we are testing the null hypothesis, then what is *z*?~~

$$\cancel{P=0}$$

$$\cancel{z=\frac{\bar{X}}{SE}}$$

$$\bar{X}-P=0$$

$$z=\frac{\bar{X}-P}{SE}$$

Follow up from last time:

Martin’s point: if population variance (the assumption of the *z*-test) is known, population mean is also known

If we are testing the null hypothesis (that the population mean is 0, (and that the variance is unknown), then what is *t*?

$$P=0$$

$$t=\frac{\bar{X}}{SE}$$

Follow up from last time:

- Sigurd's point: there is a difference between likelihood and probability
- First glance

like·li·hood (līk′lē-ho͝od′)

n.

1. The state of being probable; probability.

2. Something probable.

Follow up from last time:

- Sigurd’s point: there is a difference between likelihood and probability
- Second glance

likelihood (ˈlaɪklɪˌhʊd) *or* likeliness

n

1. the condition of being likely or probable; probability

2. something that is probable

3. (Statistics) *statistics* the probability of a given sample being randomly drawn regarded as a function of the parameters of the population. The likelihood ratio is the ratio of this to the maximized likelihood. See also **maximum likelihood**

Follow up from last time:

Likelihood: $L(\theta \mid O)$

Probability: $P(O \mid \theta)$

θ: the unknown parameters, e.g. $\hat{\beta}$

O: the observations from a given sample

- Likelihoods measures to what degree *O* provides support for θ
- Probabilities indicate which observations we can expect given the parameters (which are mostly unknown)

Follow up from last time coin tossing

Assuming a fair coin (θ = 0.5), the **probability** of two heads (*O*):

$P(O \mid \theta) = 0.25$

On the other hand, what is the **likelihood** ($L(\theta \mid O)$) of $\theta = 0.5$, given 7 heads in a row (*O*)?

$L(\theta \mid O) = 0.01563$

Follow up from last time coin tossing

```
n.heads <- 7
n.flips <- 7
theta <- 0.5

binom.test(n.heads, n.flips, theta)
```

```
##
##  Exact binomial test
##
## data:  n.heads and n.flips
## number of successes = 7, number of trials = 7, p-value = 0.01563
## alternative hypothesis: true probability of success is not equal to 0.
5
## 95 percent confidence interval:
##  0.5903836 1.0000000
## sample estimates:
## probability of success
##                      1
```

Follow up from last time – gremlins bowling in the attic

- You hear noise in the attic (*O*):
  - What is the **probability** of there being gremlins bowling up there?
  - What is the **likelihood** of the noise (*O*) given that there are gremlins bowling up there?

Example courtesy of Elliott Sober

Questions from last time?

Learning goals and outline –

Linear Mixed Effects Models (LMM)

1) Why can it be a good idea to do mixed effects modelling?
2) Understanding the basics of multilevel modelling

   - also known as linear mixed effects modelling

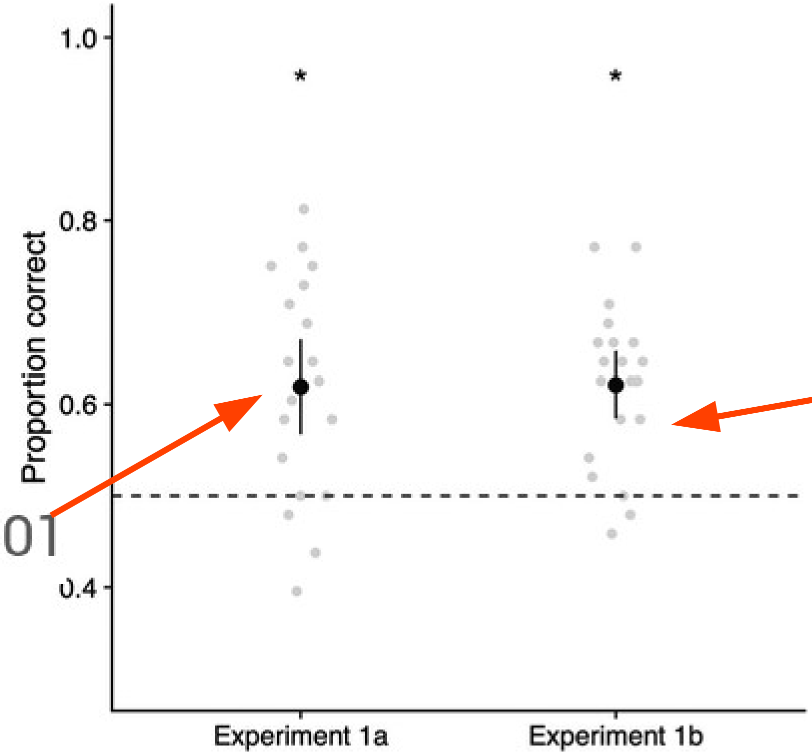
3) Appreciating the difference between the different levels of effects
   - or *random* and *fixed* effects, as they are also called

This week’s practical exercise

- The *politeness* dataset from Winter and Grawunder (2012) – looking into how politeness requirements change the pitch of speech
  - Exercise 1 - describing the dataset and making some initial plots
  - Exercise 2 - comparison of models
  - Exercise 3 - now with attitude

Motivation – (not) using the available information

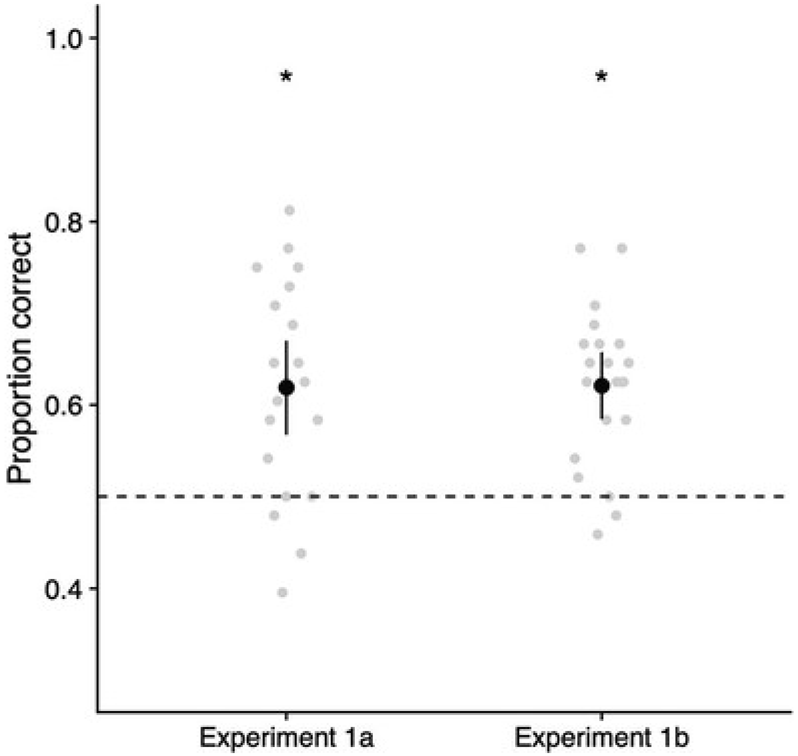
$t(19) = 4.54, p < .001$



$t(20) = 6.50, p < .001$

(Krishnan et al., 2021)

Motivation – (not) using the available information



Sequence Recognition

A 2AFC test was used to assess statistical learning of the probabilistic structure of the tone sequences previously presented. Tone words from L1 were paired exhaustively with tone words from L2 to create 16 distinct pairs of tone words. Each pair was presented three times, producing 48 trials in total, [...]

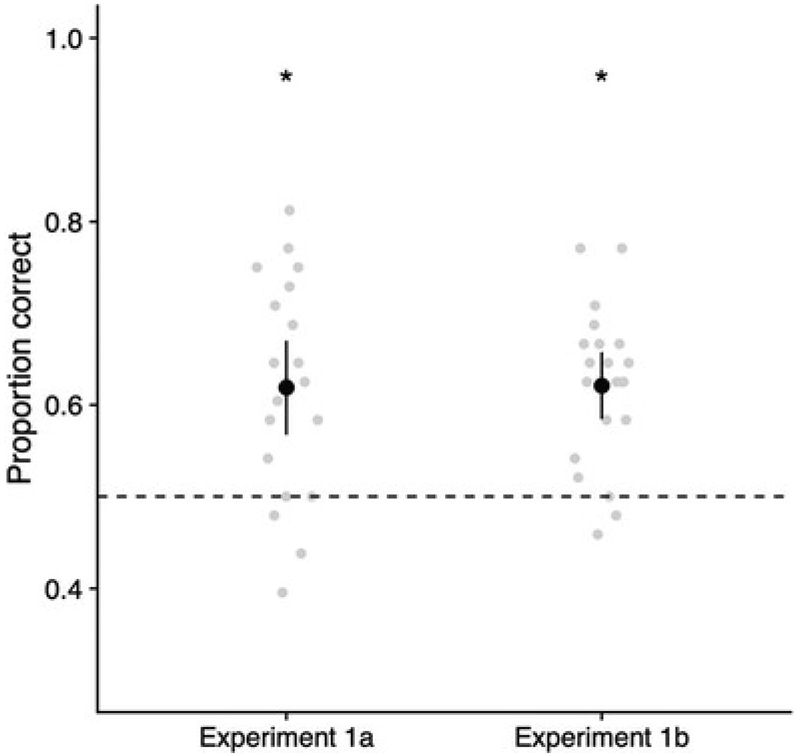
(Krishnan et al., 2021)

The general linear model

$$Y = X\beta + \epsilon$$

Y : a column vector with J observations (known)
X : the design matrix (known), size : $J \times L$
β : a column vector with L (unknown) model parameters
ϵ : a column vector with J residuals, normally distributed, mean=0

Motivation – (not) using the available information



Group discussions – write your answers in CryptPad and send:

- What do each of the grey dots represent?
- Within the General Linear Model framework, what would be an appropriate model to fit at the single subject level?
  - Why would it be problematic to do this at the group level?
- At what levels of performance would the assumption of the normal distribution of the residuals be unfounded?

(Krishnan et al., 2021)

Motivation for multilevel modelling:
We want to use all the information in the data while fulfilling the assumptions necessary for the residuals

The general linear model

$$Y = X\beta + \epsilon$$

$Y:$ a column vector with J observations (known)
$X:$ the design matrix (known), size: $J \times L$
$\beta:$ a column vector with L (unknown) model parameters
$\epsilon:$ a column vector with J residuals, normally distributed, mean=0

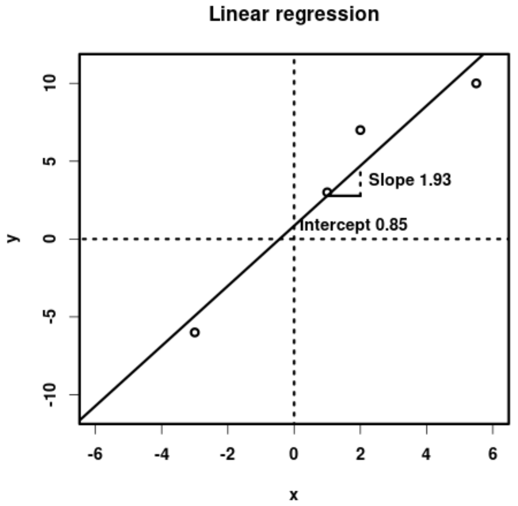
Linear regression is a special case of the general linear model

$$y = \alpha x + \beta + \epsilon$$

x	y
-3	-6
1	3
2	7
5.5	10

$$Y = X\ \beta + \epsilon$$

y : *the observed values*
$\hat{y}$: *the estimated values*
α: *slope of the line*
x : *the independent variable*
β: *the intercept , the value when* $x = 0$
ϵ: *the difference between* y and $\hat{y}$



X:
(design matrix 4 rows and 2 columns)

x^1 (α)	x^0 (β)
-3	1
1	1
2	1
5.5	1

Quadratic regression

$$y = ax^2 + bx + c + \epsilon$$

$$Y = X\,\beta + \epsilon$$

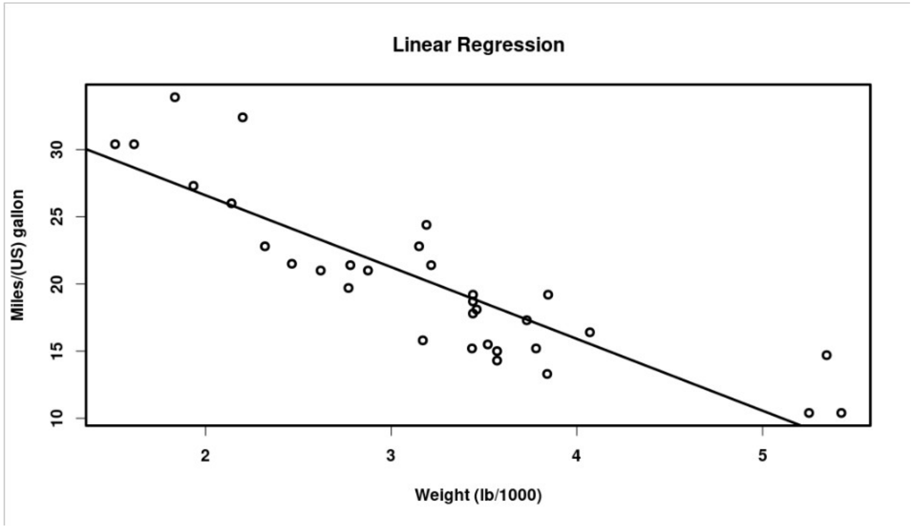
X:
(design matrix 4 rows and 3 columns)

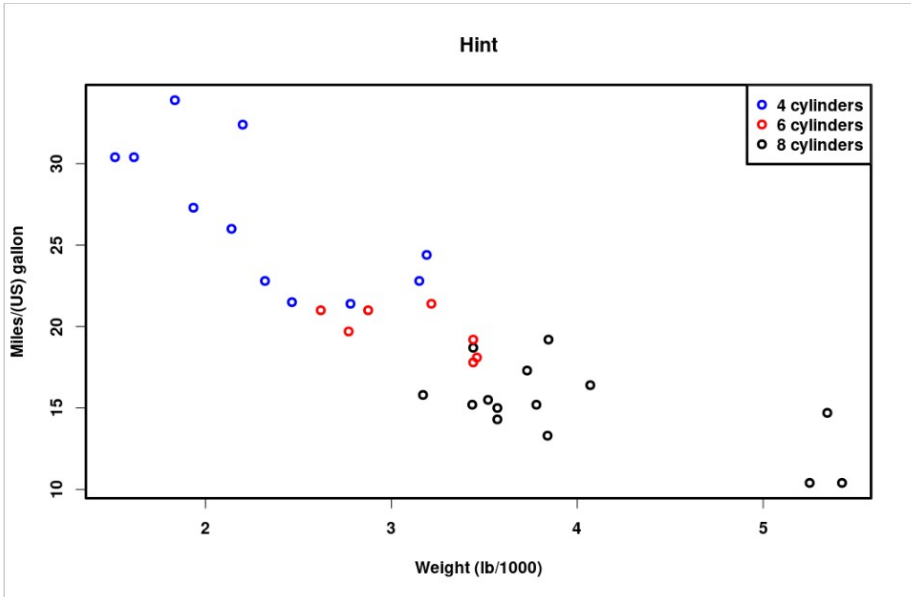
x	*y*
-3	-6
1	3
2	7
5.5	10

x^2 *(a)*	x^1 *(b)*	x^0 *(c)*
9	-3	1
1	1	1
4	2	1
30.25	5.5	1

... and we can carry on *ad nauseam*, making cubic models, models of the fourth-order etc.

What’s a hidden assumption of linear regression? (let me know if you want a hint)

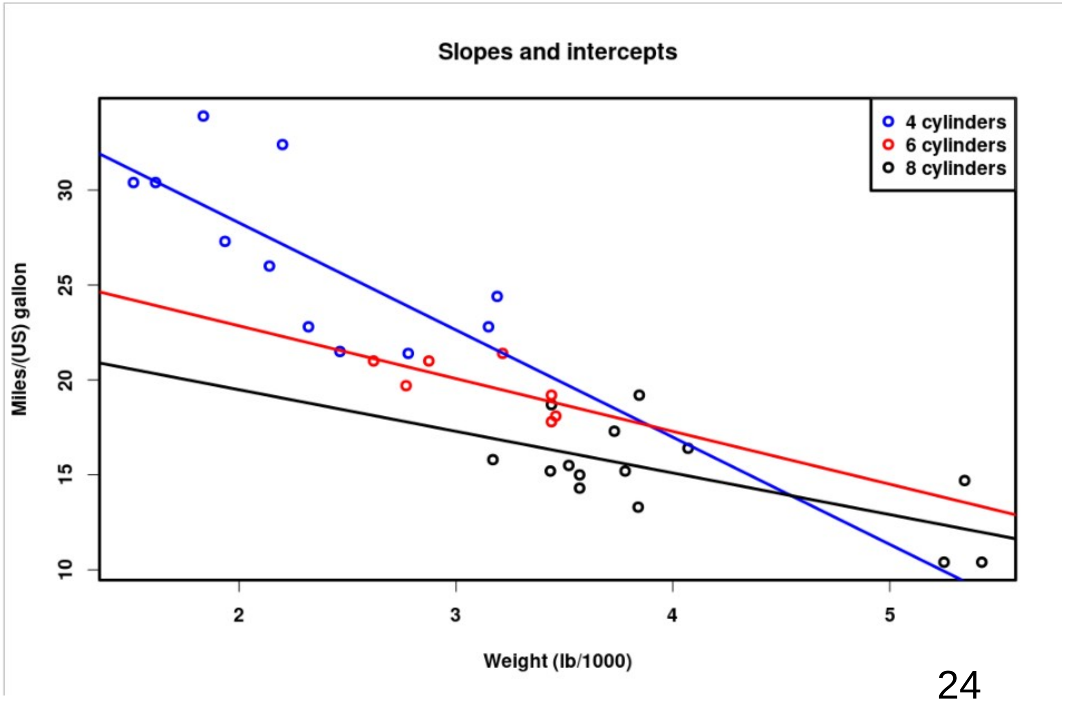




Hidden assumption: the *intercept* and the *slope* are the same for the whole sample, despite what other groupings there are, e.g. *number of cylinders* or *number of carburettors*

NB! The lines fitted here are based on three separate linear models – not the best way to do it

```
model.4 <- lm(mpg ~ wt, data=mtcars, subset=mtcars$cyl == 4)
model.6 <- lm(mpg ~ wt, data=mtcars, subset=mtcars$cyl == 6)
model.8 <- lm(mpg ~ wt, data=mtcars, subset=mtcars$cyl == 8)
```



Introducing multilevel modelling –
modelling individual slopes and intercepts

Level 1 $y_{cyl} = \alpha_{cyl} x_{cyl} + \beta_{cyl} + \epsilon_{cyl}$

Level 2
$$\alpha_{cyl} = \gamma_1 + S_{\alpha,cyl}$$
$$\beta_{cyl} = \gamma_2 + S_{\beta,cyl}$$

looks scary...

Variance components $\langle S_{\alpha,cyl}, S_{\beta,cyl} \rangle \sim N(\langle 0, 0 \rangle, \Sigma)$

$$\Sigma = \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix}$$

$$\epsilon_{cyl} \sim N(0, \sigma^2)$$

Level 1

very similar to what we have seen before

$$y_{cyl} = \alpha_{cyl} x_{cyl} + \beta_{cyl} + \epsilon_{cyl}$$

$$y_{cyl} : observed\, values\, for\, each\, cylinder\, group$$
$$\alpha_{cyl} : slope\, for\, each\, cylinder\, group\, ; defined\, at\, Level\, 2$$
$$x_{cyl} : independent\, data\, for\, each\, cylinder\, group$$
$$\beta_{cyl} : intercept\, for\, each\, cylinder\, group\, ; defined\, at\, Level\, 2$$
$$\epsilon_{cyl} : residuals\, for\, each\, cylinder\, group$$

Level 2

$$\alpha_{cyl} = \gamma_1 + S_{\alpha,cyl}$$
$$\beta_{cyl} = \gamma_2 + S_{\beta,cyl}$$

$\alpha_{cyl}: level\,1\,slope$

$\beta_{cyl}: level\,1\,intercept$

$\gamma_1: grand\ slope\,(\text{fixed}\ effect)$

$\gamma_2: grand\ intercept\,(\text{fixed}\ effect)$

$S_{\alpha,cyl}: slope\ for\ each\ cylinder\ group\,(\text{random}\ effect)$

$S_{\beta,cyl}: intercept\ for\ each\ cylinder\ group\,(\text{random}\ effect)$

Variance components

$$\langle S_{\alpha,cyl}, S_{\beta,cyl} \rangle \sim N\left(\langle 0, 0 \rangle, \Sigma\right)$$

$$\Sigma = \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix}$$

$$\epsilon_{cyl} \sim N\left(0, \sigma^2\right)$$

The random effects follow a bivariate normal distribution with mean=0, and variation specified by the covariance matrix, Σ. (can include as many dimensions as you like)

Σ has the variance of the slope, τ_α^2, and the variance of the intercept, τ_β^2 on the diagonal and the covariance off-diagonal with ρ being the correlation factor between the individual slopes and intercepts

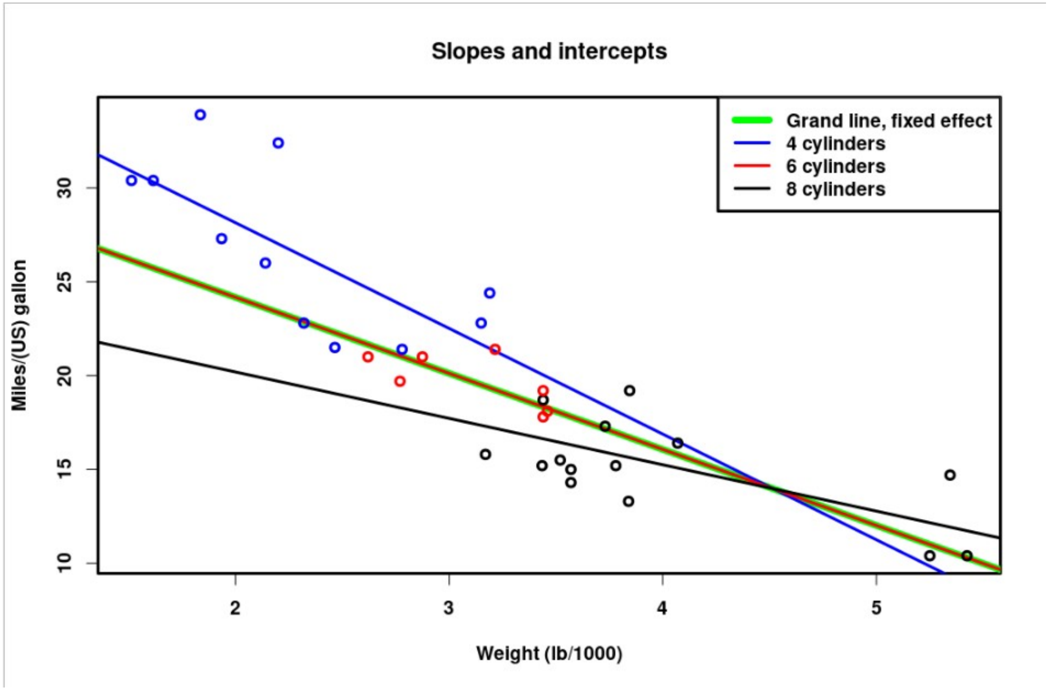
Finally, the Level 1 error, is modelled as normally distributed, mean=0 (just as in the General Linear Model).

Fixed and random effects

- Fixed effects
  - exhaust the population
  - express average effects
  - can be categorical or continuous
- Random effects
  - sample the population
  - express individual effects
  - has to be categorical

Let’s model it again

```
mixed.effects.model <- lmer(mpg ~ wt + (wt | cyl), data=mtcars)
```



mpg: dependent variable

wt: independent variable (fixed effect)

(wt | cyl): individual slopes (and intercepts) (random effects)

(+ 1): implicit intercept (fixed effect)

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   32.273      4.746   6.801
## wt            -4.052      1.102  -3.676
```

```
ranef(mixed.effects.model)
```

```
## $cyl
##    (Intercept)            wt
## 4  7.135243949 -1.5795832665
## 6  0.003908138 -0.0008645995
## 8 -7.139152086  1.5804478660
##
## with conditional variances for "cyl"
```

mean(7.135; 0.003908; -7.139) ≈ 0

mean(-1.580; -0.0008645; 1.580) ≈ 0

$$\langle S_{\alpha,cyl}, S_{\beta,cyl} \rangle \sim N(\langle 0,0 \rangle, \Sigma)$$

$$\Sigma = \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix}$$

Variance and correlation

τ_α^2 τ_β^2 ρ

```
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  cyl      (Intercept) 56.92    7.545
##           wt           2.79    1.670    -1.00
##  Residual              5.61    2.369
## Number of obs: 32, groups:  cyl, 3
```

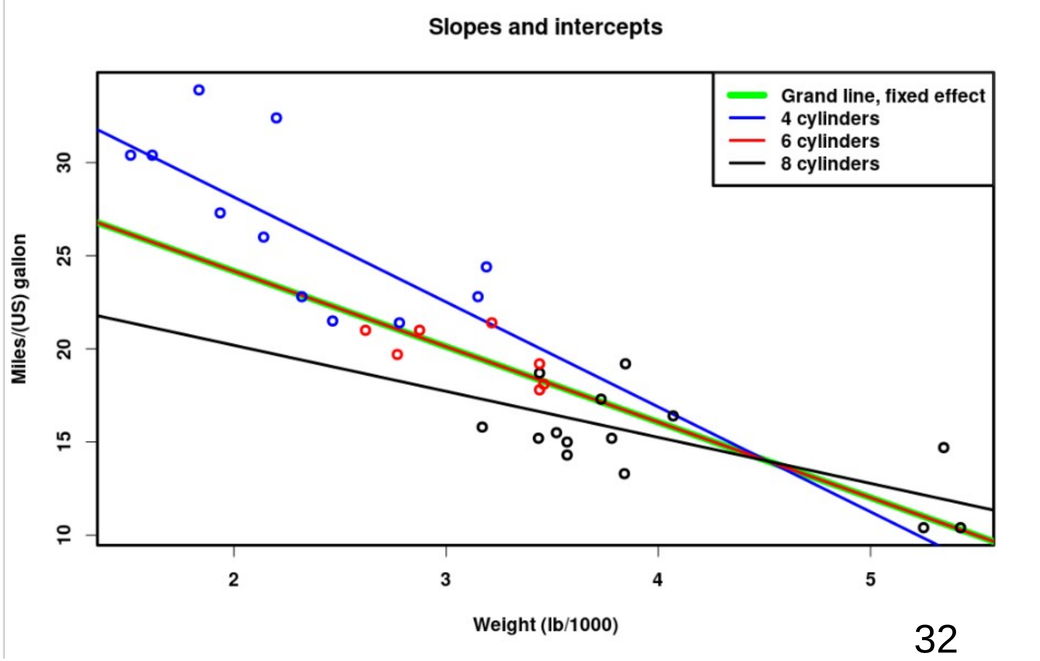
$$\alpha_{cyl} = \gamma_1 + S_{\alpha,cyl}$$
$$\beta_{cyl} = \gamma_2 + S_{\beta,cyl}$$

```
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) γ2  32.273      4.746   6.801
## wt          γ1  -4.052      1.102  -3.676
```

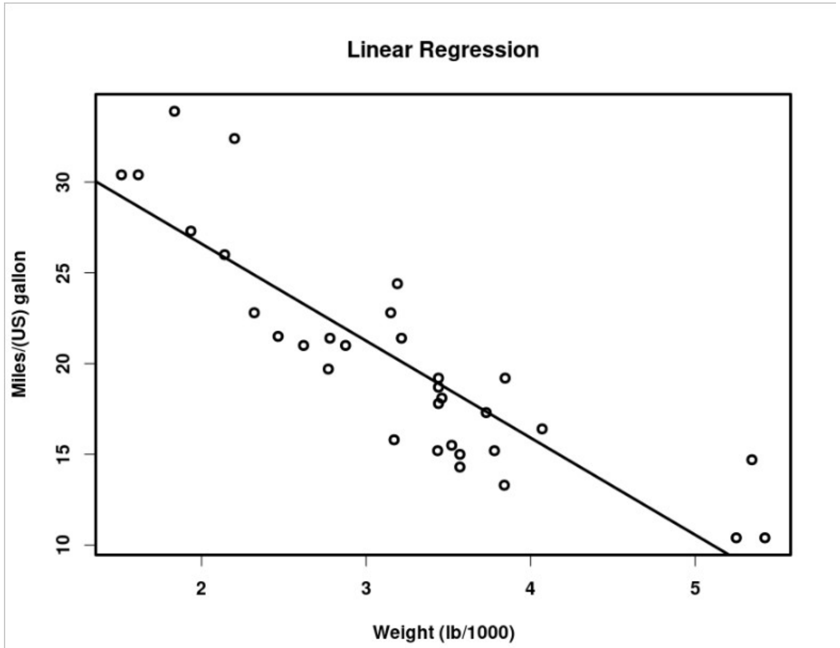
+

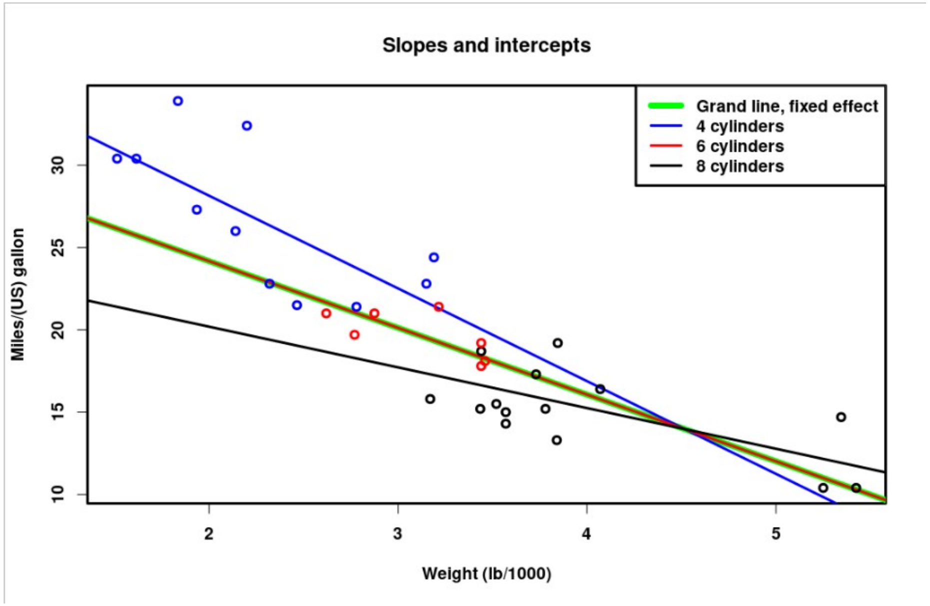
```
ranef(mixed.effects.model)
```

```
## $cyl
##     (Intercept)            wt
## 4   7.135243949 -1.5795832665
## 6   0.003908138 -0.0008645995
## 8  -7.139152086  1.5804478660
##
## with conditional variances for "cyl"
```



Comparison with single-level model



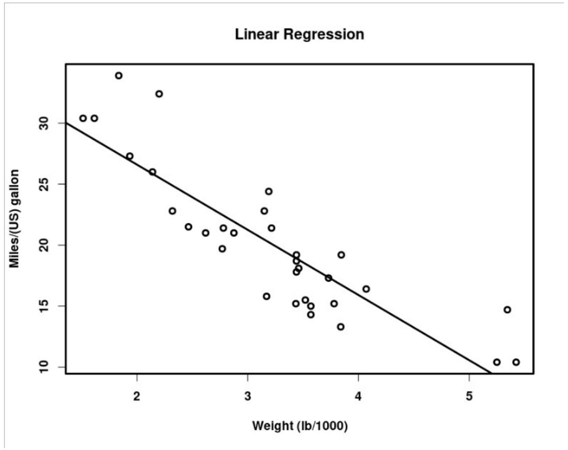


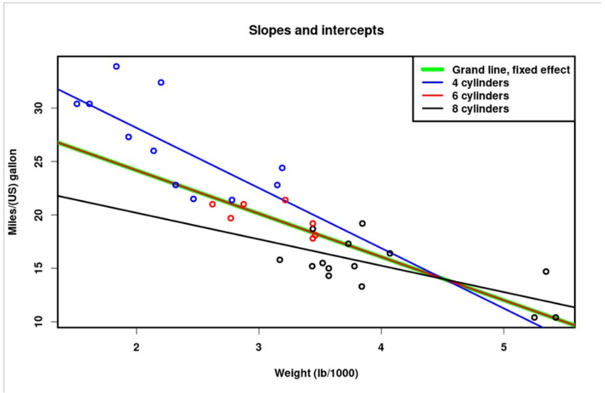

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

```
fixef(lmer(mpg ~ wt +(wt | cyl), data=mtcars ))
```

```
## (Intercept)          wt
##   32.273493   -4.052103
```

Comparison with single-level model





$$\text{residual variance} = \sum_{i=1}^{n} \epsilon_i^2$$

```
sum(residuals(model.basic)^2)
```

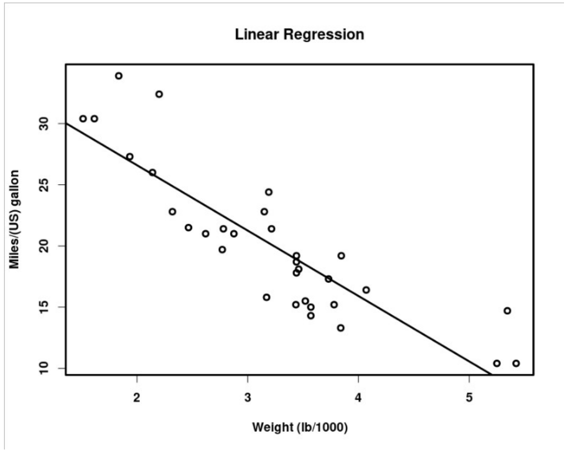
```
## [1] 278.3219
```

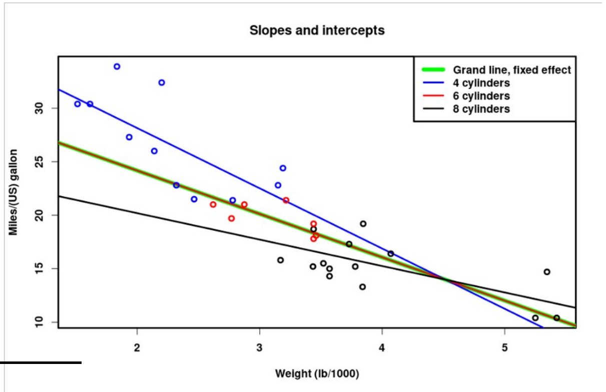
```
sum(residuals(mixed.effects.model)^2)
```

```
## [1] 158.2694
```

Which model is better according to this measure?
And why is residual variance a biased measure (hint: think back on the exercise comparing quadratic and cubic fits)

Comparison with single-level model





$$\text{residual standard deviation} = \sqrt{\left(\frac{\sum_{i=1}^{n} \epsilon_i^2}{df}\right)}$$

```
sigma(model.basic)
```

```
## [1] 3.045882
```

How is the bias from before controlled here?

```
sigma(mixed.effects.model)
```

```
## [1] 2.368621
```

Residual standard deviation

$$\text{residual standard deviation} = \sqrt{\left(\frac{\sum_{i=1}^{n} \epsilon_i^2}{df}\right)}$$

$\sum_{i=1}^{n} \epsilon_1^2$: variance of the residuals: (unexplained variance)

df : degrees of freedom; n_observations minus n_model_parameters

Natural hierarchies:

- Observations nested within subjects
- Subjects nested within e.g. schools
- Schools nested within counties
- Counties nested ...
- etc.

SLEEP STUDY EXAMPLE

https://psyteachr.github.io/stat-models-v1/introducing-linear-mixed-effects-models.html

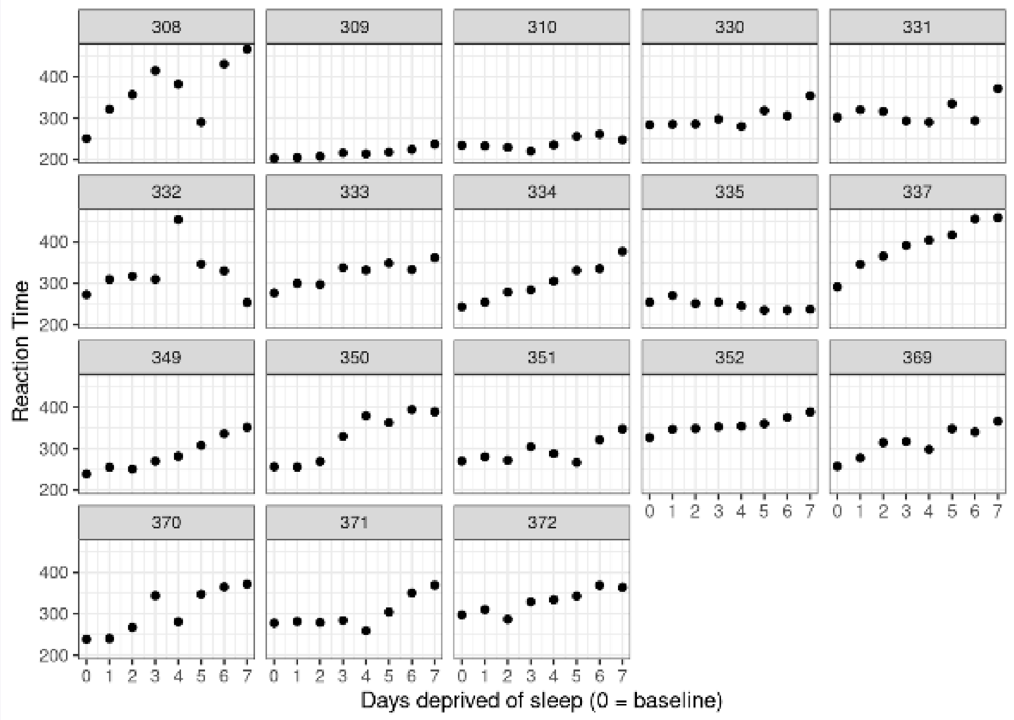
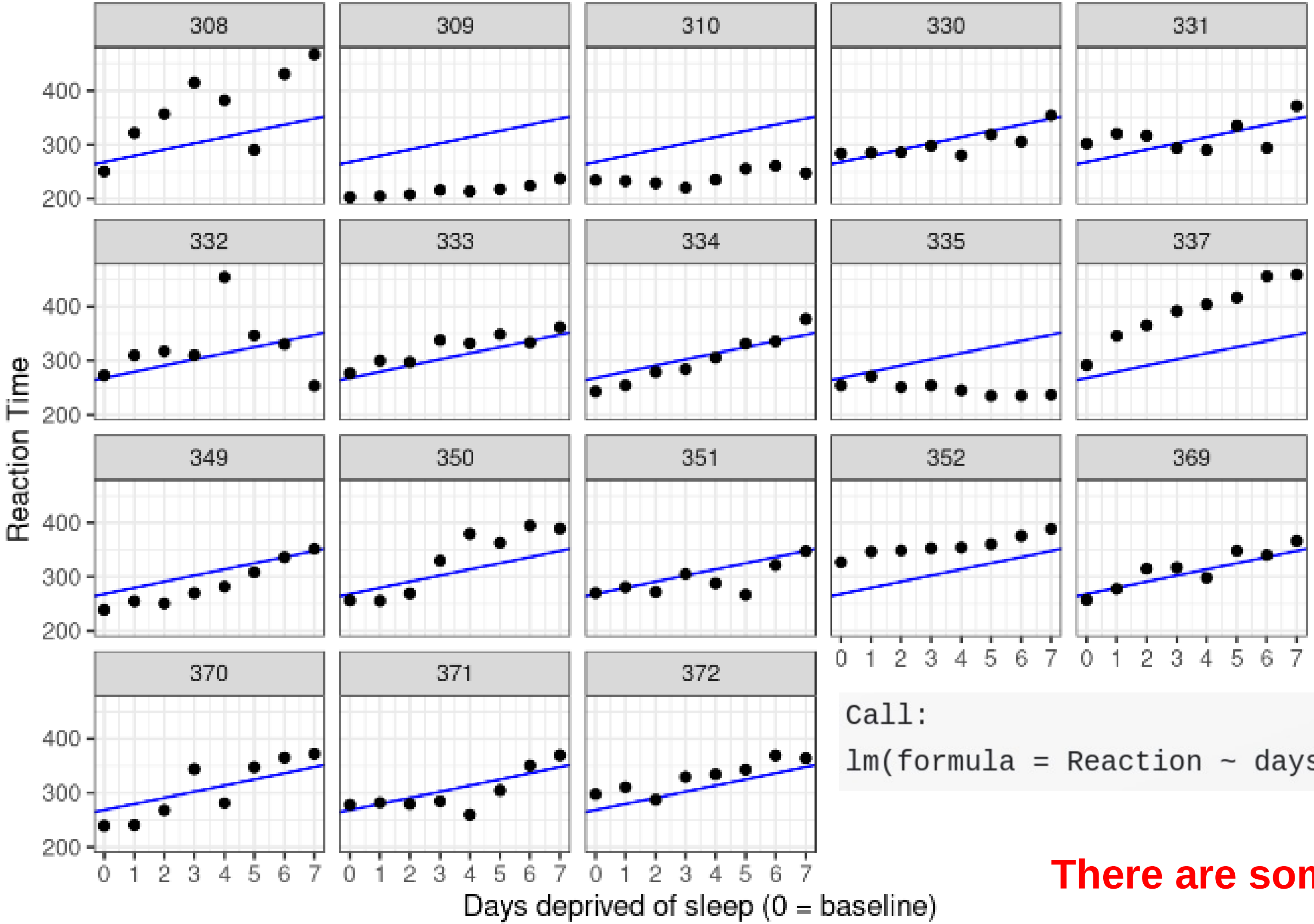


Figure 5.3: *Data from Belenky et al. (2003), showing reaction time at baseline (0) and after each day of sleep deprivation.*

Let's describe the dataset together

(Barr, University of Glasgow)



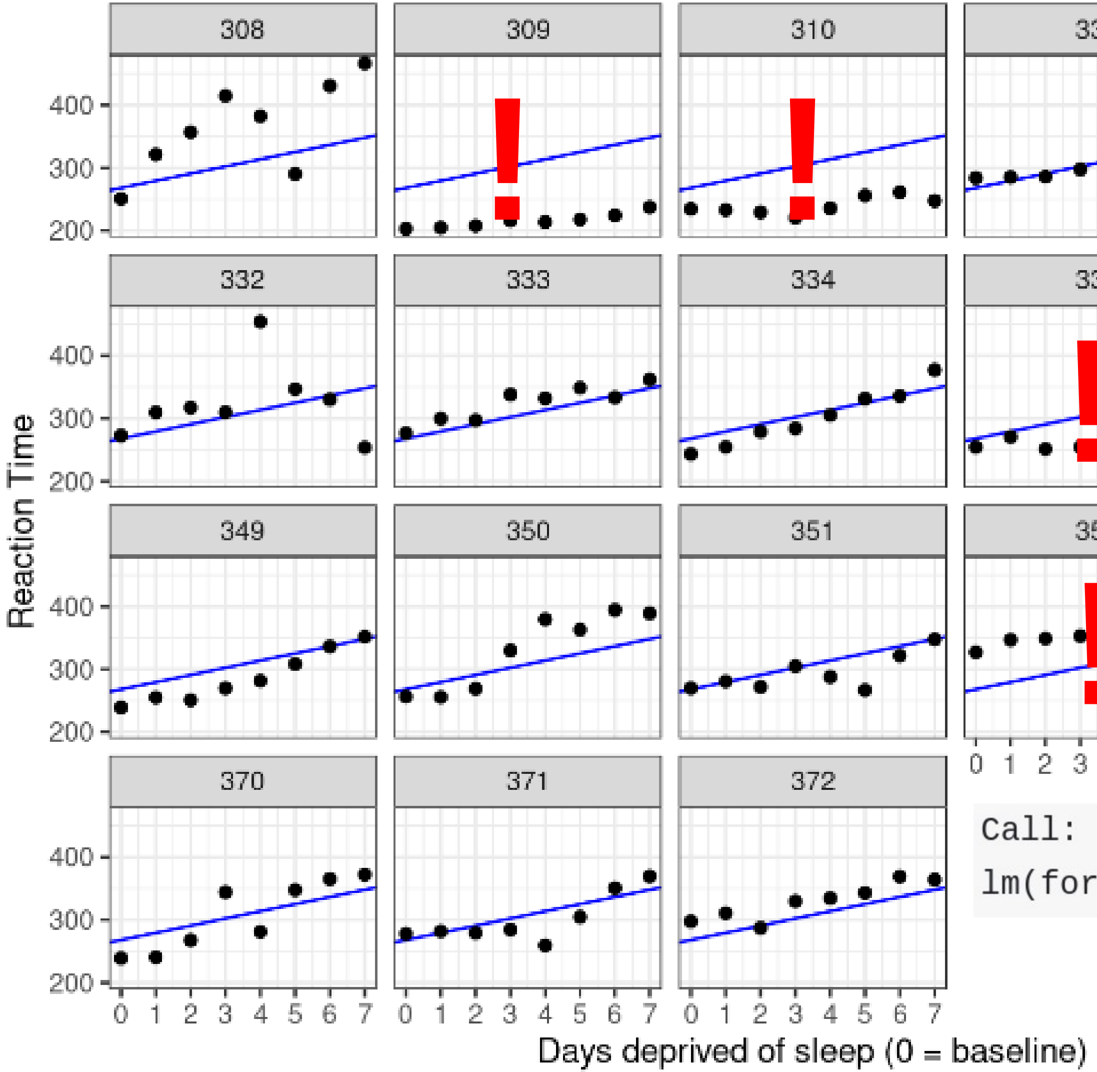
COMPLETE POOLING

```
Coefficients:
                Estimate
(Intercept)      267.967
days_deprived     11.435
```

```
Call:
lm(formula = Reaction ~ days_deprived, data = sleep2)
```

There are some bad fits

(Barr, University of Glasgow)




```
Coefficients:
             Estimate
(Intercept)   267.967
days_deprived  11.435
```

```
Call:
lm(formula = Reaction ~ days_deprived, data = sleep2)
```

There are some bad fits

(Barr, University of Glasgow)

```
lm(formula = Reaction ~ days_deprived + Subject + days_deprived:Subject,
    data = sleep2)
```

```
## Coefficients:
##                                Estimate
## (Intercept)                    288.2175
## days_deprived                   21.6905
## Subject309                     -87.9262
## Subject310                     -62.2856
## Subject330                     -14.9533
## Subject331                       9.9658
## Subject332                      27.8157
```

... and the remaining 12 subjects

```
## days_deprived:Subject309 -17.3334
## days_deprived:Subject310 -17.7915
## days_deprived:Subject330 -13.6849
## days_deprived:Subject331 -16.8231
## days_deprived:Subject332 -19.2947
## days_deprived:Subject333 -10.8151
```

... and the remaining 12 subjects

NO POOLING

(Barr, University of Glasgow)

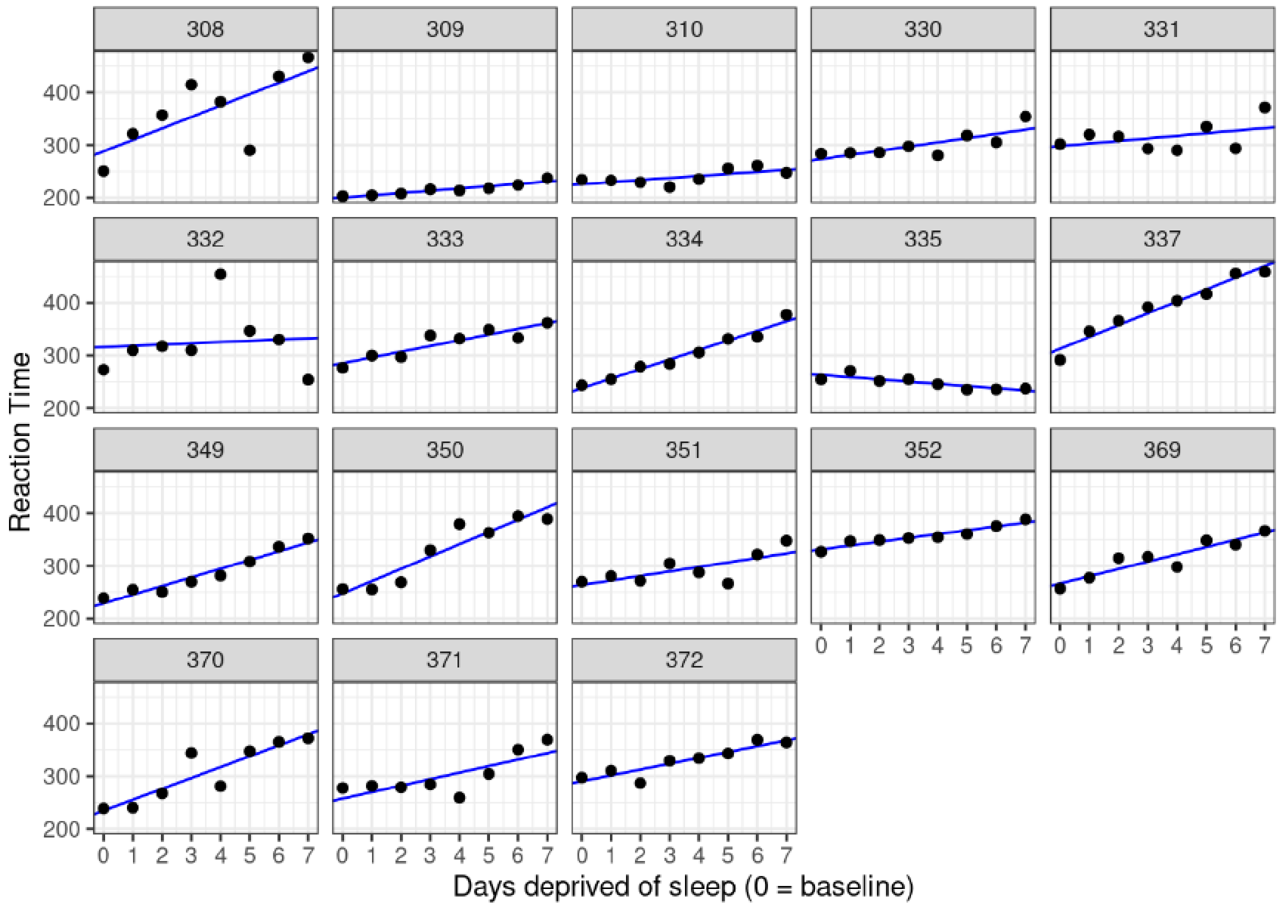


Figure 5.5: Data plotted against fits from the no-pooling approach.

NO POOLING

Good fits now:

What are the limits of this model?

(Barr, University of Glasgow)

PARTIAL POOLING

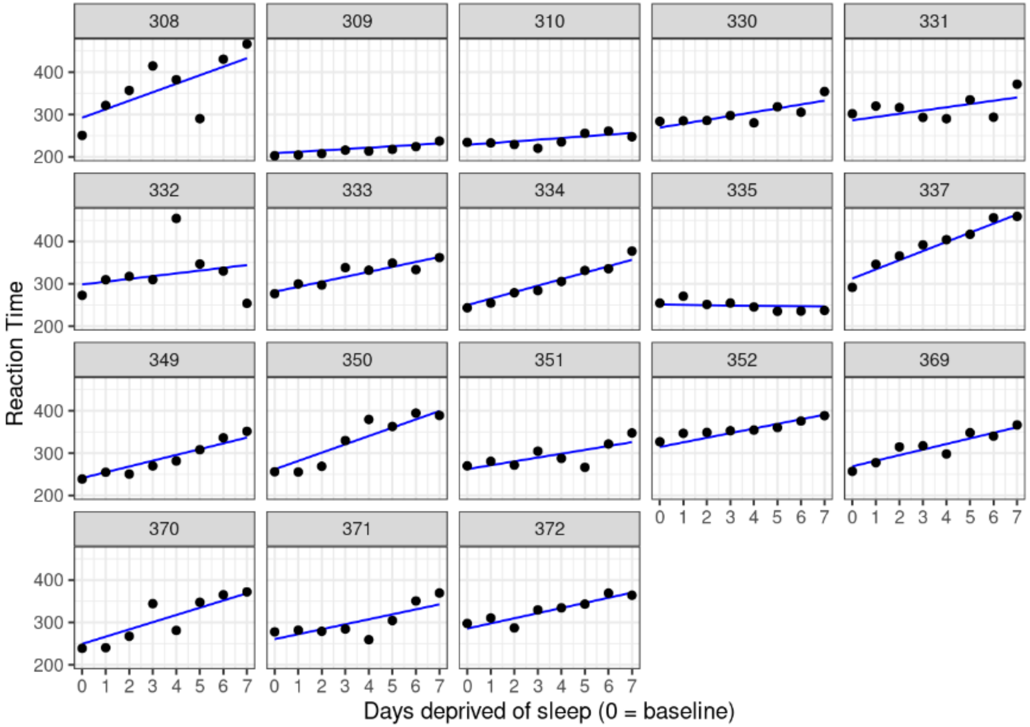


Figure 5.6: Data plotted against predictions from a partial pooling approach.

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)   267.967      8.266  32.418
days_deprived  11.435      1.845   6.197
```

```
ranef(pp_mod)[["Subject"]]
```

	(Intercept)	days_deprived
308	24.4992891	8.6020000
309	-59.3723102	-8.1277534
310	-39.4762764	-7.4292365
330	1.3500428	-2.3845976

```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ days_deprived + (days_deprived | Subject)
   Data: sleep2
```


(Barr, University of Glasgow)

No pooling vs partial pooling

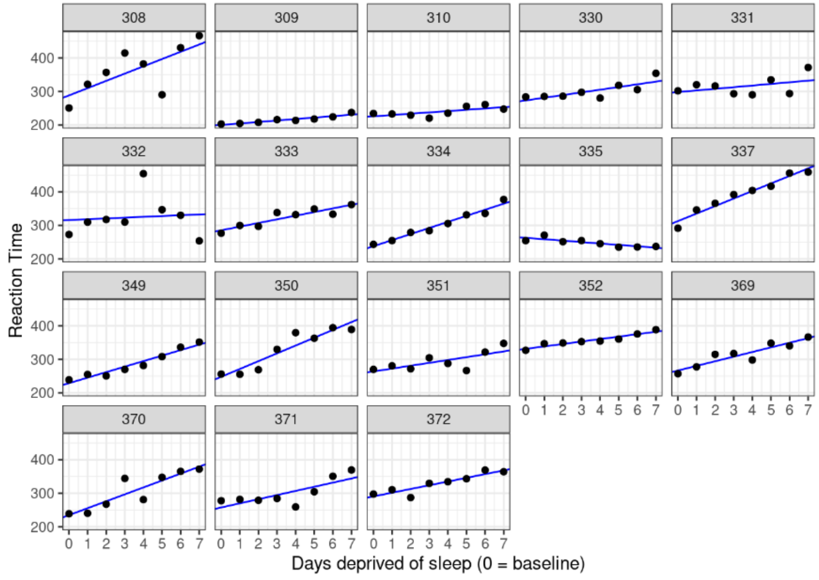


Figure 5.5: Data plotted against fits from the no-pooling approach.

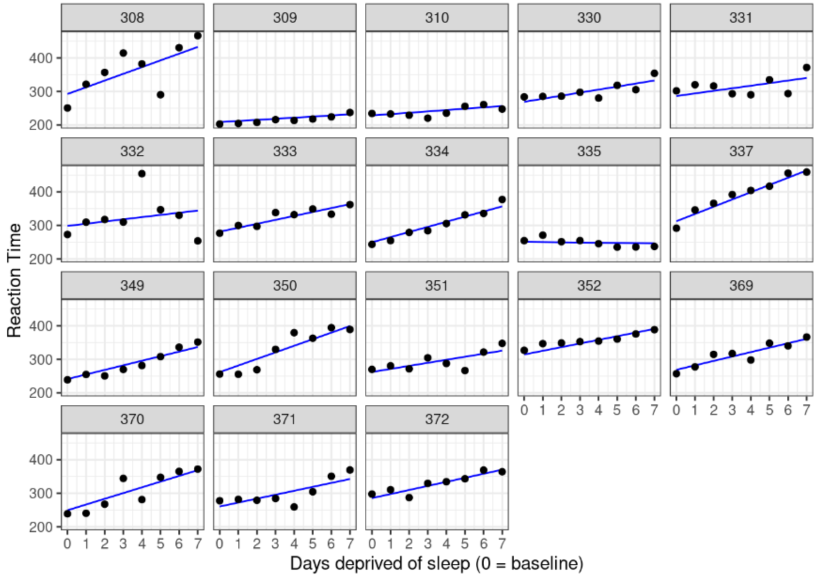


Figure 5.6: Data plotted against predictions from a partial pooling approach.

Both model the individual variance – but only one is generalisable outside the subject pool

(Barr, University of Glasgow)

Pooling - summary

- Complete pooling
  - ignoring a categorical predictor (e.g. *subject*)
- No pooling
  - model each level of the categorical predictor separately
- Partial pooling
  - we model both an average and each level

Factorial model (ANOVA) – *may be useful in exercise*

$Y =$

Observations	ON/OFF
5	ON
7	ON
1	ON
9	ON
6	ON
3	OFF
5	OFF
4	OFF
6	OFF
2	OFF

$X =$?

Factorial model (ANOVA) – *may be useful in exercise*

$Y =$

Observations	ON/OFF
5	ON
7	ON
1	ON
9	ON
6	ON
3	OFF
5	OFF
4	OFF
6	OFF
2	OFF

$X =$

Column 1	Column2
1	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1
0	1

Learning goals and outline –

Linear Mixed Effects Models (LMM)

1) Why can it be a good idea to do mixed effects modelling?

2) Understanding the basics of multilevel modelling

   - also known as linear mixed effects modelling

3) Appreciating the difference between the different levels of effects

   – or *random* and *fixed* effects, as they are also called

Next time:
Modelling binomial and count data, (or anything)
introducing generalized linear mixed models
(GLMM)

References

- Belenky, G., Wesensten, N.J., Thorne, D.R., Thomas, M.L., Sing, H.C., Redmond, D.P., Russo, M.B., Balkin, T.J., 2003. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. Journal of Sleep Research 12, 1–12. https://doi.org/10.1046/j.1365-2869.2003.00337.x
- Krishnan, S., Carey, D., Dick, F., Pearce, M.T., 2021. Effects of statistical learning in passive and active contexts on reproduction and recognition of auditory sequences. Journal of Experimental Psychology: General. https://doi.org/10.1037/xge0001091
- Sober, E., 2018. Philosophy Of Biology. Routledge.
- Winter, B., Grawunder, S., 2012. The phonetic profile of Korean formal and informal speech registers. Journal of Phonetics 40, 808–815. https://doi.org/10.1016/j.wocn.2012.08.006